ЦИФРОВАЯ СВЯЗЬ

Теоретические основы и практическое применение

Второе издание

DIGITAL COMMUNICATIONS

Fundamentals and Applications

Second Edition

BERNARD SKLAR
Communications Engineering Services, Tarzana, California
and
University of California, Los Angeles

Prentice Hall P T R
Upper Saddle River, New Jersey 07458
www.phptr.com

ЦИФРОВАЯ СВЯЗЬ

Теоретические основы и практическое применение

Второе издание

БЕРНАРД СКЛЯР
Communications Engineering Services, Тарзана, Калифорния
и
University of California, Лос-Анджелес

Москва · Санкт-Петербург · Киев
2003

ББК 32.973.26-018.2.75
С43
УДК 681.3.07

Издательский дом "Вильямс"

Зав. редакцией *А.В. Слепцов*

Перевод с английского *Е.Г. Грозы*, *В.В. Марченко*, *А.В. Назаренко*, канд.физ.-мат.наук *О.М. Ядренко*

Под редакцией *А.В. Назаренко* и *Л.А. Худяковой*

Под общей редакцией *А.В. Назаренко*

По общим вопросам обращайтесь в Издательский дом "Вильямс" по адресу: info@williamspublishing.com, http://www.williamspublishing.com

Скляр, Бернард.

С43 Цифровая связь. Теоретические основы и практическое применение, 2-е издание. : Пер. с англ. – М. : Издательский дом "Вильямс", 2003. – 1104 с. : ил. – Парал. тит. англ.

ISBN 5-8459-0386-6 (рус.)

Данную книгу стоит прочесть всем, кто интересуется цифровой связью. Это учебник, в котором математически строго описаны все преобразования, которым подвергается информация на пути от источника к адресату. Это также и справочник, в котором дано описание схем, необходимых для практической реализации соответствующих математических абстракций. И наконец, это просто хорошая и интересная книга для всех тех, кто хочет узнать все о цифровой связи, прочитав всего одну серьезную и в то же время доступную работу.

ББК 32.973.26-018.2.75

ISBN 5-8459-0386-6 (рус.)
ISBN 0-1308-4788-7 (англ.)

ОГЛАВЛЕНИЕ

СОДЕРЖАНИЕ

ГЛАВА 9. КОМПРОМИССЫ ПРИ ИСПОЛЬЗОВАНИИ МОДУЛЯЦИИ И КОДИРОВАНИЯ 543

Предисловие

Книга *Цифровая связь: теоретические основы и практическое применение* является обновленной редакцией предыдущего издания. Сюда внесены следующие изменения.

- Расширены главы, посвященные кодам коррекции ошибок, особенно это относится к кодам Рида-Соломона, турбокодам и решетчатому кодированию.
- Введена глава, посвященная каналам с замираниями и способам смягчения последствий замирания.
- Расширены описания необходимых понятий области цифровой связи.
- Увеличено число задач, предлагаемых в завершение глав. Кроме того, добавлены вопросы для самопроверки (также указано, где искать ответы на них).

Структура семестрового университетского курса сильно отличается от структуры краткого курса по тому же предмету. В первом случае имеем достаточно времени на приобретение необходимых навыков, усвоение математического аппарата и применение теорий на практике (чему способствуют домашние задания). При чтении краткого курса преподаватель вынужден "пробегаться" по необходимым понятиям и приложениям. Как я обнаружил, определить структуру краткого курса помогают контрольные вопросы, предложенные слушателям курса. Эти вопросы — не только схематическое изображение учебного плана. Они представляют собой набор понятий и терминов, которые в настоящее время не очень корректно освещены в литературе, а иногда вообще неверно трактуются. При таком подходе студенты, прослушивающие краткий курс, заранее получают общее представление о нем. Со временем они смогут описать конкретную проблему и будут осведомлены об области цифровой связи вообще. (Личное наблюдение: предлагаемый перечень вопросов пригоден как для полного, так и сокращенного курса обучения.) Итак, я предлагаю следующий "контрольный список" вопросов по цифровой связи.

1. Какая математическая дилемма является причиной существования нескольких определений ширины полосы (см. раздел 1.7.2)?
2. Почему отношение энергии бита к спектральной плотности мощности шума (E_b/N_0) является естественным критерием качества систем цифровой связи (см. раздел 3.1.5)?
3. При представлении упорядоченных во времени событий какая дилемма может легко привести к путанице между самым старшим и самым младшим битами (см. раздел 3.2.3.1)?
4. Ухудшение качества сигнала определяется двумя основными факторами: *снижением* отношения сигнал/шум и *искажением*, приводящим к не поддающейся улучшению вероятности возникновения ошибки. Чем отличаются эти факторы (см. раздел 3.3.2)?
5. Иногда увеличение отношения E_b/N_0 не останавливает ухудшение качества, вызванное *межсимвольной интерференцией*. Когда это происходит (см. раздел 3.3.2)?

6. В какой точке системы определяется отношение E_b/N_0 (см. раздел 4.3.2)?
7. Схемы цифровой модуляции относятся к одному из двух классов с противоположными поведенческими характеристиками: а) ортогональная передача сигналов, б) передача с использованием модуляции фазы/амплитуды. Опишите поведение каждого класса (см. разделы 4.8.2 и 9.7).
8. Почему двоичная фазовая манипуляция (binary phase shift keying — BPSK) и четверичная фазовая манипуляция (quaternary phase shift keying — QPSK) имеют одинаковую вероятность битовой ошибки? Справедливо ли то же самое для M-арной амплитудно-импульсной модуляции (M-ary pulse amplitude modulation — M-PAM) и M^2-арной квадратурной амплитудной модуляции (M^2-ary quadrature amplitude modulation — M^2-QAM) (см. разделы 4.8.4 и 9.8.3.1)?
9. Почему при ортогональной передаче сигналов достоверность передачи растет с увеличением размерности (см. раздел 4.8.5)?
10. Почему *потери в свободном пространстве* — это функция длины волны (см. раздел 5.3.3)?
11. Какая связь существует между отношением сигнал/шум (S/N) в принятом сигнале и отношением мощности несущей к шуму (C/N) (см. раздел 5.4)?
12. Опишите четыре типа компромиссов, которые могут быть достигнуты при использовании кода коррекции ошибок (см. раздел 6.3.4).
13. Почему эффективность традиционных кодов коррекции ошибок снижается при низких значениях E_b/N_0 (см. раздел 6.3.4)?
14. Каково значение *нормальной матрицы* в понимании блочного кода и оценке его возможностей (см. раздел 6.6.5)?
15. Почему при разработке реальных систем не стремятся достигнуть предела Шеннона, равного –1,6 дБ (см. раздел 8.4.5.2)?
16. Что вытекает из того, что алгоритм декодирования Витерби не дает *апостериорных* вероятностей? Какое более характерное название имеет алгоритм Витерби (см. раздел 8.4.6)?
17. Почему связь ширины полосы с эффективностью ее использования одинакова для ортогональных двоичной и четверичной частотных манипуляций (frequency shift keying — FSK) (см. раздел 9.5.1)?
18. Опишите преобразования скрытой энергии и скоростей принимаемых сигналов: при переходе информационных битов в канальные, затем — в символы и элементарные сигналы (см. раздел 9.7.7.)?
19. Дайте определения следующим терминам: *бод*, *состояние*, *ресурс связи*, *элементарный сигнал*, *устойчивый сигнал* (см. разделы 1.1.3 и 7.2.2, главу 11, а также разделы 12.3.2 и 12.4.2).
20. Почему в канале с замираниями дисперсия сигнала не зависит от скорости замирания (см. главу 15)?

Надеюсь, что для вас полезно было таким образом представить проблемы рассматриваемой области. Перейдем теперь к более методичному описанию целей данной книги. В предлагаемом издании я попытался представить системы цифровой связи в доступном виде для старшекурсников, аспирантов и практикующих инженеров. Хотя

основное внимание здесь уделено цифровой связи, все же в этом издании представлены необходимые базовые знания по аналоговым системам (причиной включения такого материала послужило использование аналоговых волн для радиопередачи цифровых сигналов). Особенность систем цифровой связи заключается в том, что они имеют дело с конечным набором дискретных сообщений, тогда как в системах аналоговой связи сообщения определены как непрерывные. Задача приемника цифровой системы — не точное воспроизведение сигнала, а определение, каким из конечного набора сигналов является принятый искаженный сигнал. Для выполнения этого и было разработано впечатляющее множество технологий обработки сигналов.

В данной книге все эти технологии рассматриваются в контексте унифицированной структуры. Эта структура, в форме блочной диаграммы, демонстрируется в начале каждой главы. При необходимости блоки на диаграмме выделяются, чтобы указать на соответствующие цели главы. Основные задачи книги — ввести понятие об организации и структуре отрасли, которая быстро развивается, а также обеспечить осведомленность о "общей картине" (иногда, вдаваясь в подробности). Сигналы и ключевые этапы их обработки прослеживаются, начиная от источника информации через передатчик, канал, приемник и заканчивая, в конечном итоге, ее адресатом. Преобразования сигналов сгруппированы согласно девяти функциональным классам: форматирование и кодирование источника, передача узкополосного сигнала, передача полосового сигнала, выравнивание, канальное кодирование, уплотнение и множественный доступ, расширение спектра, шифрование, синхронизация. В этой книге основное внимание уделяется задачам системы цифровой связи и необходимости альтернатив между основными параметрами системы, такими как отношение сигнал/шум, вероятность ошибки и эффективность использования полосы пропускания.

Структура книги

В главе 1 вводятся основные понятия систем цифровой связи и называются основные преобразования сигналов, которые подробно будут рассмотрены в последующих главах. Даются некоторые основные сведения относительно случайных величин и *аддитивного белого гауссового шума* (additive white Gaussian noise — AWGN). Кроме того, устанавливается связь между спектральной плотностью мощности и автокорреляцией, а также рассматривается передача сигналов через линейные системы. В главе 2 рассмотрен такой этап обработки сигналов, как форматирование; он необходим для формирования информационного сигнала, совместимого с цифровой системой. Глава 3 посвящена вопросам *узкополосной передачи*, обнаружения сигналов в гауссовом шуме и оптимизации приемника. В главе 4 рассмотрена *полосовая передача* и связанные с ней технологии модуляции и демодуляции/обнаружения. В главе 5 дан *анализ канала передачи данных*, позволяющий составить общее представление о системе. В этой главе представлено несколько "тонких" моментов, которые в литературе обычно пропускаются. В главах 6–8 рассмотрено *канальное кодирование* — рентабельный способ реализации разнообразных компромиссов, связанных с производительностью системы. В главе 6 основное внимание уделяется *линейным блочным кодам*, в главе 7 — *сверточным кодам*, а в главе 8 — *кодам Рида-Соломона* и *каскадным кодам*, в частности *турбокодам*.

В главе 9 рассматриваются различные проектные компромиссы при использовании модуляции/кодирования, связанные с вероятностью битовой ошибки, эффективностью использования полосы и отношением сигнал/шум. Освещаются также важные аспекты кодовой модуляции, в частности *решетчатое кодирование*. Глава 10 посвящена *синхрониза-*

ции цифровых систем. В ней рассмотрено использование контура фазовой автоподстройки частоты (ФАПЧ) для синхронизации несущей. Описана также битовая синхронизация, кадровая синхронизация и сетевая синхронизация. Кроме того, вводятся некоторые способы обеспечения синхронизации с использованием цифровых методов.

В главе 11 рассматривается *уплотнение* и *множественный доступ*. Здесь исследуются доступные методы эффективного использования ресурса связи. В главе 12 вводятся методы *расширения спектра* и их применение в таких областях, как множественный доступ, масштабирование и подавление интерференции. Эта технология важна как для военных, так и коммерческих приложений. В главе 13 рассматривается *кодирование источника*, представляющее собой особый класс форматирования данных. И форматирование, и кодирование источника включают оцифровывание данных; основное отличие состоит в том, что кодирование источника дополнительно включает снижение избыточности данных. Несмотря на сходство этих преобразований сигнала, кодирование источника не рассматривается непосредственно после форматирования, оно умышленно представлено в отдельной главе, дабы не прерывать поток представления основных этапов обработки. Глава 14 включает основные идеи *шифрования/дешифрования*. В ней изложены некоторые классические концепции, а также рассмотрен класс систем, известных как системы шифрования с открытым ключом, и широко используемое программное обеспечение для шифровки сообщений электронной почты, называемое *Pretty Good Privacy* (PGP). В главе 15 рассматриваются *каналы с замираниями*. Здесь мы рассмотрим приложения, такие как сотовая радиосвязь, где характеристики канала связи имеют намного более важное значение, чем в незамирающих каналах. Вообще, проектирование систем связи, противостоящих ухудшающему эффекту замирания, может оказаться более перспективным, чем разработка их незамирающих эквивалентов. В данной главе описываются технологии, которые могут снизить эффект замирания, и рассматривается несколько проектов, которые уже были успешно реализованы.

Предполагается, что читатель знаком с методами Фурье-анализа и операцией свертки. Краткий обзор этих методов предлагается в приложении А, где основное внимание обращается на моменты, полезные в теории связи. Также предполагается, что читатель имеет необходимые знания из области теории вероятностей и случайных переменных. В приложении Б на основе этих дисциплин дана краткая трактовка теории принятия статистических решений с акцентом на критериях проверки гипотез — весьма важных для понимания теории обнаружения. В данное издание было добавлено приложение Д, в котором приведен краткий обучающий материал по s-области, z-области и цифровой фильтрации.

При использовании данной книги для двусеместрового курса, предлагается первые семь глав представить в первом семестре, а следующие восемь — во втором. При чтении семестрового вводного курса предлагается выбрать материал из следующих глав: 1–7, 9, 10, 12.

Благодарности

Написать техническую книгу без чьей-либо помощи чрезвычайно трудно. Я весьма признателен всем, кто помог мне в создании данной книги. За содействие в работе я благодарю д-ра Эндрю Витерби (Andrew Viterbi), д-ра Чака Уитли (Chuck Wheatley), д-ра Эда Тайдмэна (Ed Tiedeman), д-ра Джо Оденуолдера (Joe Odenwalder) и Сержа Уиллинеггера (Serge Willinegger) из Qualcomm. Также хочу поблагодарить д-ра Дариуша Дивсалара (Dariush Divsalar) из Jet Propulsion Laboratory (JPL), д-ра Боба Богуша

(Bob Bogusch) из Mission Research, д-ра Тома Стэнли (Tom Stanley) из Federal Communication Commission, профессора Ларри Милстейна (Larry Milstein) из University of California, San Diego, профессора Рея Пикхольца (Ray Pickholtz) из Gerge Washington University, профессора Даниеля Костелло (Daniel Costello) из Notre Dame University, профессора Теда Раппапорта (Ted Rappaport) из Virginia Polytechnic Institute, Фила Коссина (Phil Kossin) из Lincom, Леса Брауна (Les Brown) из Motorola, а также д-ра Боба Прайса (Bob Price) и Франка Аморосо (Frank Amoroso).

Мне также хотелось бы поблагодарить людей, которые помогли мне с выпуском первого издания данной книги. Это — д-р Морис Кинг (Maurice King), Дон Мартин (Don Martin) и Нэд Фельдман (Ned Feldman) из The Aerospace Corporation, д-р Марв Симон (Marv Simon) из JPL, д-р Билл Линдсей (Bill Lindsey) из Lincom, профессор Вейн Старк (Wayne Stark) из University of Michigan, а также д-р Джим Омура (Jim Omura), д-р Адам Лендер (Adam Lender) и д-р Тодд Цитрон (Todd Citron).

Хотелось бы выразить признательность доктору Морис Кинг (Maurice King) за вклад в главу 10, посвященную синхронизации, и профессору Фреду Харрису (Fred Harris) из San Diego University за написание главы 13, посвященной кодированию источника. Спасибо также Мишель Ландри (Michelle Landry) за создание разделов по Pretty Good Privacy в главе 14 и Эндрю Гвиди (Andrew Guidi) за вклад в задачи главы 15.

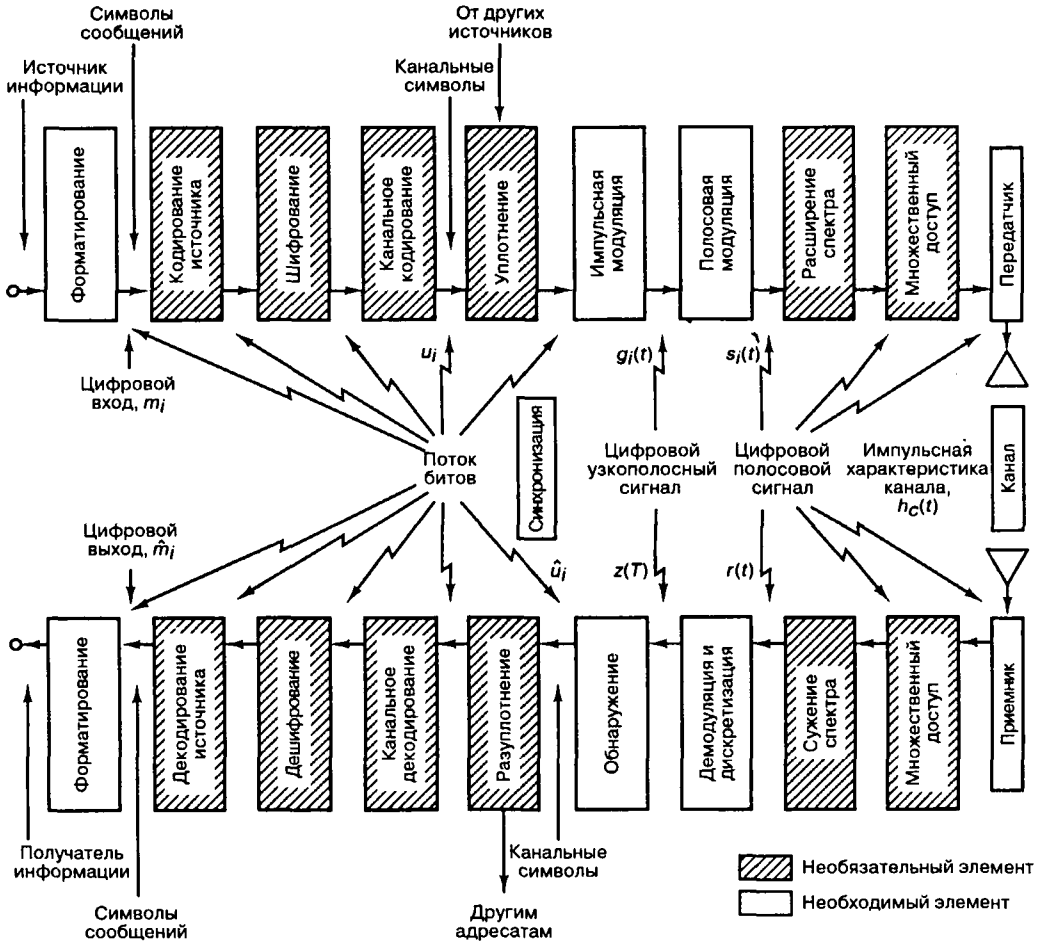
Я в неоплатном долгу перед моими друзьями и коллегами Фрэдом Харрисом (Fred Harris), профессором Дэном Буковцером (Dan Bukofzer) из California State University в Fresno и д-ром Маури Шифф (Maury Schiff) из Elanix, которые терпеливо выслушивали меня всякий раз, когда я к ним обращался. Также хочу поблагодарить моих лучших учителей — моих студентов из University of California (Los Angeles), а также всех студентов, которые уделили внимание моим кратким курсам. Их вопросы направляли меня и побудили написать данное (второе) издание. Надеюсь, что я сумел доходчиво ответить на все их вопросы.

Отдельно хотел бы поблагодарить моего сына, Дина Скляра (Dean Sklar), за технические предложения; он взял на себя роль главного критика работы своего отца и "адвоката дьявола". Я многим обязан профессору Бобу Стюарту (Bob Stewart) из University of Strathclyde, который провел бесчисленные часы за написанием и подготовкой компакт-диска и разработкой приложения Д. Я благодарен Роуз Кернан (Rose Kernan), моему редактору, за помощь в создании проекта и Бернарду Гудвину (Bernard Goodwin), издателю из Prentice Hall, за снисходительное отношение ко мне и веру в меня. Его рекомендации были бесценными. Я чрезвычайно благодарен моей жене, Гвен (Gwen), за ее одобрение, преданность и ценные советы. Она хранила меня от "стрел и камней" повседневной жизни, что дало мне возможность закончить данное издание.

Bernard Sklar
Tarzana, California

ГЛАВА 1

Сигналы и спектры



В данной книге излагаются идеи и технологии, являющиеся фундаментальными для систем цифровой связи. Основное внимание обращается на вопросы проектирования систем и необходимость компромиссов между основными параметрами системы, такими как отношение сигнал/шум (signal-to-noise ratio — SNR), вероятность появления ошибки и эффективность использования полосы. Мы рассмотрим передачу информации (речь, видео или данные) по каналу связи, где средой передачи является проводник, волновод или окружающая среда.

Системы цифровой связи становятся все более привлекательными вследствие постоянно растущего спроса и из-за того, что цифровая передача предлагает возможности обработки информации, не доступные при использовании аналоговой передачи. В данной книге цифровые системы часто рассматриваются в контексте спутникового канала связи. Иногда это трактуется в контексте систем мобильной радиосвязи, в этом случае передача сигнала обычно ухудшается вследствие явления, называемого *замиранием*. Здесь стоит отметить, что спроектировать и описать систему связи, противостоящую замиранию, сложнее, чем выполнить то же для системы без замирания.

Отличительной особенностью систем цифровой связи (digital communication system — DCS) является то, что за конечный промежуток времени они посылают сигнал, состоящий из конечного набора элементарных сигналов (в отличие от систем аналоговой связи, где сигнал состоит из бесконечного множества элементарных сигналов). В системах DCS задачей приемника является *не* точное воспроизведение переданного сигнала, а определение на основе искаженного шумами сигнала, какой именно сигнал из конечного набора был послан передатчиком. Важным критерием производительности системы DCS является вероятность ошибки (P_E).

1.1. Обработка сигналов в цифровой связи

1.1.1. Почему "цифровая"

Почему в военных и коммерческих системах связи используются "цифры"? Существует множество причин. Основным преимуществом такого подхода является легкость восстановления цифровых сигналов по сравнению с аналоговыми. Рассмотрим рис. 1.1, на котором представлен идеальный двоичный цифровой импульс, распространяющийся по каналу передачи данных. На форму сигнала влияют два основных механизма: (1) поскольку все каналы и линии передачи имеют неидеальную частотную характеристику, идеальный импульс искажается; и (2) нежелательные электрические шумы или другое воздействие со стороны еще больше искажает форму импульса. Чем протяженнее канал, тем существеннее эти механизмы искажают импульс (рис. 1.1). В тот момент, когда переданный импульс все еще может быть достоверно определен (прежде чем он ухудшится до неоднозначного состояния), импульс усиливается цифровым усилителем, восстанавливающим его первоначальную идеальную форму. Импульс "возрождается" или восстанавливается. За восстановление сигнала отвечают *регенеративные ретрансляторы*, расположенные в канале связи на определенном расстоянии друг от друга.

Цифровые каналы менее подвержены искажению и интерференции, чем аналоговые. Поскольку двоичные цифровые каналы дают значимый сигнал только при работе в одном из двух состояний — включенном или выключенном — возмущение должно быть достаточно большим, чтобы перевести операционную точку канала из одного со-

стояния в другое. Наличие всего двух состояний облегчает восстановление сигнала и, следовательно, предотвращает накопление в процессе передачи шумов или других возмущений. Аналоговые сигналы, наоборот, *не являются* сигналами с двумя состояниями; они могут принимать *бесконечное множество* форм. В аналоговых каналах даже небольшое возмущение может неузнаваемо исказить сигнал. После искажения аналогового сигнала возмущение нельзя убрать путем усиления. Поскольку накопление шума неразрывно связано с аналоговыми сигналами, как следствие, они не могут воспроизводиться идеально. При использовании цифровых технологий очень низкая частота возникновения ошибок плюс применение процедур выявления и коррекции ошибок делают возможным высокую точность сигнала. Остается только отметить, что с аналоговыми технологиями подобные процедуры недоступны.

Рис. 1.1. Искажение и восстановление импульса

Существуют и другие важные преимущества цифровой связи. Цифровые каналы *надежнее* и могут производиться по более низким ценам, чем аналоговые. Кроме того, цифровое программное обеспечение позволяет *более гибкую* реализацию, чем аналоговое (например, микропроцессоры, цифровая коммутация и большие интегральные схемы (large-scale integrated circuit — LSI)). Использование цифровых сигналов и уплотнения с временным разделением (time-division multiplexing — TDM) *проще* применения аналоговых сигналов и уплотнения с частотным разделением (frequency-division multiplexing — FDM). При передаче и коммутации различные типы цифровых сигналов (данные, телеграф, телефон, телевидение) могут рассматриваться как идентичные: *ведь бит — это и есть бит.* Кроме того, для удобства коммутации и обработки, цифровые сообщения могут группироваться в автономные единицы, называемые *пакетами.* В цифровые технологии естественным образом внедряются функции, защищающие от интерференции и подавления сигнала либо обеспечивающие шифрование или секретность. (Подобные технологии рассматриваются в главах 12 и 14.) Кроме того, обмен данными в основном производится между двумя компьютерами или между компьютером и цифровыми устройствами или терминалом. Подобные цифровые оконечные устройства лучше (и естественнее!) обслуживаются цифровыми каналами связи.

Чем же мы платим за преимущества систем цифровой связи? Цифровые системы требуют более интенсивной обработки, чем аналоговые. Кроме того, для цифровых систем необходимо выделение значительной части ресурсов для синхронизации на различных уровнях (см. главу 10). Аналоговые системы, наоборот, легче синхронизировать. Еще одним недостатком систем цифровой связи является то, что *ухудше-*

ние качества носит пороговый характер. Если отношение сигнал/шум падает ниже некоторого порога, качество обслуживания может внезапно измениться от очень хорошего до очень плохого. В аналоговых же системах ухудшение качества происходит более плавно.

1.1.2. Типичная блочная диаграмма и основные преобразования

Функциональная блочная диаграмма, приведенная на рис. 1.2, иллюстрирует распространение сигнала и этапы его обработки в типичной системе цифровой связи (DCS). Этот рисунок является чем-то вроде плана, направляющего читателя по главам данной книги. Верхние блоки — форматирование, кодирование источника, шифрование, канальное кодирование, уплотнение, импульсная модуляция, полосовая модуляция, расширение спектра и множественный доступ — отражают преобразования сигнала на пути от источника к передатчику. Нижние блоки диаграммы — преобразования сигнала на пути от приемника к получателю информации, и, по сути, они противоположны верхним блокам. Блоки *модуляции* и *демодуляции/обнаружения* вместе называются *модемом.* Термин "модем" часто объединяет несколько этапов обработки сигналов, показанных на рис. 1.2; в этом случае модем можно представлять как "мозг" системы. Передатчик и приемник можно рассматривать как "мускулы" системы. Для беспроводных приложений передатчик состоит из схемы повышения частоты в область радиочастот (radio frequency — RF), усилителя мощности и антенны, а приемник — из антенны и малошумящего усилителя (low-noise amplifier — LNA). Обратное понижение частоты производится на выходе приемника и/или демодулятора.

На рис. 1.2 иллюстрируется соответствие блоков верхней (передающей) и нижней (принимающей) частей системы. Этапы обработки сигнала, имеющие место в передатчике, являются преимущественно обратными к этапам приемника. На рис. 1.2 исходная информация преобразуется в двоичные цифры (*биты*); после этого биты группируются в *цифровые сообщения* или *символы сообщений.* Каждый такой символ (m_i, где $i = 1, ..., M$) можно рассматривать как элемент *конечного алфавита*, содержащего M элементов. Следовательно, для $M = 2$ символ сообщения m_i является бинарным (т.е. состоит из одного бита). Несмотря на то что бинарные символы можно классифицировать как M-арные (с $M = 2$), обычно названис "M-арный" используется для случаев $M > 2$; значит, такие символы состоят из последовательности двух или большего числа битов. (Сравните подобный конечный алфавит систем DCS с тем, что мы имеем в аналоговых системах, когда сигнал сообщения является элементом бесконечного множества возможных сигналов.) Для систем, использующих *канальное кодирование* (коды коррекции ошибок), последовательность символов сообщений преобразуется в последовательность *канальных символов* (кодовых символов), и каждый канальный символ обозначается u_i. Поскольку символы сообщений или канальные символы могут состоять из одного бита или группы битов, последовательность подобных символов называется *потоком битов* (рис. 1.2).

Рассмотрим ключевые блоки обработки сигналов, изображенные на рис. 1.2; необходимыми для систем DCS являются только этапы форматирования, модуляции, демодуляции/обнаружения и синхронизации.

Форматирование преобразовывает исходную информацию в биты, обеспечивая, таким образом, совместимость информации и функций обработки сигналов с системой DCS. С этой точки рисунка и вплоть до блока импульсной модуляции информация остается в форме *потока битов.*

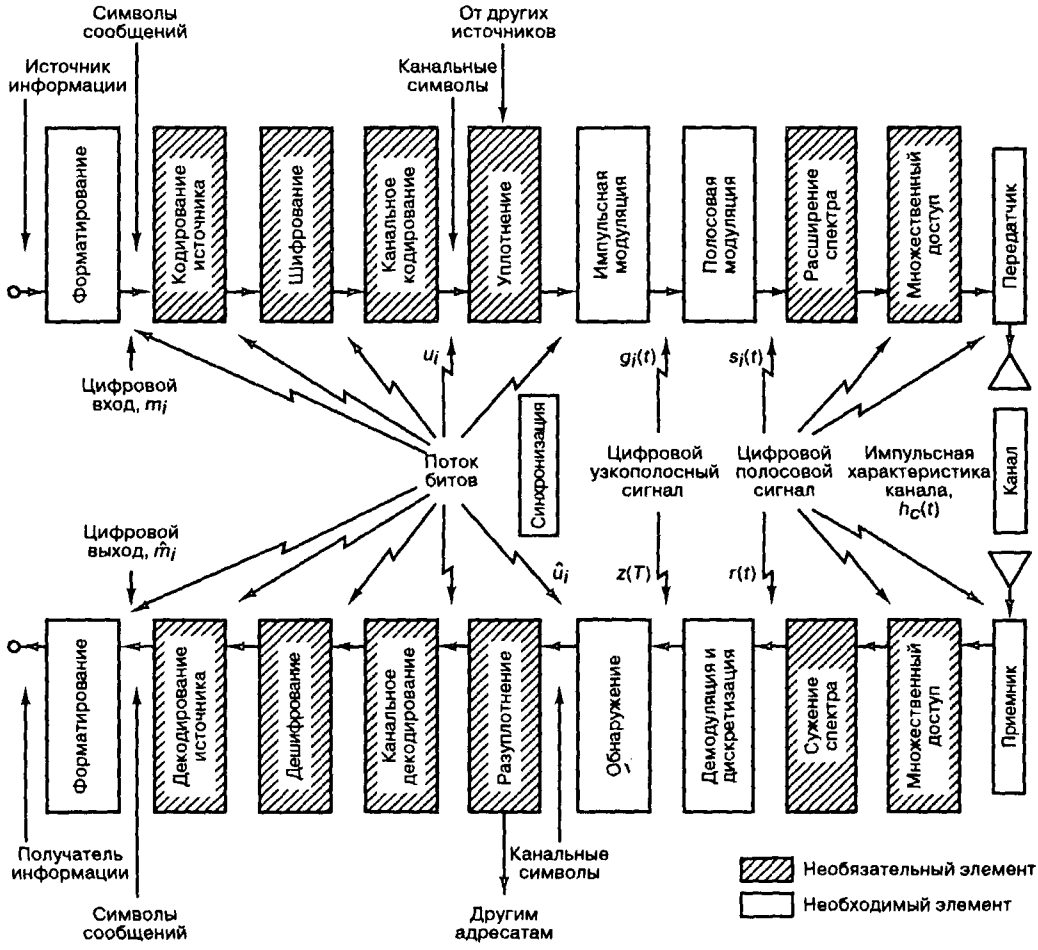


Рис. 1.2. Блочная диаграмма типичной системы цифровой связи

Модуляция — это процесс, посредством которого символы сообщений или канальные символы (если используется канальное кодирование) преобразуются в *сигналы*, совместимые с требованиями, налагаемыми каналом передачи данных. *Импульсная модуляция* — это еще один необходимый этап, поскольку каждый символ, который требуется передать, вначале нужно преобразовать из двоичного представления (уровни напряжений представляют двоичные нули и единицы) в форму *узкополосного* сигнала. Термин "узкополосный" (baseband) определяет сигнал, спектр которого начинается от (или около) постоянной составляющей и заканчивается некоторым конечным значением (обычно, не более нескольких мегагерц). Блок импульсно-кодовой модуляции обычно включает фильтрацию, направленную на минимизацию полосы передачи. При применении импульсной модуляции к двоичным символам результирующий двоичный сигнал называется сигналом в кодировке PCM (pulse-code modulation — импульсно-кодовая модуляция). Существует несколько типов сигналов PCM (описанных в главе 2); в приложениях телефонной связи эти сигналы часто называются *кодами канала*. При применении импульсной модуляции к небинарным симво-

лам результирующий сигнал именуется M-арным импульсно-модулированным. Существует несколько типов подобных сигналов, которые также описаны в главе 2, где основное внимание уделяется *амплитудно-импульсной модуляции* (pulse-amplitude modulation — PAM). После импульсной модуляции каждый символ сообщения или канальный символ принимает форму полосового сигнала $g_i(t)$, где $i = 1, ..., M$. В любой электронной реализации поток битов, предшествующий импульсной модуляции, представляется уровнями напряжений. Может возникнуть вопрос, почему существует отдельный блок для импульсной модуляции, когда фактически уровни напряжения для двоичных нулей и единиц уже можно рассматривать как идеальные прямоугольные импульсы, длительность каждого из которых равна времени передачи одного бита? Существует два важных отличия между подобными уровнями напряжения и полосовыми сигналами, используемыми для модуляции. Во-первых, блок импульсной модуляции позволяет использовать бинарные и M-арные сигналы. В разделе 2.8.2 описаны различные полезные параметры этих типов сигналов. Во-вторых, фильтрация, производимая в блоке импульсной модуляции, формирует импульсы, длительность которых больше времени передачи одного бита. Фильтрация позволяет использовать импульсы большей длительности; таким образом, импульсы расширяются на соседние временные интервалы передачи битов. Этот процесс иногда называется формированием импульсов; он используется для поддержания полосы передачи в пределах некоторой желаемой области спектра.

Для приложений, включающих передачу в диапазоне радиочастот, следующим важным этапом является *полосовая модуляция* (bandpass modulation); она необходима всегда, когда среда передачи не поддерживает распространение сигналов, имеющих форму импульсов. В таких случаях среда требует полосового сигнала $s_i(t)$, где $i = 1, ..., M$. Термин "полосовой" (bandpass) используется для отражения того, что узкополосный сигнал $g_i(t)$ сдвинут несущей волной на частоту, гораздо большую спектральных составляющих $g_i(t)$. По мере распространения сигнала $s_i(t)$ по каналу, на него воздействуют характеристики канала, которые можно выразить через *импульсную характеристику* $h_c(t)$ (см. раздел 1.6.1). Кроме того, в различных точках вдоль маршрута сигнала дополнительные случайные шумы искажают принятый сигнал $r(t)$, поэтому прием должен выражаться через поврежденную версию сигнала $s_i(t)$, поступающего от передатчика. Принятый сигнал $r(t)$ можно выразить следующим образом:

$$r(t) = s_i(t) * h_c(t) + n(t) \quad i = 1, ..., M, \tag{1.1}$$

где знак "*" представляет собой операцию свертки (см. приложение А), а $n(t)$ — процесс шума (см. раздел 1.5.5).

В обратном направлении входной каскад приемника и/или демодулятор обеспечивают понижение частоты каждого полосового сигнала $r(t)$. В качестве подготовки к обнаружению демодулятор восстанавливает $r(t)$ в виде оптимального огибающего узкополосного сигнала $z(t)$. Обычно с приемником и демодулятором связано несколько фильтров — фильтрование производится для удаления нежелательных высокочастотных составляющих (в процессе преобразования полосового сигнала в узкополосный) и формирования импульса. Выравнивание можно описать как разновидность фильтрации, используемой в демодуляторе (или после демодулятора) для удаления всех эффектов ухудшения качества сигнала, причиной которых мог быть канал. Выравнивание (equalization) необходимо в том случае, если импульсная характеристика канала

$h_c(t)$ настолько плоха, что принимаемый сигнал сильно искажен. Эквалайзер (устройство выравнивания) реализуется для компенсации (т.е. для удаления или ослабления) всех искажений сигнала, вызванных неидеальной характеристикой $h_c(t)$. И последнее, этап дискретизации преобразовывает сформированный импульс $z(t)$ в выборку $z(T)$ для восстановления (приблизительно) символа канала $\hat{u}_i$ или символа сообщения m_i (если не используется канальное кодирование). Некоторые авторы используют термины "демодуляция" и "обнаружение" как синонимы. В данной книге под *демодуляцией* (demodulation) подразумевается восстановление сигнала (полосового импульса), а под *обнаружением* (detection) — принятие решения относительно цифрового значения этого сигнала.

Остальные этапы обработки сигнала в модеме являются необязательными и направлены на удовлетворение специфических системных нужд. *Кодирование источника* (source coding) — это преобразование аналогового сигнала в цифровой (для аналоговых источников) *и* удаление избыточной (ненужной) информации. Отметим, что типичная система DCS может использовать либо *кодирование источника* (для оцифровывания и сжатия исходной информации), либо более простое преобразование *форматирование* (только для оцифровывания). Система не может одновременно применять и кодирование источника, и форматирование, поскольку первое уже включает необходимый этап оцифровывания информации. Шифрование, которое используется для обеспечения секретности связи, предотвращает понимание сообщения несанкционированным пользователем и введение в систему ложных сообщений. *Канальное кодирование* (channel coding) при данной скорости передачи данных может снизить вероятность ошибки P_E или уменьшить отношение сигнал/шум, необходимое для получения желаемой вероятности P_E за счет увеличения полосы передачи или усложнения декодера. Процедуры *уплотнения* (multiplexing) и *множественного доступа* (multiple access) объединяют сигналы, которые могут иметь различные характеристики или могут поступать от разных источников, с тем, чтобы они могли совместно использовать часть ресурсов связи (например, спектр, время). *Расширение частоты* (frequency spreading) может давать сигнал, относительно неуязвимый для интерференции (как естественной, так и умышленной), и может использоваться для повышения конфиденциальности сообщающихся сторон. Также оно является ценной технологией, используемой для множественно доступа.

Блоки обработки сигналов, показанные на рис. 1.2, представляют типичную схему системы цифровой связи; впрочем, эти блоки иногда реализуются в несколько ином порядке. Например, уплотнение может происходить до канального кодирования *или* модуляции *либо* — при двухэтапном процессе модуляции (поднесущая и несущая) — оно может выполняться между двумя этапами модуляции. Подобным образом блок расширения частоты может находиться в различных местах верхнего ряда рис. 1.2; точное его местонахождение зависит от конкретной используемой технологии. Синхронизация и ее ключевой элемент, синхронизирующий сигнал, задействованы во всех этапах обработки сигнала в системе DCS. Для простоты блок синхронизации на рис. 1.2 показан безотносительно к чему-либо, хотя фактически он участвует в регулировании операций практически в каждом блоке, приведенном на рисунке.

На рис. 1.3 показаны основные функции обработки сигналов (которые можно рассматривать как преобразования сигнала), разбитые на следующие девять групп.

Рис. 1.3. Основные преобразования цифровой связи

1. Форматирование и кодирование источника
2. Узкополосная передача сигналов
3. Полосовая передача сигналов
4. Выравнивание
5. Канальное кодирование
6. Уплотнение и множественный доступ
7. Расширение спектра
8. Шифрование
9. Синхронизация

Хотя пункты такого разделения частично перекрываются, все же это позволяет удобно упорядочить материал книги. Начиная с главы 2 поочередно рассматриваются все девять основных преобразований. В главе 2 исследуются основные методы форматирования, используемые для преобразования исходной информации в символы сообщений. Кроме того, здесь описывается выбор узкополосного сигнала и фильтрация импульса, обеспечивающая совместимость символов сообщений с узкополосной передачей. Обратные этапы демодуляции, выравнивания, дискретизации и обнаружения представлены в главе 3. Форматирование и кодирование источника являются подобными процессами, поскольку оба включают оцифровывание данных. Впрочем, термин "кодирование источника" подразумевает дополнительное сжатие данных и рассматривается позднее (глава 13) как частный случай форматирования.

На рис. 1.3 блок *Узкополосная передача сигналов* содержит перечень бинарных альтернатив при использовании модуляции PCM или линейных кодов. В этом блоке также указана небинарная категория сигналов, называемая *M*-арной импульсной модуляцией. Еще одно преобразование на рис. 1.3, помеченное как *Полосовая передача сигналов*, разделено на два основных блока, когерентный и некогерентный. Демодуляция обычно выполняется с помощью *опорных* сигналов. При использовании известных сигналов в качестве меры всех параметров сигнала (особенно фазы) процесс демодуляции называется *когерентным*; когда информация о фазе не используется, процесс именуется *некогерентным*. Обе технологии подробно описаны в главе 4.

Глава 5 посвящена *анализу канала связи*. Среди множества спецификаций, анализов и табличных представлений, поддерживающих разработку систем связи, анализ канала связи занимает особое место, поскольку позволяет представить общую картину системы. В главе 5 воедино сводятся все основные понятия канала связи, необходимые для анализа большинства систем связи.

Канальное кодирование связано с методами, используемыми для улучшения цифровых сигналов, которые в результате становятся менее уязвимыми к таким факторам ухудшения качества, как шум, замирание и подавление сигнала. На рис. 1.3 канальное кодирование разделено на два блока, блок кодирования формой сигнала и блок структурированных последовательностей. *Кодирование формой сигнала* включает использование новых сигналов, привносящих улучшенное качество обнаружения по сравнению с исходным сигналом. Структурированные последовательности включают применение дополнительных битов для определения наличия ошибки, вызванной шумом в канале. Одна из таких технологий, *автоматический запрос повторной передачи* (automatic repeat request — ARQ), просто распознает появление ошибки и запрашивает отправителя повторно передать сообщение; другая технология, известная как *прямая коррекция ошибок*

(forward error correction — FEC), позволяет автоматически исправлять ошибки (с определенными ограничениями). При рассмотрении структурированных последовательностей мы обсудим три распространенных метода — блочное, сверточное и турбокодирование. Вначале в главе 6 описывается *линейное блочное кодирование*. В главе 7 мы рассмотрим *сверточное кодирование*, декодирование Витерби (и другие алгоритмы декодирования) и сравним аппаратные и программные процедуры кодирования. В главе 8 представлено каскадное кодирование, которое привело к созданию класса кодов, известных как *турбокоды*, а также подробно рассмотрены коды *Рида-Соломона*.

В главе 9 обобщаются вопросы проектирования систем связи и представляются различные компромиссы из областей модуляции и кодировки, которые обязательно должны быть рассмотрены при проектировании системы. Обсуждаются теоретические ограничения, такие как критерий Найквиста и предел Шеннона. Также исследуются схемы модуляции, позволяющие эффективно использовать полосу, такие как решетчатое кодирование.

Глава 10 посвящена *синхронизации*. В цифровой связи синхронизация включает вычисление как времени, так и частоты. Как показано на рис. 1.3, синхронизация выполняется на пяти уровнях. Эталонные частоты когерентных систем требуется синхронизировать с несущей (и возможно, поднесущей) по частоте и фазе. Для некогерентных систем синхронизация фазы не обязательна. Основной процесс синхронизации по времени — это символьная синхронизация (или битовая синхронизация для бинарных символов). Демодулятор и детектор должны знать, когда начинать и заканчивать процесс обнаружения символа и бита; ошибка синхронизации приводит к снижению эффективности обнаружения. Следующий уровень синхронизации по времени, кадровая синхронизация, позволяет перестраивать сообщения. И последний уровень, сетевая синхронизация, позволяет скоординировать действия с другими пользователями с целью эффективного использования ресурсов. В главе 10 мы рассмотрим синхронизацию пространственно разделенных периодических процессов.

В главе 11 описаны *уплотнение* и *множественный доступ*. Значения этих двух терминов очень похожи; оба связаны с идеей совместного использования ресурсов. Основным отличием является то, что уплотнение реализуется локально (например, на печатной плате, в компоновочном узле или даже на аппаратном уровне), а множественный доступ — удаленно (например, нескольким пользователям требуется совместно использовать спутниковый транспондер). При уплотнении применяется алгоритм, известный априорно; обычно он внедрен непосредственно в систему. Множественный доступ, наоборот, обычно адаптивен и может требовать для работы некоторых служебных издержек. В главе 11 мы рассмотрим классические способы совместного использования ресурсов связи: частотное, временное и кодовое разделение. Кроме того, будут описаны некоторые технологии множественного доступа, возникшие в результате использования спутниковой связи.

В главе 12 вводится преобразование, изначально разработанное для военной связи и известное как *расширение* (spreading). Здесь рассмотрены методы расширения спектра, важные для получения защиты от интерференции и обеспечения секретности. Сигналы могут расширяться по частоте, времени или по частоте и времени. В основном в главе обсуждается расширение частоты. Также глава иллюстрирует применение методом расширения частоты для совместного использования ресурсов с ограниченной полосой в коммерческой переносной телефонии.

В главе 13 рассматривается *кодирование источника*, которое включает эффективное описание исходной информации. Оно связано с процессом компактного описания

сигнала согласно заданным критериям точности. Кодирование источника может применяться и к цифровым, и аналоговым сигналам; путем уменьшения избыточности информации коды источника могут снизить системную скорость передачи данных. Следовательно, основным преимуществом кодирования источника является возможность уменьшения объема требуемых ресурсов системы (например, ширины полосы).

Глава 14 посвящена *шифрованию* и *дешифрованию*, основными задачами которых является аутентификация и обеспечение конфиденциальности связи. Поддержание конфиденциальности означает предотвращение извлечения информации из канала несанкционированными лицами ("подслушивание"). Аутентификация подразумевает предотвращение ввода в канал ложных сигналов несанкционированными лицами. В этой главе значительное внимание уделяется стандарту шифрования данных (data encryption standard — DES) и основным идеям, относящимся к классу систем шифрования, называемых *системы с открытым ключом*. Кроме того, здесь рассмотрена новая схема, названная Pretty Good Privacy ("достаточно хорошая секретность"), которая позволяет эффективно шифровать файлы, предназначенные для отправки по электронной почте.

В последней главе 15 рассмотрены каналы с замираниями. Здесь мы обсудим замирание, которое воздействует на мобильные системы, такие как переносные и персональные системы связи (personal communication system — PCS). В главе перечисляются основные механизмы замирания, типы ухудшения качества и методы борьбы с этим ухудшением. Подробно исследуются два метода: эквалайзер Витерби, реализованный в системе GSM (Global Systems for Mobile Communication — глобальная система мобильной связи), и приемник Рейка, используемый в системах CDMA (Code Division Multiple Access — множественный доступ с кодовым разделением каналов).

1.1.3. Основная терминология области цифровой связи

Ниже приведены некоторые основные термины, часто используемые в области цифровой связи.

Источник информации (information source). Устройство, передающее информацию посредством системы DCS. Источник информации может быть *аналоговым* или *дискретным*. Выход аналогового источника может принимать любое значение из непрерывного диапазона амплитуд, тогда как выход дискретного источника информации — значения из конечного множества амплитуд. Аналоговые источники информации преобразуются в цифровые посредством *дискретизации* или *квантования*. Методы дискретизации и квантования, называемые форматированием и кодированием источника (рис. 1.3), описаны в главах 2 и 13.

Текстовое сообщение (textual message). Последовательность символов (рис. 1.4, *а*). При цифровой передаче данных сообщение представляет собой последовательность цифр или символов, принадлежащих конечному набору символов или алфавиту.

Знак (Character). Элемент алфавита или набора символов (рис. 1.4, *б*). Знаки могут отображаться в последовательность двоичных цифр. Существует несколько стандартизованных кодов, используемых для знакового кодирования, в том числе код ASCII (American Standard Code for Information Interchange — Американский стандартный код для обмена информацией), код EBCDIC (Extended Binary Coded Decimal Interchange Code — расширенный двоичный код обмена инфор-

мацией), код Холлерита (Hollerith code), код Бодо (Baudot code), код Муррея (Murray code) и код (азбука) Морзе (Morse code).

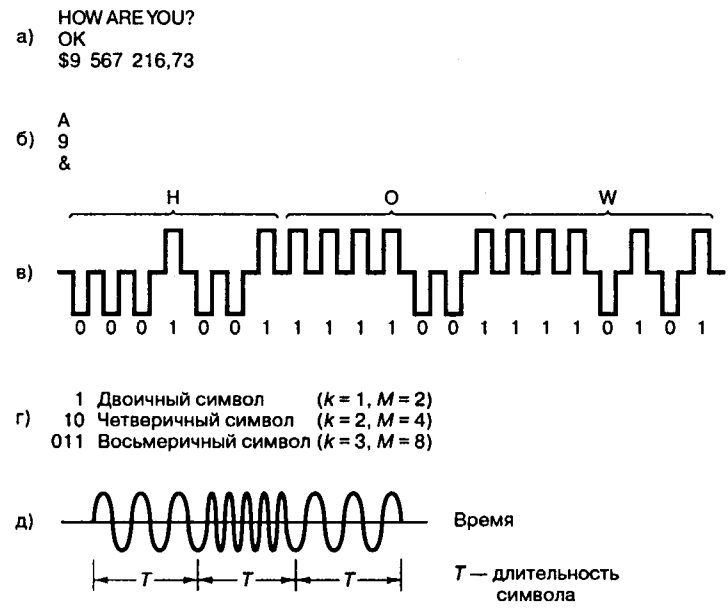


Рис. 1.4. Иллюстрация терминов: а) текстовые сообщения; б) символы; в) поток битов (7-битовый код ASCII); г) символы m_i, $i = 1, \ldots, M$, $M = 2^k$; д) полосовой цифровой сигнал $s_i(t)$, $i = 1, \ldots, M$

Двоичная цифра (binary digit) (*бит*) (bit). Фундаментальная единица информации для всех цифровых систем. Термин "бит" также используется как единица объема информации, что описывается в главе 9.

Поток битов (bit stream). Последовательность двоичных цифр (нулей и единиц). Поток битов часто называют *узкополосным* (baseband) сигналом; это подразумевает, что его спектральные составляющие размещены от (или около) постоянной составляющей до некоторого конечного значения, обычно не превышающего несколько мегагерц. На рис. 1.4, *в* сообщение "HOW" представлено с использованием семибитового кода ASCII, а поток битов показан в форме двухуровневых импульсов. Последовательность импульсов изображена посредством крайне стилизованных (идеально прямоугольных) сигналов с промежутками между соседними импульсами. В реальной системе импульсы никогда не будут выглядеть так, поскольку подобные промежутки абсолютно бесполезны. При данной скорости передачи данных промежутки увеличат ширину полосы, необходимую для передачи; или, при данной ширине полосы, они увеличат временную задержку, необходимую для получения сообщения.

Символ (symbol) (*цифровое сообщение*) (digital message). Символ — это группа из k бит, рассматриваемых как единое целое. Далее мы будем называть этот блок *символом сообщения* (message symbol) m_i ($i = 1, \ldots, M$) из конечного набора символов или алфавита (рис. 1.4, *г*.) Размер алфавита M равен 2^k, где k — число битов в символе. При *узкополосной* передаче каждый из символов m_i будет

представлен одним из набора узкополосных импульсных сигналов $g_1(t)$, $g_2(t)$, ..., $g_M(t)$. Иногда при передаче последовательности таких импульсов для выражения скорости передачи импульсов (скорости передачи символов) используется единица *бод* (baud). Для типичной *полосовой* (bandpass) передачи каждый импульс $g_i(t)$ будет представляться одним из набора полосовых импульсных сигналов $s_1(t)$, $s_2(t)$, ..., $s_M(t)$. Таким образом, для беспроводных систем символ m_i посылается путем передачи цифрового сигнала $s_i(t)$ в течение T секунд. Следующий символ посылается в течение следующего временного интервала, T. То, что набор символов, передаваемых системой DCS, является конечным, и есть главное отличие этих систем от систем аналоговой связи. Приемник DCS должен всего лишь определить, какой из M возможных сигналов был передан; тогда как аналоговый приемник должен точно определять значение, принадлежащее непрерывному диапазону сигналов.

Цифровой сигнал (digital waveform). Описываемый уровнем напряжения или тока, сигнал (импульс — для узкополосной передачи или синусоида — для полосовой передачи), представляющий цифровой символ. Характеристики сигнала (для импульсов — амплитуда, длительность и расположение или для синусоиды — амплитуда, частота и фаза) позволяют его идентифицировать как один из символов конечного алфавита. На рис. 1.4, *д* приведен пример полосового цифрового сигнала. Хотя сигнал является синусоидальным и, следовательно, имеет аналоговый вид, все же он именуется *цифровым*, поскольку кодирует цифровую информацию. На данном рисунке цифровое значение указывается посредством передачи в течение каждого интервала времени T сигнала определенной частоты.

Скорость передачи данных (data rate). Эта величина в битах в секунду (бит/с) дается формулой $R = k/T = (1/T)\log_2 M$ (бит/с), где k бит определяют символ из $M = 2^k$-символьного алфавита, а T — это длительность k-битового символа.

1.1.4. Цифровые и аналоговые критерии производительности

Принципиальное отличие систем аналоговой и цифровой связи связано со способом оценки их производительности. Сигналы аналоговых систем составляют континуум, так что приемник должен работать с бесконечным числом возможных сигналов. Критерием производительности аналоговых систем связи является точность, например отношение сигнал/шум, процент искажения или ожидаемая среднеквадратическая ошибка между переданным и принятым сигналами.

В отличие от аналоговых, цифровые системы связи передают сигналы, представляющие цифры. Эти цифры формируют конечный набор или алфавит, и этот набор известен приемнику априорно. Критерием качества цифровых систем связи является вероятность неверного обнаружения цифры или вероятность ошибки (P_E).

1.2. Классификация сигналов

1.2.1. Детерминированные и случайные сигналы

Сигнал можно классифицировать как *детерминированный* (при отсутствии неопределенности относительно его значения в любой момент времени) или *случайный*, в про-

тивном случае. Детерминированные сигналы моделируются математическим выражением $x(t) = 5\cos 10t$. Для случайного сигнала такое выражение написать *невозможно.* Впрочем, при наблюдении случайного сигнала (также называемого *случайным процессом*) в течение достаточно длительного периода времени, могут отмечаться некоторые закономерности, которые можно описать через вероятности и среднее статистическое. Такая модель, в форме вероятностного описания случайного процесса, особенно полезна для описания характеристик сигналов и шумов в системах связи.

1.2.2. Периодические и непериодические сигналы

Сигнал $x(t)$ называется *периодическим во времени*, если существует постоянное $T_0 > 0$, такое, что

$$x(t) = x(t + T_0) \quad \text{для } -\infty < t < \infty, \tag{1.2}$$

где через t обозначено время. Наименьшее значение T_0, удовлетворяющее это условие, называется *периодом* сигнала $x(t)$. Период T_0 определяет длительность одного полного цикла функции $x(t)$. Сигнал, для которого не существует значения T_0, удовлетворяющего уравнение (1.2), именуется *непериодическим.*

1.2.3. Аналоговые и дискретные сигналы

Аналоговый сигнал $x(t)$ является непрерывной функцией времени, т.е. $x(t)$ однозначно определяется для всех t. Электрический аналоговый сигнал возникает тогда, когда физический сигнал (например, речь) некоторым устройством преобразовывается в электрический. Для сравнения, *дискретный сигнал* $x(kT)$ является сигналом, существующим в дискретные промежутки времени; он характеризуется последовательностью чисел, определенных для каждого момента времени, kT, где k — целое число, а T — фиксированный промежуток времени.

1.2.4. Сигналы, выраженные через энергию или мощность

Электрический сигнал можно представить как изменение напряжения $v(t)$ или тока $i(t)$ с мгновенной мощностью $p(t)$, подаваемой на сопротивление $\Re$:

$$p(t) = \frac{v^2(t)}{\Re} \tag{1.3,а}$$

или

$$p(t) = i^2(t)\Re \tag{1.3,б}$$

В системах связи мощность часто нормируется (предполагается, что сопротивление $\Re$ равно 1 Ом, хотя в реальном канале оно может быть любым). Если требуется определить действительное значение мощности, оно получается путем "денормирования" нормированного значения. В нормированном случае уравнения (1.3,а) и (1.3,б) имеют одинаковый вид. Следовательно, вне зависимости от того, представлен сигнал через напряжение или ток, нормированная форма позволяет нам выразить мгновенную мощность как

$$p(t) = x^2(t), \tag{1.4}$$

где $x(t)$ — это либо напряжение, либо ток. Рассеивание энергии в течение промежутка времени $(-T/2, T/2)$ реального сигнала с мгновенной мощностью, полученной с помощью уравнения (1.4), может быть записано следующим образом.

$$E_x^T = \int_{-T/2}^{T/2} x^2(t)dt \tag{1.5}$$

Средняя мощность, рассеиваемая сигналом в течение этого интервала, равна следующему.

$$P_x^T = \frac{1}{T}E_x^T = \frac{1}{T}\int_{-T/2}^{T/2} x^2(t)dt \tag{1.6}$$

Производительность системы связи зависит от *энергии* принятого сигнала; сигналы с более высокой энергией обнаруживаются более достоверно (с меньшим числом ошибок) — *работу по обнаружению выполняет принятая энергия.* С другой стороны, *мощность* — это *скорость* поступления энергии. Этот момент важен по нескольким причинам. Мощность определяет напряжение, которое необходимо подать на передатчик, и напряженность электромагнитных полей, которые следует учитывать в радиосистемах (т.е. поля в волноводах, соединяющих передатчик с антенной, и поля вокруг излучающих элементов антенны).

При анализе сигналов связи зачастую желательно работать с *энергией сигнала.* Будем называть $x(t)$ *энергетическим сигналом* тогда и только тогда, когда он в любой момент времени имеет ненулевую конечную энергию $(0 < E_x < \infty)$, где

$$E_x = \lim_{T\to\infty}\int_{-T/2}^{T/2} x^2(t)dt = \int_{-\infty}^{\infty} x^2(t)dt \tag{1.7}$$

В реальной ситуации мы всегда передаем сигналы с конечной энергией $(0 < E_x < \infty)$. Впрочем, для описания *периодических сигналов*, которые по определению (уравнение (1.2)) существуют всегда и, следовательно, имеют бесконечную энергию, и для работы со случайными сигналами, также имеющими неограниченную энергию, удобно определить класс сигналов, выражаемых через *мощность.* Итак, сигнал удобно представить с использованием мощности, если он является *периодическим* и в любой момент времени имеет ненулевую конечную мощность $(0 < P_x < \infty)$, где

$$P_x = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} x^2(t)dt \tag{1.8}$$

Определенный сигнал можно отнести *либо* к энергетическому, *либо* периодическому. Энергетический сигнал имеет конечную энергию, но *нулевую среднюю мощность*, тогда как периодический сигнал имеет нулевую среднюю мощность, но *бесконечную энергию.* Сигнал в системе может выражаться либо через его энергетические, либо периодические значения. *Общее правило*: периодические и случайные сигналы выражаются через мощность, а сигналы, являющиеся детерминированными и непериодическими, — через энергию [1, 2].

Энергия и мощность сигнала — это два важных параметра в описании системы связи. Классификация сигнала либо как энергетического, либо как периодического является удобной моделью, облегчающей математическую трактовку различных сигналов и шумов. В разделе 3.1.5 эти идеи развиваются в контексте цифровых систем связи.

1.2.5. Единичная импульсная функция

Полезной функцией в теории связи является единичный импульс, или *дельта-функция Дирака* $\delta(t)$. Импульсная функция — это абстракция, импульс с бесконечно большой амплитудой, нулевой шириной и единичным весом (площадью под импульсом), сконцентрированный в точке, в которой значение его аргумента равно нулю. Единичный импульс задается следующими соотношениями.

$$\int_{-\infty}^{\infty}\delta(t)dt = 1 \tag{1.9}$$

$$\delta(t) = 0 \text{ для } t \neq 0 \tag{1.10}$$

$$\delta(t) \text{ не ограничена в точке } t = 0 \tag{1.11}$$

$$\int_{-\infty}^{\infty}x(t)\delta(t-t_0)dt = x(t_0) \tag{1.12}$$

Единичный импульс $\delta(t)$ — это не функция в привычном смысле этого слова. Если $\delta(t)$ входит в какую-либо операцию, его удобно считать импульсом конечной амплитуды, единичной площади и ненулевой длительности, после чего нужно рассмотреть предел при стремлении длительности импульса к нулю. Графически $\delta(t-t_0)$ можно изобразить как пик, расположенный в точке $t=t_0$, высота которого равна интегралу от него или его площади. Таким образом, $A\delta(t-t_0)$ с постоянной A представляет импульсную функцию, площадь которой (или вес) равна A, а значение везде нулевое, за исключением точки $t=t_0$.

Уравнение (1.12) известно как *просеивающее* (или *квантующее*) *свойство* единичной импульсной функции; интеграл от единичного импульса и произвольной функции дает выборку функции $x(t)$ в точке $t=t_0$.

1.3. Спектральная плотность

Спектральная плотность (spectral density) характеристик сигнала — это распределение энергии или мощности сигнала по диапазону частот. Особую важность это понятие приобретает при рассмотрении фильтрации в системах связи. Мы должны иметь возможность оценить сигнал и шум на выходе фильтра. При проведении подобной оценки используется спектральная плотность энергии (energy spectral density — ESD) или спектральная плотность мощности (power spectral density — PSD).

1.3.1. Спектральная плотность энергии

Общая энергия действительного энергетического сигнала $x(t)$, определенного в интервале $(-\infty, \infty)$, описывается уравнением (1.7). Используя теорему Парсеваля [1], мы мо-

жем связать энергию такого сигнала, выраженную во временной области, с энергией, выраженной в частотной области:

$$E_x = \int_{-\infty}^{\infty} x^2(t)dt = \int_{-\infty}^{\infty} |X(f)|^2 \, df \,, \tag{1.13}$$

где $X(f)$ — Фурье-образ непериодического сигнала $x(t)$. (Краткие сведения об анализе Фурье можно найти в приложении А.) Обозначим через $\psi_x(f)$ прямоугольный амплитудный спектр, определенный как

$$\psi_x(f) = |X(f)|^2 \tag{1.14}$$

Величина $\psi_x(f)$ является *спектральной плотностью энергии* (ESD) сигнала $x(t)$. Следовательно, из уравнения (1.13) можно выразить общую энергию $x(t)$ путем интегрирования спектральной плотности по частоте.

$$E_x = \int_{-\infty}^{\infty} \psi_x(f)df \tag{1.15}$$

Данное уравнение показывает, что энергия сигнала равна площади под $\psi_x(f)$ на графике в частотной области. Спектральная плотность энергии описывает энергию сигнала на единицу ширины полосы и измеряется в Дж/Гц. Положительные и отрицательные частотные компоненты дают равные энергетические вклады, поэтому, для реального сигнала $x(t)$, величина $|X(f)|$ представляет собой четную функцию частоты. Следовательно, спектральная плотность энергии симметрична по частоте относительно начала координат, а общую энергию сигнала $x(t)$ можно выразить следующим образом.

$$E_x = 2\int_{0}^{\infty} \psi_x(f)df \tag{1.16}$$

1.3.2. Спектральная плотность мощности

Средняя мощность P_x действительного сигнала в периодическом представлении $x(t)$ определяется уравнением (1.8). Если $x(t)$ — это периодический сигнал с периодом T_0, он классифицируется как сигнал в периодическом представлении. Выражение для средней мощности периодического сигнала дается формулой (1.6), где среднее по времени берется за один период T_0.

$$P_x = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} x^2(t)dt \tag{1.17,а}$$

Теорема Парсеваля для действительного периодического сигнала [1] имеет вид

$$P_x = \frac{1}{T}\int_{-T_0/2}^{T_0/2} x^2(t)dt = \sum_{n=-\infty}^{\infty} |c_n|^2 \,, \tag{1.17,б}$$

где члены $|c_n|$ являются комплексными коэффициентами ряда Фурье для периодического сигнала (см. приложение А).

Чтобы использовать уравнение (1.17,б), необходимо знать только значение коэффициентов $|c_n|$. Спектральная плотность мощности (PSD) $G_x(f)$ периодического сигнала $x(t)$, которая является действительной, четной и неотрицательной функцией частоты и дает распределение мощности сигнала $x(t)$ по диапазону частот, определяется следующим образом.

$$G_x(f) = \sum_{n=-\infty}^{\infty} |c_n|^2 \delta(f - nf_0) \tag{1.18}$$

Уравнение (1.18) определяет спектральную плотность мощности периодического сигнала $x(t)$ как последовательность взвешенных дельта-функций. Следовательно, PSD периодического сигнала является дискретной функцией частоты. Используя PSD, определенную в уравнении (1.18), можно записать среднюю нормированную мощность действительного сигнала.

$$P_x = \int_{-\infty}^{\infty} G_x(f)df = 2\int_{0}^{\infty} G_x(f)df \tag{1.19}$$

Уравнение (1.18) описывает PSD только периодических сигналов. Если $x(t)$ — непериодический сигнал, он *не может быть* выражен через ряд Фурье; если он является непериодическим сигналом в периодическом представлении (имеющим бесконечную энергию), он *может не иметь* Фурье-образа. Впрочем, мы по-прежнему можем выразить спектральную плотность мощности таких сигналов в *пределе*. Если сформировать *усеченную версию* $x_T(t)$ непериодического сигнала в периодическом представлении $x(t)$, взяв для этого только его значения из интервала $(-T/2, T/2)$, то $x_T(t)$ будет иметь конечную энергию и соответствующий Фурье-образ $X_T(f)$. Можно показать [2], что спектральная плотность мощности непериодического сигнала $x(t)$ определяется как предел.

$$G_x(f) = \lim_{T\to\infty} \frac{1}{T} |X_T(f)|^2 \tag{1.20}$$

Пример 1.1. Средняя нормированная мощность

а) Найдите среднюю нормированную мощность сигнала $x(t) = A\cos \pi f_0 t$, используя усреднение по времени.

б) Выполните п. а путем суммирования спектральных коэффициентов.

Решение

а) Используя уравнение (1.17,а), имеем следующее.

$$P_x = \frac{1}{T}\int_{-T_0/2}^{T_0/2} (A^2\cos^2 2\pi f_0 t)\,dt =$$

$$= \frac{A^2}{2T_0}\int_{-T_0/2}^{T_0/2} (1 + \cos 4\pi f_0 t)\,dt =$$

$$= \frac{A^2}{2T_0}(T_0) = \frac{A^2}{2}$$

б) Используя уравнения (1.18) и (1.19), получаем следующее.

$$G_x(f) = \sum_{n=-\infty}^{\infty} |c_n|^2 \delta(f - nf_0)$$

$$\left.\begin{array}{r} c_1 = c_{-1} = \dfrac{A}{2} \\ c_n = 0 \quad \text{для } n = 0, \pm 2, \pm 3, \ldots \end{array}\right\} \text{(см. приложение А)}$$

$$G_x(f) = \left(\frac{A}{2}\right)^2 \delta(f - f_0) + \left(\frac{A}{2}\right)^2 \delta(f + f_0)$$

$$P_x = \int_{-\infty}^{\infty} G_x(f)\, df = \frac{A^2}{2}$$

1.4. Автокорреляция

1.4.1. Автокорреляция энергетического сигнала

Корреляция — это процесс согласования; *автокорреляцией* называется согласование сигнала с собственной запаздывающей версией. Автокорреляционная функция действительного энергетического сигнала *x(t) определяется следующим образом.*

$$R_x(\tau) = \int_{-\infty}^{\infty} x(t)x(t+\tau)dt \quad \text{для } -\infty < \tau < \infty \tag{1.21}$$

Автокорреляционная функция $R_x(\tau)$ дает меру похожести сигнала с собственной копией, смещенной на τ единиц времени. Переменная τ играет роль параметра сканирования или поиска. $R_x(\tau)$ — это не функция времени; это всего лишь функция разности времен τ между сигналом и его смещенной копией.

Автокорреляционная функция действительного *энергетического* сигнала имеет следующие свойства.

1. $R_x(\tau) = R_x(-\tau)$	симметрия по τ относительно нуля
2. $R_x(\tau) \le R_x(0)$ для всех τ	максимальное значение в нуле
3. $R_x(\tau) \leftrightarrow \psi_x(f)$	автокорреляция и ESD являются Фурье-образами друг друга, что обозначается двусторонней стрелкой
4. $R_x(0) = \int_{-\infty}^{\infty} x^2(t)\, dt$	значение в нуле равно энергии сигнала

При удовлетворении пп. 1–3 $R_x(\tau)$ является автокорреляционной функцией. Условие 4 — следствие условия 3, поэтому его не обязательно включать в основной набор для проверки на автокорреляционную функцию.

1.4.2. Автокорреляция периодического сигнала

Автокорреляция действительного периодического сигнала $x(t)$ определяется следующим образом.

$$R_x(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\tau)dt \quad \text{для } -\infty < \tau < \infty \tag{1.22}$$

Если сигнал $x(t)$ является периодическим с периодом T_0, среднее по времени в уравнении (1.22) можно брать по *одному периоду* T_0, а автокорреляцию выражать следующим образом.

$$R_x(\tau) = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x(t)x(t+\tau)dt \quad \text{для } -\infty < \tau < \infty \tag{1.23}$$

Автокорреляция *периодического* сигнала, принимающего действительные значения, имеет свойства, сходные со свойствами энергетического сигнала.

1. $R_x(\tau) = R_x(-\tau)$	симметрия по τ относительно нуля
2. $R_x(\tau) \le R_x(0)$ для всех τ	максимальное значение в нуле
3. $R_x(\tau) \leftrightarrow G_x(f)$	автокорреляция и PSD являются Фурье-образами друг друга
4. $R_x(0) = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x^2(t)\,dt$	значение в нуле равно средней мощности сигнала

1.5. Случайные сигналы

Основной задачей системы связи является передача информации по каналу связи. Все полезные сигналы сообщений появляются случайным образом, т.е. приемник не знает заранее, какой из возможных символов сообщений будет передан. Кроме того, вследствие различных электрических процессов возникают шумы, которые сопровождают информационные сигналы. Следовательно, нам нужен эффективный способ описания случайных сигналов.

1.5.1. Случайные переменные

Пусть *случайная переменная* $X(A)$ представляет функциональное отношение между случайным событием A и действительным числом. Для удобства записи обозначим случайную переменную через X, а ее функциональную зависимость от A будем считать явной. Случайная переменная может быть дискретной или непрерывной. *Распределение* $F_X(x)$ случайной переменной X находится выражением:

$$F_X(x) = P(X \le x), \tag{1.24}$$

где $P(X \le x)$ — вероятность того, что значение принимаемой случайной переменной X меньше действительного числа x или равно ему. Функция распределения $F_X(x)$ имеет следующие свойства.

1. $0 \le F_X(x) \le 1$
2. $F_X(x_1) \le F_X(x_2)$, если $x_1 \le x_2$
3. $F_X(-\infty) = 0$
4. $F_X(+\infty) = 1$

Еще одной полезной функцией, связанной со случайной переменной X, является *плотность вероятности*, которая записывается следующим образом.

$$p_X(x) = \frac{dF_X(x)}{dx} \tag{1.25,а}$$

Как и в случае функции распределения, плотность вероятности — это функция действительного числа x. Название "функция плотности" появилось вследствие того, что вероятность события $x_1 \le X \le x_2$ равна следующему.

$$\begin{aligned} P(x_1 \le X \le x_2) &= P(X \le x_2) - P(X \le x_1) = \\ &= F_X(x_2) - F_X(x_1) = \\ &= \int_{x_1}^{x_2} p_X(x)dx \end{aligned} \tag{1.25,б}$$

Используя уравнение (1.25,б), можно приближенно записать вероятность того, что случайная переменная X имеет значение, принадлежащее очень малому промежутку между x и $x + \Delta x$.

$$P(x \le X \le x + \Delta x) \approx p_X(x)\Delta x \tag{1.25,в}$$

Таким образом, в пределе при Δx, стремящемся к нулю, мы можем записать следующее.

$$P(X = x) = p_X(x)dx \tag{1.25,г}$$

Плотность вероятности имеет следующие свойства.

1. $p_X(x) \ge 0$.
2. $\int_{-\infty}^{\infty} p_X(x)dx = F_X(+\infty) - F_X(-\infty) = 1$.

Таким образом, плотность вероятности всегда неотрицательна и имеет единичную площадь. В тексте книги мы будем использовать запись $p_X(x)$ для обозначения плотности вероятности для непрерывной случайной переменной. Для удобства записи мы часто будем опускать индекс X и писать просто $p(x)$. Если случайная переменная X

может принимать только *дискретные* значения, для обозначения плотности вероятности мы будем использовать запись $p(X = x_i)$.

1.5.1.1. Среднее по ансамблю

Среднее значение (mean value) m_X, или *математическое ожидание* (expected value), случайной переменной X определяется выражением

$$m_X = \mathbf{E}\{X\} = \int_{-\infty}^{\infty} x p_X(x)\, dx, \tag{1.26}$$

где $\mathbf{E}\{\cdot\}$ именуется *оператором математического ожидания* (expected value operator). *Моментом n-го порядка* распределения вероятностей случайной переменной X называется следующая величина.

$$\mathbf{E}\{X^n\} = \int_{-\infty}^{\infty} x^n p_X(x)\, dx \tag{1.27}$$

Для анализа систем связи важны первые два момента переменной X. Так, при $n = 1$ уравнение (1.27) дает момент m_X, рассмотренный выше, а при $n = 1$ — среднеквадратическое значение X.

$$\mathbf{E}\{X^2\} = \int_{-\infty}^{\infty} x^2 p_X(x)\, dx \tag{1.28}$$

Можно также определить *центральные моменты*, представляющие собой моменты разности X и m_X. Центральный момент второго порядка (называемый также *дисперсией*) равен следующему.

$$\mathrm{var}(X) = \mathbf{E}\{(X - m_X)^2\} = \int_{-\infty}^{\infty} (x - m_X)^2 p_X(x)\, dx \tag{1.29}$$

Дисперсия X также записывается как σ_X^2, а квадратный корень из этой величины, σ_X, называется *среднеквадратическим отклонением* X. Дисперсия — это мера "разброса" случайной переменной X. Задание дисперсии случайной переменной ограничивает ширину функции плотности вероятности. Дисперсия и среднеквадратическое значение связаны следующим соотношением.

$$\begin{aligned}\sigma_X^2 &= \mathbf{E}\{X^2 - 2m_X X + m_X^2\} = \\ &= \mathbf{E}\{X^2\} - 2m_X \mathbf{E}\{X\} + m_X^2 = \\ &= \mathbf{E}\{X^2\} - m_X^2\end{aligned}$$

Таким образом, дисперсия равна разности среднеквадратического значения и квадрата среднего значения.

1.5.2. Случайные процессы

Случайный процесс $X(A, t)$ можно рассматривать как функцию двух переменных: *события A* и *времени*. На рис. 1.5 представлен пример случайного процесса. Показаны N

выборочных функций времени $\{X_j(t)\}$. Каждую из выборочных функций можно рассматривать как выход отдельного генератора шума. Для каждого события A_j имеем единственную функцию времени $X(A_j, t) = X_j(t)$ (т.е. выборочную функцию). Совокупность всех выборочных функций называется *ансамблем*. В любой определенный момент времени t_k, $X(A, t_k)$ — это случайная переменная $X(t_k)$, значение которой зависит от события. И последнее, для конкретного события $A = A_j$ и для конкретного момента времени $t = t_k$, $X(A_j, t_k)$ — это обычное *число*. Для удобства записи будем обозначать случайный процесс через $X(t)$, а функциональную зависимость от A будем считать явной.

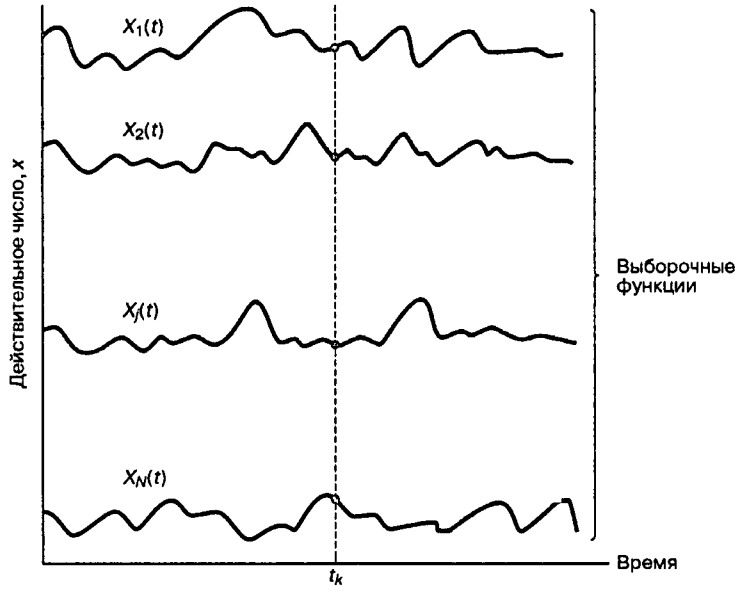


Рис. 1.5. Случайный процесс шума

1.5.2.1. Статистическое среднее случайного процесса

Поскольку значение случайного процесса в каждый последующий момент времени неизвестно, случайный процесс, функции распределения которого непрерывны, можно описать статистически через плотность вероятности. Вообще, в различные моменты времени эта функция для случайного процесса будет иметь разный вид. В большинстве случаев эмпирически определить распределение вероятностей случайного процесса нереально. В то же время для нужд систем связи часто достаточно частичного описания, включающего среднее и функцию автокорреляции. Итак, определим среднее случайного процесса $X(t)$ как

$$\mathrm{E}\{X(t_k)\} = \int_{-\infty}^{\infty} x p_{X_k}(x)\, dx = m_X(t_k), \tag{1.30}$$

где $X(t_k)$ — случайная переменная, полученная при рассмотрении случайного процесса в момент времени t_k, а $p_{X_k}(x)$ — плотность вероятности $X(t_k)$ (плотность по ансамблю событий в момент времени t_k).

Определим автокорреляционную функцию случайного процесса $X(t)$ как функцию двух переменных t_1 и t_2

$$R_X(t_1, t_2) = \mathbf{E}\{X(t_1)X(t_2)\}, \tag{1.31}$$

где $X(t_1)$ и $X(t_2)$ — случайные переменные, получаемые при рассмотрении $X(t)$ в моменты времени t_1 и t_2 соответственно. Автокорреляционная функция — это мера связи двух временных выборок одного случайного процесса.

1.5.2.2. Стационарность

Случайный процесс $X(t)$ называется *стационарным в строгом смысле*, если ни на одну из его статистик не влияет перенос начала отсчета времени. Случайный процесс именуется *стационарным в широком смысле*, если две его статистики, среднее и автокорреляционная функция, не меняются при переносе начала отсчета времени. Таким образом, процесс является стационарным в широком смысле, если

$$\mathbf{E}\{X(t)\} = m_X = \text{константа} \tag{1.32}$$

и

$$R_X(t_1, t_2) = R_X(t_1 - t_2) \tag{1.33}$$

Стационарность в строгом смысле подразумевает стационарность в широком смысле, но не наоборот. Большинство полезных результатов теории связи основывается на предположении, что случайные информационные сигналы и шум являются стационарными в широком смысле. С практической точки зрения случайный процесс не обязательно всегда должен быть стационарным, достаточно стационарности в некотором наблюдаемом интервале времени, представляющем практический интерес.

Для стационарных процессов автокорреляционная функция в уравнении (1.33) зависит не от времени, а только от разности $t_1 - t_2$. Иными словами, все пары значений $X(t)$ в моменты времени, разделенные промежутком $\tau = t_1 - t_2$, имеют одинаковое корреляционное значение. Следовательно, для стационарных систем функцию $R_X(t_1, t_2)$ можно записывать просто как $R_X(\tau)$.

1.5.2.3. Автокорреляция случайных процессов, стационарных в широком смысле

Как дисперсия предлагает меру случайности для случайных переменных, так и автокорреляционная функция предлагает подобную меру для случайных процессов. Для процессов, стационарных в широком смысле, автокорреляционная функция зависит только от разности времен $\tau = t_1 - t_2$.

$$R_X(\tau) = \mathbf{E}\{X(t)X(t+\tau)\} \text{ для } -\infty < \tau < \infty \tag{1.34}$$

Для стационарного в широком смысле процесса с нулевым средним, функция $R_X(\tau)$ показывает, насколько статистически коррелируют случайные величины процесса, разделенные τ секундами. Другими словами, $R_X(\tau)$ дает информацию о частотной характеристике, связанной со случайным процессом. Если $R_X(\tau)$ меняется медленно по мере увеличения τ от нуля до некоторого значения, это показывает, что в среднем выборочные значения $X(t)$, взятые в моменты времени $t = t_1$ и $t = t_2$, практически равны. Следовательно, мы вправе ожидать, что в частотном представлении $X(t)$ будут преоб-

ладать низкие частоты. С другой стороны, если $R_X(\tau)$ быстро уменьшается по мере увеличения τ, стоит ожидать, что $X(t)$ будет быстро меняться по времени и, следовательно, будет включать преимущественно высокие частоты.

Автокорреляционная функция стационарного в широком смысле процесса, принимающего действительные значения, имеет следующие свойства.

1. $R_X(\tau) = R_X(-\tau)$	симметрия по τ относительно нуля
2. $R_X(\tau) \le R_X(0)$ для всех τ	максимальное значение в нуле
3. $R_X(\tau) \leftrightarrow G_X(f)$	автокорреляция и спектральная плотность мощности являются Фурье-образами друг друга
4. $R_X(0) = \mathbf{E}\{X^2(t)\}$	значение в нуле равно средней мощности сигнала

1.5.3. Усреднение по времени и эргодичность

Для вычисления m_X и $R_X(\tau)$ путем усреднения по ансамблю нам нужно усреднить их по всем выборочным функциям процесса, и, значит, нам потребуется полная информация о взаимном распределении функций плотности вероятности в первом и втором приближениях. В общем случае, как правило, такая информация недоступна.

Если случайный процесс принадлежит к особому классу, называемому классом *эргодических процессов*, его среднее по времени равно среднему по ансамблю и статистические свойства процесса можно определить путем *усреднения по времени одной выборочной функции* процесса. Чтобы случайный процесс был эргодическим, он должен быть стационарным в строгом смысле (обратное не обязательно). Впрочем, для систем связи, где нам достаточно стационарности в широком смысле, нас интересуют только среднее и автокорреляционная функция.

Говорят, что случайный процесс является *эргодическим по отношению к среднему значению*, если

$$m_X = \lim_{T \to \infty} 1/T \int_{-T/2}^{T/2} X(t)\, dt, \tag{1.35}$$

и *эргодическим по отношению к автокорреляционной функции*, если

$$R_X(\tau) = \lim_{T \to \infty} 1/T \int_{-T/2}^{T/2} X(t)\, X(t+\tau)\, dt. \tag{1.36}$$

Проверка случайного процесса на эргодичность обычно весьма непроста. На практике, как правило, используется интуитивное предположение о целесообразности замены средних по ансамблю средними по времени. При анализе большинства сигналов в каналах связи (при отсутствии импульсных эффектов) разумным будет предположение, что случайные сигналы являются эргодическими по отношению к автокорреляционной функции. Поскольку для эргодических процессов средние по времени равны средним по ансамблю, фундаментальные электротехнические параметры, такие как амплитуда постоянной составляющей, среднеквадратическое значение и средняя мощность, могут быть связаны с моментами эргодического случайного процесса.

1. Величина $m_X = \mathbf{E}\{X(t)\}$ равна постоянной составляющей сигнала.
2. Величина m_x^2 равна нормированной мощности постоянной составляющей.
3. Момент второго порядка $X(t)$, $\mathbf{E}\{X_2(t)\}$, равен общей средней нормированной мощности.
4. Величина $\sqrt{\mathbf{E}\{X^2(t)\}}$ равна среднеквадратическому значению сигнала, выраженного через ток или напряжение.
5. Дисперсия σ_X^2 равна средней нормированной мощности переменного сигнала.
6. Если среднее процесса равно нулю (т.е. $m_X = m_X^2 = 0$), то $\sigma_X^2 = \mathbf{E}\{X^2\}$, а дисперсия равна среднеквадратическому значению или (другая формулировка) дисперсия представляет общую мощность в нормированной нагрузке.
7. Среднеквадратическое отклонение σ^2 является среднеквадратическим значением переменного сигнала.
8. Если $m_X = 0$, то σ_X — это среднеквадратическое значение сигнала.

1.5.4. Спектральная плотность мощности и автокорреляция случайного процесса

Случайный процесс $X(t)$ можно отнести к периодическому сигналу, имеющему такую спектральную плотность мощности $G_X(f)$, как указано в уравнении (1.20). Функция $G_X(f)$ особенно полезна в системах связи, поскольку она описывает распределение мощности сигнала по диапазону частот. Спектральная плотность мощности позволяет оценить мощность сигнала, который будет передаваться через сеть с известными частотными характеристиками. Основные свойства функций спектральной плотности мощности можно сформулировать следующим образом.

1. $G_X(f) \geq 0$	всегда принимает действительные значения
2. $G_X(f) = G_X(-f)$	для $X(t)$, принимающих действительные значения
3. $G_X(f) \leftrightarrow R_X(\tau)$	автокорреляция и спектральная плотность мощности являются Фурье-образами друг друга
4. $P_X = \int_{-\infty}^{\infty} G_X(f)\, df$	связь между средней нормированной мощностью и спектральной плотностью мощности

На рис. 1.6 приведено визуальное представление автокорреляционной функции и функции спектральной плотности мощности. Что означает термин "корреляция"? Когда мы интересуемся корреляцией двух явлений, спрашиваем, насколько близко они соотносятся по поведению или виду и насколько они совпадают. В математике автокорреляционная функция сигнала (во временной области) описывает соответствие сигнала самому себе, смещенному на некоторый промежуток времени. Точная копия считается созданной и локализированной на минус бесконечности. Затем мы последовательно перемещаем копию в положительном направлении временной оси и задаем вопрос, насколько они (исходная версия и копия) соответствуют друг другу. Затем мы перемещаем копию еще на один шаг в положительном направлении и задаем вопрос, насколько они совпа-

дают теперь, и т.д. Корреляция между двумя сигналами изображается как функция времени, обозначаемого τ; при этом время τ можно рассматривать как параметр сканирования.

На рис. 1.6, *а–г* изображена описанная выше ситуация в некоторые моменты времени. Рис. 1.6, *а* иллюстрирует отдельный сигнал стационарного в широком смысле случайного процесса $X(t)$. Сигнал представляет собой случайную двоичную последовательность с положительными и отрицательными (биполярными) импульсами единичной амплитуды. Положительные и отрицательные импульсы появляются с равной вероятностью. Длительность каждого импульса (двоичной цифры) равна T секунд, а среднее, или величина постоянной составляющей случайной последовательности, равно нулю. На рис. 1.6, *б* показана та же последовательность, смещенная во времени на τ_1 секунд. Согласно принятым обозначениям, эта последовательность обозначается $X(t-\tau_1)$. Предположим, что процесс $X(t)$ является эргодическим по отношению к автокорреляционной функции, поэтому для нахождения $R_X(\tau)$ мы можем использовать усреднение по времени вместо усреднения по ансамблю. Значение $R_X(\tau)$ получается при перемножении двух последовательностей $X(t)$ и $X(t-\tau_1)$ с последующим нахождением среднего с помощью уравнения (1.36), которое справедливо для эргодических процессов *только в пределе*. Впрочем, интегрирование по целому числу периодов может дать нам некоторую оценку $R_X(\tau)$. Отметим, что $R_X(\tau_1)$ может быть получено при смещении $X(t)$ как в положительном, так и отрицательном направлении. Подобный случай иллюстрирует рис. 1.6, *в*, на котором использована исходная выборочная последовательность (рис. 1.6, *а*) и ее смещенная копия (рис. 1.6, *б*). Заштрихованные области под кривой произведения $X(t)X(t-\tau_1)$ вносят положительный вклад в произведение, а серые области — отрицательный. Интегрирование $X(t)X(t-\tau_1)$ по времени передачи импульсов дает точку $R_X(\tau_1)$ на кривой $R_X(\tau)$. Последовательность может далее смещаться на τ_2, τ_3, ..., и каждое такое смещение будет давать точку на общей автокорреляционной функции $R_X(\tau)$, показанной на рис. 1.6, *г*. Иными словами, каждой случайной последовательности биполярных импульсов соответствует автокорреляционная точка на общей кривой, приведенной на рис. 1.6, *г*. Максимум функции находится в точке $R_X(0)$ (наилучшее соответствие имеет место при τ, равном нулю, поскольку для всех τ $R(\tau) \le R(0)$), и функция спадает по мере роста τ. На рис. 1.6, *г* показаны точки, соответствующие $R_X(0)$ и $R_X(\tau_1)$.

Аналитическое выражение для автокорреляционной функции $R_X(\tau)$, приведенной на рис. 1.6, *г*, имеет следующий вид [1].

$$R_X(\tau) = \begin{cases} 1 - \dfrac{|\tau|}{T} & \text{для } |\tau| \le T \\ 0 & \text{для } |\tau| > T \end{cases} \tag{1.37}$$

Отметим, что автокорреляционная функция дает нам информацию о частоте; она сообщает нам кое-что о полосе сигнала. В то же время автокорреляция — это временная функция; в формуле (1.37) отсутствуют члены, зависящие от частоты. Так как же она дает нам информацию о полосе сигнала?

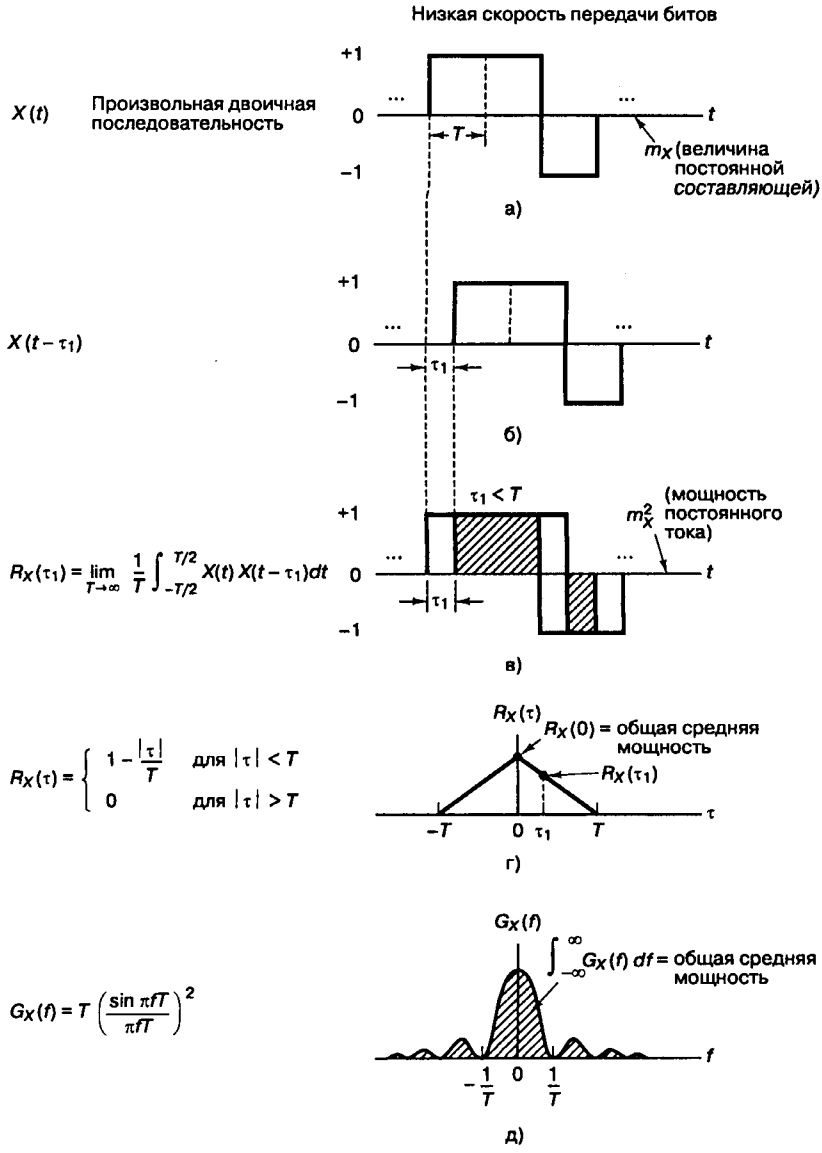


Рис. 1.6. Автокорреляция и спектральная плотность мощности

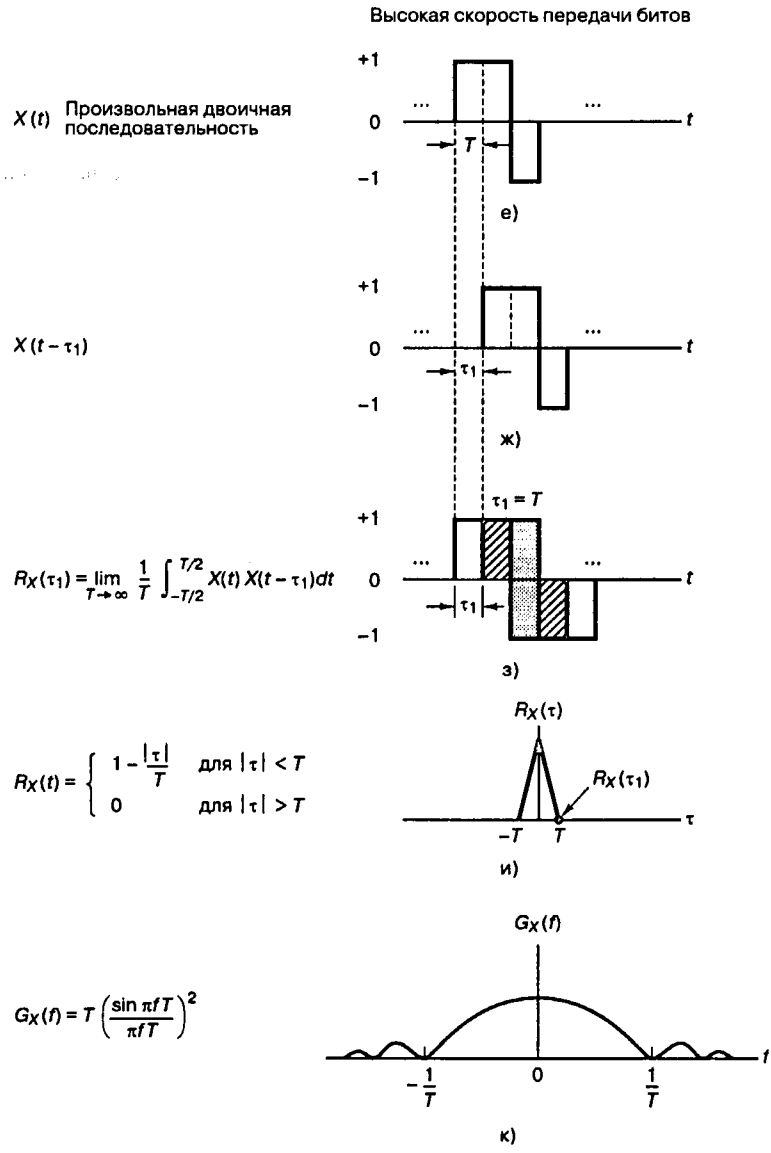


Рис. 1.6. Автокорреляция и спектральная плотность мощности (окончание)

Предположим, что сигнал перемещается очень медленно (сигнал имеет малую ширину полосы). Если мы будем смещать копию сигнала вдоль оси τ, задавая на каждом этапе смещения вопрос, насколько соответствуют друг другу копия и оригинал, соответствие достаточно долго будет довольно сильным. Другими словами, треугольная автокорреляционная функция (рис. 1.6, *г* и формула 1.37) будет медленно спадать с ростом τ. Предположим теперь, что сигнал меняется достаточно быстро (т.е. имеем большую полосу). В этом случае даже небольшое изменение τ приведет к тому, что корреляция будет нулевой и автокорреляционная функция будет иметь очень узкую форму. Следовательно, сравнение автокорреляционных функций по форме дает нам некоторую инфор-

мацию о ширине полосы сигнала. Функция спадает постепенно? В этом случае имеем сигнал с узкой полосой. Форма функции напоминает узкий пик? Тогда сигнал имеет широкую полосу.

Автокорреляционная функция позволяет явно выражать спектральную плотность мощности случайного сигнала. Поскольку спектральная плотность мощности и автокорреляционная функция являются Фурье-образами друг друга, спектральную плотность мощности, $G_X(f)$, случайной последовательности биполярных импульсов можно найти как Фурье-преобразование функции $R_X(\tau)$, аналитическое выражение которой дано в уравнении (1.37). Для этого можно использовать табл. А.1. Заметим, что

$$G_X(f) = T\left(\frac{\sin \pi fT}{\pi fT}\right)^2 = T\,\mathrm{sinc}^2\, fT, \tag{1.38}$$

где

$$\mathrm{sinc}\, y = \frac{\sin \pi y}{\pi y}. \tag{1.39}$$

Общий вид функции $G_X(f)$ показан на рис. 1.6, *д*.

Отметим, что площадь под кривой спектральной плотности мощности представляет собой среднюю мощность сигнала. Одной из удобных мер *ширины полосы* является ширина основного спектрального лепестка (см. раздел 1.7.2). На рис. 1.6, *д* показано, что ширина полосы сигнала связана с обратной длительностью символа или шириной импульса. Рис. 1.6, *е*–*к* формально повторяют рис. 1.6, *а*–*д*, за исключением того, что на последующих рисунках длительность импульса меньше. Отметим, что для более коротких импульсов функция $R_X(\tau)$ уже (рис. 1.6, *и*), чем для более длительных (рис. 1.6, *г*). На рис. 1.6, *и* $R_X(\tau_1) = 0$; другими словами, в случае меньшей длительности импульса смещения на τ_1 достаточно для создания нулевого соответствия или для полной потери корреляции между смещенными последовательностями. Поскольку на рис. 1.6, *е* длительность импульса T меньше (выше скорость передачи импульса), чем на рис. 1.6, *а*, занятость полосы на рис. 1.6, *к* больше занятости полосы для более низкой частоты импульсов, показанной на рис. 1.6, *д*.

1.5.5. Шум в системах связи

Термин "шум" обозначает *нежелательные* электрические сигналы, которые всегда присутствуют в электрических системах. Наличие шума, наложенного на сигнал, "затеняет", или маскирует, сигнал; это ограничивает способность приемника принимать точные решения о значении символов, а следовательно, ограничивает скорость передачи информации. Природа шумов различна и включает как естественные, так и искусственные источники. *Искусственные шумы* — это шумы искрового зажигания, коммутационные импульсные помехи и шумы от других родственных источников электромагнитного излучения. *Естественные шумы* исходят от атмосферы, солнца и других галактических источников.

Хорошее техническое проектирование может устранить большинство шумов или их нежелательные эффекты посредством фильтрации, экранирования, выбора модуляции и оптимального местоположения приемника. Например, чувствительные ра-

диоастрономические измерения проводятся, как правило, в отдаленных пустынных местах, вдали от естественных источников шума. Впрочем, существует один естественный шум, называемый *тепловым*, который устранить нельзя. Тепловой шум [4, 5] вызывается тепловым движением электронов во всех диссипативных компонентах — резисторах, проводниках и т.п. Те же электроны, которые отвечают за электропроводимость, являются причиной теплового шума.

Тепловой шум можно описать как *гауссов* случайный процесс с нулевым средним. Гауссов процесс $n(t)$ — это случайная функция, значение которой n в произвольный момент времени t статистически характеризуется гауссовой функцией плотности вероятностей:

$$p(n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{n}{\sigma}\right)^2\right], \tag{1.40}$$

где σ^2 — дисперсия n. Нормированная гауссова функция плотности процесса с нулевым средним получается в предположении, что $\sigma = 1$. Схематически нормированная функция плотности вероятностей показана на рис. 1.7.

Далее мы часто будем представлять случайный сигнал как сумму случайной переменной, выражающей гауссов шум, и сигнала канала связи.

$$z = a + n$$

Здесь z — случайный сигнал, a — сигнал в канале связи, а n — случайная переменная, выражающая гауссов шум. Тогда функция плотности вероятности $p(z)$ выражается как

$$p(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z-a}{\sigma}\right)^2\right], \tag{1.41}$$

где, как и выше, σ^2 — дисперсия n.

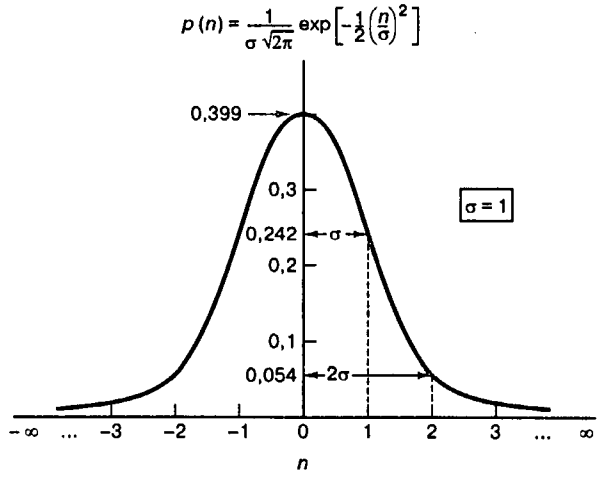


Рис. 1.7. Нормированная ($\sigma = 1$) *гауссова функция плотности вероятности*

Гауссово распределение часто используется как модель шума в системе, поскольку существует *центральная граничная теорема* [3], утверждающая, что при весьма общих условиях распределение вероятностей суммы j статистически независимых случайных переменных подчиняется гауссовому распределению при $j \to \infty$, причем вид отдельных функций распределения не имеет значения. Таким образом, даже если отдельные механизмы шума будут иметь негауссово распределение, совокупность многих таких механизмов будет стремиться к гауссовому распределению.

1.5.5.1. Белый шум

Основной спектральной характеристикой теплового шума является то, что его спектральная плотность мощности *одинакова* для всех частот, представляющих интерес для большинства систем связи; другими словами, источник теплового шума на всех частотах излучает с равной мощностью на единицу ширины полосы — от постоянной составляющей до частоты порядка 10^{12} Гц. Следовательно, простая модель теплового шума предполагает, что его спектральная плотность мощности $G_n(f)$ равномерна для всех частот, как показано на рис. 1.8, *а*, и записывается в следующем виде.

$$G_n(f) = \frac{N_0}{2} \text{ Вт/Гц} \tag{1.42}$$

Здесь коэффициент 2 включен для того, чтобы показать, что $G_n(f)$ — *двусторонняя* спектральная плотность мощности. Когда мощность шума имеет такую единообразную спектральную плотность, мы называем этот шум *белым*. Прилагательное "белый" используется в том же смысле, что и для белого света, содержащего равные доли всех частот видимого диапазона электромагнитного излучения.

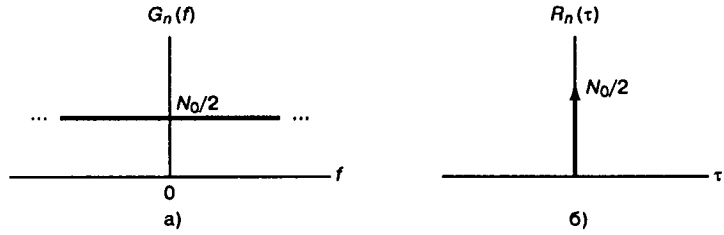


Рис. 1.8. Белый шум: а) спектральная плотность мощности; б) автокорреляционная функция

Автокорреляционная функция белого шума дается обратным преобразованием Фурье спектральной плотности мощности шума (см. табл. А.1) и записывается следующим образом.

$$R_n(\tau) = \mathfrak{F}^{-1}\{G_n(f)\} = \frac{N_0}{2}\delta(\tau) \tag{1.43}$$

Таким образом, автокорреляция белого шума — это дельта-функция, взвешенная множителем $N_0/2$ и находящаяся в точке $\tau = 0$, как показано на рис. 1.8, *б*. Отметим, что $R_n(\tau)$ равна нулю для $\tau \neq 0$, т.е. две различные выборки белого шума не коррелируют, вне зависимости от того, насколько близко они находятся.

Средняя мощность P_n белого шума *бесконечна*, поскольку бесконечна ширина полосы белого шума. Это можно увидеть, получив из уравнений (1.19) и (1.42) следующее выражение.

$$P_n = \int_{-\infty}^{\infty} \frac{N_0}{2} df = \infty \tag{1.44}$$

Хотя белый шум представляет собой весьма полезную абстракцию, ни один процесс шума в действительности не может быть белым; впрочем, шум, появляющийся во многих реальных системах, можно предположительно считать белым. Наблюдать такой шум мы можем только после того, как он пройдет через реальную систему, имеющую конечную ширину полосы. Следовательно, пока ширина полосы шума существенно больше ширины полосы, используемой системой, можно считать, что шум имеет бесконечную ширину полосы.

Дельта-функция в уравнении (1.43) означает, что сигнал шума $n(t)$ абсолютно не коррелирует с собственной смещенной версией для любого $\tau > 0$. Уравнение (1.43) показывает, что *любые* две выборки процесса белого шума не коррелируют. Поскольку тепловой шум — это гауссов процесс и его выборки не коррелируют, выборки шума также являются независимыми [3]. Таким образом, воздействие канала с *аддитивным белым гауссовым шумом* на процесс обнаружения состоит в том, что шум *независимо* воздействует на каждый переданный символ. Такой канал называется *каналом без памяти*. Термин "аддитивный" означает, что шум просто накладывается на сигнал или добавляется к нему — никаких мультипликативных механизмов не существует.

Поскольку тепловой шум присутствует во всех системах связи и для большинства систем является заметным источником шума, характеристики теплового шума (аддитивный, белый и гауссов) часто применяются для моделирования шума в системах связи. Поскольку гауссов шум с нулевым средним полностью характеризуется его *дисперсией*, эту модель особенно просто использовать при обнаружении сигналов и проектировании оптимальных приемников. В данной книге мы будем считать (если не оговорено противное), что система подвергается искажению со стороны *аддитивного белого гауссового шума с нулевым средним*, хотя иногда такое упрощение будет чересчур сильным.

1.6. Передача сигнала через линейные системы

После того как мы разработали набор моделей для сигнала и шума, рассмотрим характеристики систем и их воздействие на сигналы и шумы. Поскольку систему с равным успехом можно охарактеризовать как в частотной, так и во временной области, в обоих случаях были разработаны методы, позволяющие анализировать отклик линейной системы на произвольный входной сигнал. Сигнал, поданный на вход системы (рис. 1.9), можно описать либо как временной сигнал, $x(t)$, либо через его Фурье-образ, $X(f)$. Использование временного анализа дает временной выход $y(t)$, и в процессе будет определена функция $h(t)$, *импульсная характеристика*, или *импульсный отклик*, сети. При рассмотрении ввода в частотной области мы должны определить для системы *частотную характеристику*, или *передаточную функцию* $H(f)$, которая определит частотный выход $Y(f)$. Предполагается, что система линейна и инвариантна относительно времени. Также предполагается, что система не имеет скрытой энергии на момент подачи сигнала на вход.

Рис. 1.9. Линейная система и ее ключевые параметры

1.6.1. Импульсная характеристика

Линейная, инвариантная относительно времени система или сеть, показанная на рис. 1.9, описывается (во временной области) импульсной характеристикой $h(t)$, представляющей собой реакцию системы при подаче на ее вход единичного импульса $\delta(t)$.

$$h(t) = y(t) \text{ при } x(t) = \delta(t) \tag{1.45}$$

Рассмотрим термин "импульсный отклик", крайне подходящий для данного события. Описание характеристик системы через ее импульсный отклик имеет прямую физическую интерпретацию. На вход системы мы подаем единичный импульс (нереальный сигнал, имеющий бесконечную амплитуду, нулевую ширину и единичную площадь), как показано на рис. 1.10, *а*. Подачу такого импульса в систему можно рассматривать как "мгновенный удар". Как отреагирует ("откликнется") система на такое применение силы (импульс)? Выходящий сигнал $h(t)$ — это и есть импульсный отклик системы. (Возможный вид этого отклика показан на рис. 1.10, *б*.)

Отклик сети на произвольный сигнал $x(t)$ является сверткой $x(t)$ с $h(t)$, что записывается следующим образом.

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)\,d\tau \tag{1.46}$$

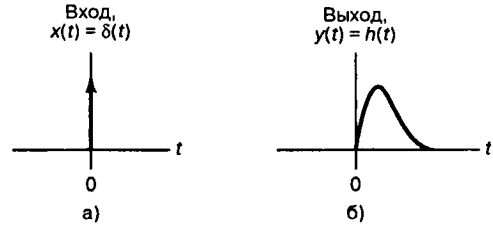


Рис. 1.10. Иллюстрация понятия "импульсный отклик": а) входной сигнал x(t) является единичной импульсной функцией; б) выходной сигнал y(t) — импульсным откликом системы h(t)

Здесь знак "*" обозначает операцию свертки (см. раздел А.5). Система предполагается *причинной*, что означает *отсутствие* сигнала на выходе до момента времени $t = 0$, когда сигнал подается на вход. Следовательно, нижняя граница интегрирования может быть взята равной нулю, и выход $y(t)$ можно выразить несколько иначе.

$$y(t) = \int_{0}^{\infty} x(\tau)h(t-\tau)\,d\tau \tag{1.47,а}$$

или в виде

$$y(t) = \int_0^\infty x(t-\tau)h(\tau)\,d\tau \qquad (1.47,б)$$

Выражения в уравнениях (1.46) и (1.47) называются *интегралами свертки*. Свертка (convolution) — это фундаментальный математический аппарат, играющий важную роль в понимании всех систем связи. Если читатель не знаком с этой операцией, ему стоит обратиться к разделу А.5, где приводится вывод уравнений (1.46) и (1.47).

1.6.2. Частотная передаточная функция

Частотный выходной сигнал $Y(f)$ получаем при применении преобразования Фурье к обеим частям уравнения (1.46). Поскольку свертка во временной области превращается в умножение в частотной (и наоборот), из уравнения (1.46) получаем следующее.

$$Y(f) = X(f)H(f) \qquad (1.48)$$

или

$$H(f) = \frac{Y(f)}{X(f)} \qquad (1.49)$$

(Подразумевается, конечно, что $X(f) \neq 0$ для всех f.) Здесь $H(f) = \mathfrak{F}\{h(t)\}$, Фурье-образ импульсного отклика, называемый *частотной передаточной функцией*, *частотной характеристикой*, или *частотным откликом* сети. Вообще, функция $H(f)$ является комплексной и может быть записана как

$$H(f) = |H(f)|e^{i\theta(f)}, \qquad (1.50)$$

где $|H(f)|$ — модуль отклика. Фаза отклика определяется следующим образом.

$$\theta(f) = \operatorname{arctg}\frac{\operatorname{Im}\{H(f)\}}{\operatorname{Re}\{H(f)\}} \qquad (1.51)$$

(Re и Im обозначают действительную и мнимую части аргумента.)

Частотная передаточная функция линейной, инвариантной относительно времени сети может легко измеряться в лабораторных условиях — в сети с генератором гармонических колебаний на входе и осциллографом на выходе. Если входной сигнал $x(t)$ выразить как

$$x(t) = A\cos 2\pi f_0 t,$$

то выход можно записать следующим образом.

$$y(t) = A\,|H(f_0)|\cos\,[2\pi f_0 t + \theta(f_0)] \qquad (1.52)$$

Входная частота f_0 смещается на интересующее нас значение; таким образом, измерения на входе и выходе позволяют определить вид $\theta(f)$.

1.6.2.1. Случайные процессы и линейные системы

Если случайный процесс формирует вход линейной, инвариантной относительно времени системы, то на выходе этой системы получим также случайный процесс. Иными словами, каждая выборочная функция входного процесса дает выборочную функцию выходного процесса. Входная спектральная плотность мощности $G_X(f)$ и выходная спектральная плотность мощности $G_Y(f)$ связаны следующим соотношением.

$$G_Y(f) = G_X(f)\,|H(f)|^2 \tag{1.53}$$

Уравнение (1.53) предоставляет простой способ нахождения спектральной плотности мощности на выходе линейной, инвариантной относительно времени системы при подаче на вход случайного процесса.

В главах 3 и 4 мы рассмотрим обнаружение сигналов в гауссовом шуме. Основное свойство гауссовых процессов будет применено к линейной системе. Будет показано, что если гауссов процесс $X(t)$ подается на инвариантный относительно времени линейный фильтр, то случайный процесс $Y(t)$, поступающий на выход, также является гауссовым [6].

1.6.3. Передача без искажений

Что необходимо для того, чтобы сеть вела себя как *идеальный* канал передачи? Сигнал на выходе идеального канала связи может запаздывать по отношению к сигналу на входе; кроме того, эти сигналы могут иметь различные амплитуды (простое изменение масштаба), но что касается всего остального — сигнал не должен быть искажен, т.е. он должен иметь ту же форму, что и сигнал на входе. Следовательно, для идеальной неискаженной передачи выходной сигнал мы можем описать как

$$y(t) = Kx(t - t_0), \tag{1.54}$$

где K и t_0 — константы. Применив к обеим частям преобразование Фурье (см. раздел А.3.1), имеем следующее.

$$Y(f) = KX(f)e^{-2\pi i f t_0} \tag{1.55}$$

Подставляя выражение (1.55) в уравнение (1.49), видим, что необходимая передаточная функция системы для передачи без искажений имеет следующий вид.

$$H(f) = Ke^{-2\pi i f t_0} \tag{1.56}$$

Следовательно, для получения *идеальной передачи без искажений* общий отклик системы должен иметь постоянный модуль, а сдвиг фаз должен быть линейным по частоте. Недостаточно, чтобы система равно усиливала или ослабляла все частотные компоненты. Все гармоники сигнала должны поступать на выход с одинаковым запаздыванием, чтобы их можно было просуммировать. Поскольку запаздывание t_0 связано со сдвигом фаз θ и циклической частотой $\omega = 2\pi f$ соотношением

$$t_0(\text{секунд}) = \frac{\theta\ (\text{радиан})}{2\pi f\ (\text{радиан в секунду})}, \tag{1.57,а}$$

очевидно, что, для того чтобы запаздывание всех компонентов было одинаковым, сдвиг фаз должен быть пропорционален частоте. Для измерения искажения сигнала, вызванного запаздыванием, часто используется характеристика, называемая *групповой задержкой*; она определяется следующим образом.

$$\tau(f) = -\frac{1}{2\pi}\frac{d\theta(f)}{df} \tag{1.57,б}$$

Таким образом, для передачи без искажений имеем два эквивалентных требования: фаза должна быть линейной по частоте или групповая задержка $\tau(f)$ должна быть равна константе. На практике сигнал будет искажаться при проходе через некоторые час-

ти системы. Для устранения этого искажения в систему могут вводиться схемы коррекции фазы или амплитуды (*выравнивания*). Вообще, искажение — это общая характеристика ввода-вывода системы, определяющая ее производительность.

1.6.3.1. Идеальный фильтр

Построить идеальную сеть, описываемую уравнением (1.56), нереально. Проблема заключается в том, что в уравнении (1.56) предполагается бесконечная ширина полосы, причем ширина полосы системы определяется интервалом положительных частот, в которых модуль $|H(f)|$ имеет заданную величину. (Вообще, существует несколько мер ширины полосы; самые распространенные перечислены в разделе 1.7.) В качестве приближения к идеальной сети с бесконечной шириной полосы выберем усеченную сеть, без искажения пропускающую все гармоники с частотами между f_l и f_u, где f_l — нижняя частота среза, а f_u — верхняя, как показано на рис. 1.11. Все подобные сети называются *идеальными фильтрами*. Предполагается, что вне диапазона $f_l < f < f_u$, который называется *полосой пропускания* (passband), амплитуда отклика идеального фильтра равна нулю. Эффективная ширина полосы пропускания определяется шириной полосы фильтра и составляет $W_f = (f_u - f_l)$ Гц.

Если $f_l \neq 0$ и $f_u \neq \infty$, фильтр называется *пропускающим* (рис. 1.11, *а*). Если $f_l = 0$ и f_u имеет конечное значение, он именуется *фильтром нижних частот* (рис. 1.11, *б*). Если f_l имеет ненулевое значение и $f_u \to \infty$, он называется *фильтром верхних частот* (рис. 1.11, *в*).

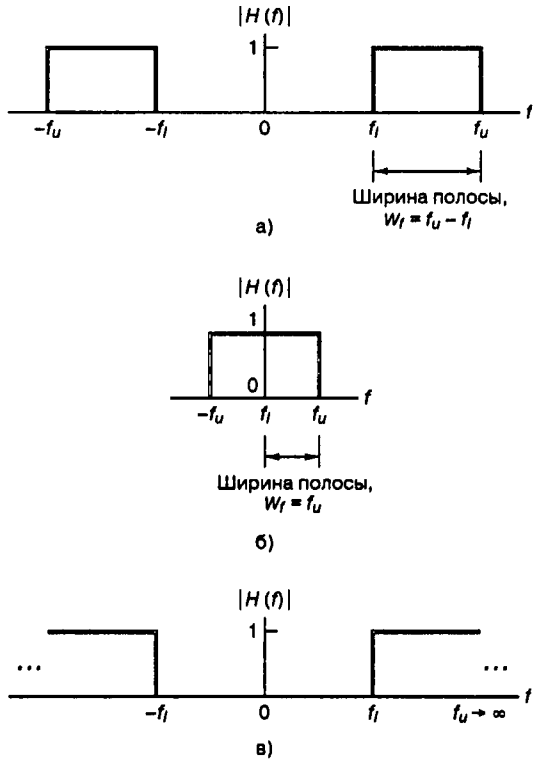


Рис. 1.11. Передаточная функция идеальных фильтров: а) идеальный пропускающий фильтр; б) идеальный фильтр нижних частот; в) идеальный фильтр верхних частот

Используя уравнение (1.59) и полагая $K = 1$ для идеального фильтра нижних частот с шириной полосы $W_f = f_u$ Гц, показанной на рис. 1.11, *б*, можно записать передаточную функцию следующим образом.

$$H(f) = |H(f)|e^{-i\theta(f)}, \tag{1.58}$$

где

$$|H(f)| = \begin{cases} 1 & \text{для } |f| < f_u \\ 0 & \text{для } |f| \geq f_u \end{cases} \tag{1.59}$$

и

$$e^{-i\theta(f)} = e^{-i2\pi f t_0} \tag{1.60}$$

Импульсный отклик идеального фильтра нижних частот, показанный на рис. 1.12, выражается следующей формулой.

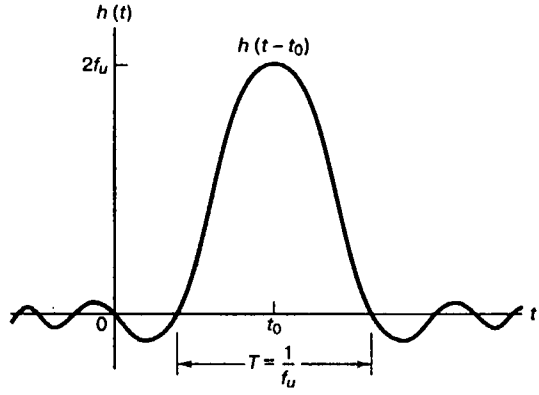


Рис. 1.12. Импульсный отклик идеального фильтра нижних частот

$$h(t) = \mathfrak{F}^{-1}\{H(f)\} = \int_{-\infty}^{\infty} H(f)e^{2\pi ift}df = \tag{1.61}$$

$$= \int_{-f_u}^{f_u} e^{-2\pi ift_0}e^{2\pi ift}df$$

или

$$= \int_{-f_u}^{f_u} e^{-2\pi if(t-t_0)}df =$$

$$= 2f_u \frac{\sin 2\pi f_u(t-t_0)}{2\pi f_u(t-t_0)} =$$

$$= 2f_u \operatorname{sinc} 2f_u(t-t_0), \tag{1.62}$$

где функция $\operatorname{sinc} x$ определена в уравнении (1.39). Импульсный отклик, показанный на рис. 1.12, является непричинным; это означает, что в момент подачи сигнала на

вход ($t = 0$), на выходе фильтра имеется ненулевой отклик. Таким образом, должно быть очевидно, что идеальный фильтр, описываемый уравнением (1.58), не реализуется в действительности.

Пример 1.2. Прохождение белого шума через идеальный фильтр

Белый шум со спектральной плотностью мощности $G_n(f) = N_0/2$, показанный на рис 1.8, *а*, подается на вход идеального фильтра нижних частот, показанного на рис. 1.11, *б*. Определите спектральную плотность мощности $G_Y(f)$ и автокорреляционную функцию $R_Y(\tau)$ выходного сигнала.

Решение

$$G_Y(f) = G_n(f)\,|H(f)|^2 =$$

$$= \begin{cases} \dfrac{N_0}{2} & \text{для } |f| < f_u \\ 0 & \text{для остальных } |f| \end{cases}$$

Автокорреляционная функция — это результат применения обратного преобразования Фурье к спектральной плотности мощности. Определяется автокорреляционная функция следующим выражением (см. табл. А.1).

$$R_Y(\tau) = N_0 f_u \frac{\sin 2\pi f_u \tau}{2\pi f_u \tau} =$$

$$= N_0 f_u \operatorname{sinc} 2f_u\tau$$

Сравнивая полученный результат с формулой (1.62), видим, что $R_Y(\tau)$ имеет тот же вид, что и импульсный отклик идеального фильтра нижних частот, показанный на рис. 1.12. В этом примере идеальный фильтр нижних частот преобразовывает автокорреляционную функцию белого шума (определенную через дельта-функцию) в функцию sinc. После фильтрации в системе уже не будет белого шума. Выходной шумовой сигнал будет иметь нулевую корреляцию с собственными смещенными копиями только при смещении на $\tau = n/2f_u$, где n — любое целое число, отличное от нуля.

1.6.3.2. Реализуемые фильтры

Простейший реализуемый фильтр нижних частот состоит из сопротивления ($\Re$) и емкости (C), как показано на рис. 1.13, *а*; этот фильтр называется $\Re C$-фильтром, и его передаточная функция может быть выражена следующим образом [7].

$$H(f) = \frac{1}{1 + 2\pi i f \Re C} = \frac{1}{\sqrt{1 + (2\pi f \Re C)^2}} e^{-i\theta(f)}, \qquad (1.63)$$

где $\theta(f) = \operatorname{arctg} 2\pi f \Re C$. Амплитудная характеристика $|H(f)|$ и фазовая характеристика $\theta(f)$ изображены на рис. 1.13, *б*, *в*. Ширина полосы фильтра нижних частот определяется в точке половинной мощности; эта точка представляет собой частоту, на которой мощность выходного сигнала равна половине максимального значения, или частоту, на которой амплитуда выходного напряжения равна $1/\sqrt{2}$ максимального значения.

В общем случае точка половинной мощности выражается в децибелах (дБ) как точка –3 дБ, или точка, находящаяся на 3 дБ ниже максимального значения. По определению величина в децибелах определяется отношением мощностей, P_1 и P_2.

$$\text{величина в дБ} = 10\lg\frac{P_2}{P_1} = 10\lg\frac{V_2^2/\Re_2}{V_1^2/\Re_1} \qquad (1.64,а)$$

Здесь V_1 и V_2 — напряжения, а $\Re_1$ и $\Re_2$ — сопротивления. В системах связи для анализа обычно используется *нормированная мощность*; в этом случае сопротивления $\Re_1$ и $\Re_2$ считаются равными 1 Ом, тогда

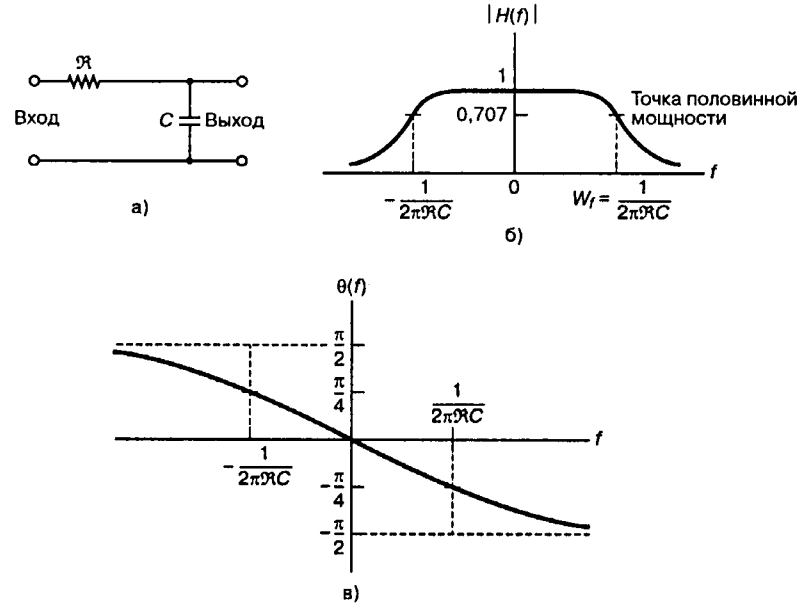


Рис. 1.13. $\Re C$-фильтр и его передаточная функция: а) $\Re C$-фильтр; б) амплитудная характеристика $\Re C$-фильтра; в) фазовая характеристика $\Re C$-фильтра

$$\text{величина в дБ} = 10\lg\frac{P_2}{P_1} = 10\lg\frac{V_2^2}{V_1^2} \qquad (1.64,б)$$

Амплитудный отклик можно выразить в децибелах как

$$|H(f)|_{\text{дБ}} = 20\lg\frac{V_2}{V_1} = 20\lg|H(f)|, \qquad (1.64,в)$$

где V_1 и V_2 — напряжения на входе и выходе, а сопротивления на входе и выходе предполагаются равными.

Из уравнения (1.63) легко проверить, что точка половинной мощности $\Re C$-фильтра нижних частот соответствует $\omega = 1/\Re C$ рад/с, или $f = 1/(2\pi\Re C)$ Гц. Таким образом, ширина полосы W_f в герцах равна $1/(2\pi\Re C)$. *Форм-фактор* фильтра — это мера того, насколько хорошо реальный фильтр аппроксимирует идеальный. Обычно он определяется как отношение ширины полос фильтров по уровню –60 дБ и –6 дБ. Достаточно малый форм-фактор (около 2) можно получить в пропускающем фильтре с

очень резким срезом. Для сравнения, форм-фактор простого $\Re C$-фильтра нижних частот составляет около 600.

Существует несколько полезных аппроксимаций характеристики идеального фильтра нижних частот. Одну из них дает фильтр Баттерворта, аппроксимирующий идеальный фильтр нижних частот функцией

$$|H_n(f)| = \frac{1}{\sqrt{1+(f/f_u)^{2n}}} \qquad n \geq 1, \tag{1.65}$$

где f_u — верхняя частота среза (−3 дБ), а n — *порядок* фильтра. Чем выше порядок, тем выше сложность и стоимость реализации фильтра. На рис. 1.14 показаны графики амплитуды $|H(f)|$ для нескольких значений n. Отметим, что по мере роста n амплитудные характеристики приближаются к характеристикам идеального фильтра. Фильтры Баттерворта популярны из-за того, что они являются лучшей аппроксимацией идеального случая в смысле *максимальной пологости* полосы пропускания фильтра.

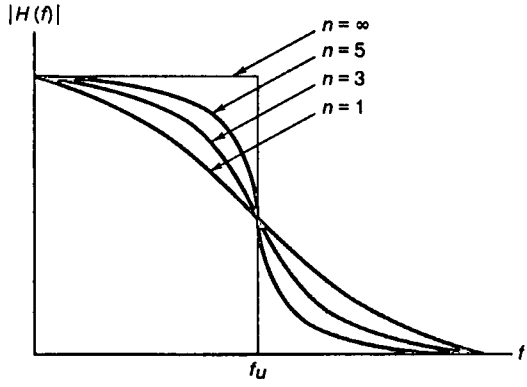


Рис. 1.14. Амплитудный отклик фильтра Баттерворта

Пример 1.3. Прохождение белого шума через $\Re C$-фильтр

Белый шум со спектральной плотностью $G_n(f) = N_0/2$, показанной на рис. 1.8, *а*, подается на вход $\Re C$-фильтра, показанного на рис. 1.13, *а*. Найдите спектральную плотность мощности $G_Y(f)$ и автокорреляционную функцию $R_Y(\tau)$ сигнала на выходе.

Решение

$$G_Y(f) = G_n(f)\,|H(f)|^2 =$$

$$= \frac{N_0}{2}\frac{1}{1+(2\pi f \Re C)^2}$$

$$R_Y(\tau) = \mathfrak{F}^{-1}\{G_Y(f)\}$$

Используя табл. А.1, находим Фурье-образ $G_Y(f)$.

$$R_Y(\tau) = \frac{N_0}{4\Re C}\exp\left(\frac{|\tau|}{\Re C}\right)$$

Как можно предположить, после фильтрации у нас уже не будет белого шума. $\Re C$-фильтр преобразовывает входную автокорреляционную функцию белого шума (определенную через дельта-функцию) в экспоненциальную функцию. Для узкополосного фильтра (большая величина $\Re C$) шум на выходе будет проявлять более высокую корреляцию между выборками шума через фиксированные промежутки времени, чем шум на выходе широкополосного фильтра.

1.6.4. Сигналы, каналы, спектры

Сигналы описываются через их спектр. Подобным образом сети или каналы связи описываются через их спектральные характеристики или частотные передаточные функции. Как ширина полосы сигнала влияет на результат передачи сигнала через фильтр? На рис. 1.15 показаны два случая, представляющие для нас практический интерес. На рис. 1.15, *а* (случай 1) входной сигнал имеет узкий спектр, а частотная передаточная функция фильтра является широкополосной. Из уравнения (1.48) видим, что спектр выходного сигнала представляет собой простое произведение этих двух спектров. Можно проверить (рис. 1.15, *а*), что перемножение двух спектральных функций дает спектр с шириной полосы, приблизительно равной меньшей из двух полос (когда одна из двух спектральных функций стремится к нулю, произведение также стремится к нулю). Следовательно, для случая 1 спектр выходного сигнала ограничен спектром входного сигнала. Подобным образом в случае 2, где входной сигнал является широкополосным, а фильтр имеет узкополосную передаточную функцию (рис. 1.15, *б*), видим, что ширина полосы выходного сигнала ограничена шириной полосы фильтра; выходной сигнал будет профильтрованным (искаженным) изображением входного сигнала.

Воздействие фильтра на сигнал также можно рассматривать во временной области. Выход $y(t)$, получаемый в результате свертки идеального входного импульса $x(t)$ (имеющего амплитуду V_m и ширину импульса T) с импульсным откликом $\Re C$-фильтра нижних частот, можно записать в следующем виде [8].

$$y(t) = \begin{cases} V_m(1 - e^{-t/\Re C}) & \text{для } 0 \le t \le T \\ V'_m\, e^{-(t-T)/\Re C} & \text{для } t > T \end{cases}, \tag{1.66}$$

где

$$V'_m = V_m(1 - e^{-T/\Re C}) \tag{1.67}$$

Определим ширину импульса как

$$W_p = \frac{1}{T}, \tag{1.68}$$

а ширину полосы $\Re C$-фильтра как

$$W_f = \frac{1}{2\pi\Re C}. \tag{1.69}$$

Идеальный входной импульс $x(t)$ и его амплитудный спектр $|X(f)|$ показаны на рис. 1.16. $\Re C$-фильтр и его амплитудная характеристика $|H(f)|$ показаны на рис. 1.13, *а*, *б*. На рис. 1.17 показаны три примера, в каждом из которых использованы уравнения (1.66)–(1.69). Пример 1 иллюстрирует случай $W_p << W_f$. Отме-

тим, что выходной отклик $y(t)$ является достаточно хорошим приближением исходного импульса $x(t)$, показанного пунктиром.

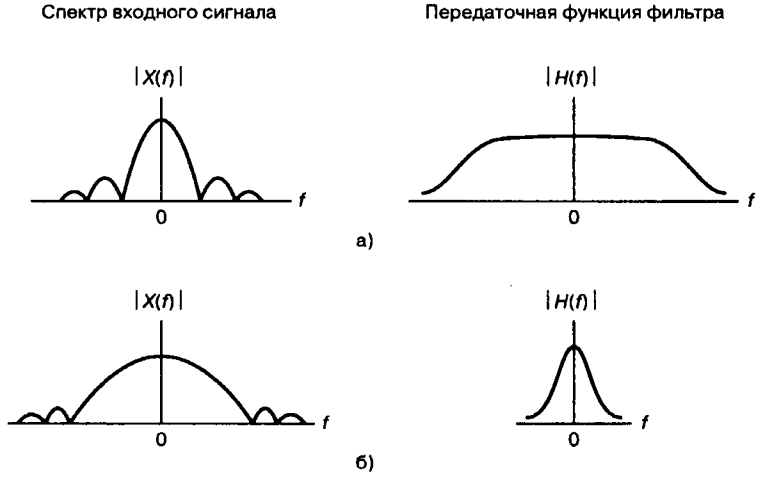


Рис. 1.15. Спектральные характеристики входного сигнала и вклад цепи в спектральные характеристики выходного сигнала: а) случай 1. Ширина выходной полосы ограничена шириной полосы входного сигнала; б) случай 2. Ширина выходной полосы ограничена шириной полосы фильтра

Этот случай является примером *хорошей точности воспроизведения*. В примере 2, где $W_p \approx W_f$, переданный импульс все еще можно распознать. Пример 3 иллюстрирует случай, когда $W_p >> W_f$. Видим, что по форме $y(t)$ импульс едва угадывается. Где может понадобиться большая ширина полосы (или хорошая точность воспроизведения), как в примере 1? Это могут быть *дистанционные приложения большой точности*, где на время прибытия импульса влияет расстояние, что требует импульсов с малыми временами нарастания. Какой пример демонстрирует двоичные приложения цифровой связи? *Пример 2.* Как указывалось ранее (рис. 1.1), одной из принципиальных особенностей двоичной цифровой связи является то, что требуется всего лишь точно *почувствовать*, к какому из двух возможных состояний принадлежит каждый принятый импульс. Пример 3 был включен для полноты обсуждения; в реальных системах подобные схемы не используются.

1.7. Ширина полосы при передаче цифровых данных

1.7.1. Узкополосные и широкополосные сигналы

Легким способом трансляции спектра низкочастотного или узкополосного сигнала $x(t)$ на более высокую частоту является умножение узкополосного сигнала на несущий сигнал $\cos 2\pi f_c t$ или *наложение колебаний*, как показано на рис. 1.18. Результирующий сигнал $x_c(t)$ называется *двухполосным* (double sideband — DSB) *модулированным сигналом* и выражается следующей формулой.

$$x_c(t) = x(t) \cos 2\pi f_c t \tag{1.70}$$

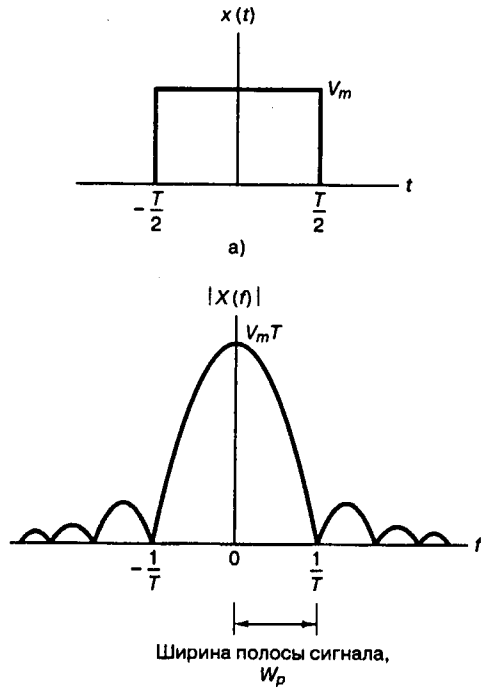


Рис. 1.16. Идеальный импульс и его амплитудный спектр

Из теоремы о модуляции (см. раздел А.3.2) спектр двухполосного сигнала $x_c(t)$ дается следующим выражением.

$$X_c(f) = \tfrac{1}{2}\left[X(f - f_c) + X(f + f_c)\right] \tag{1.71}$$

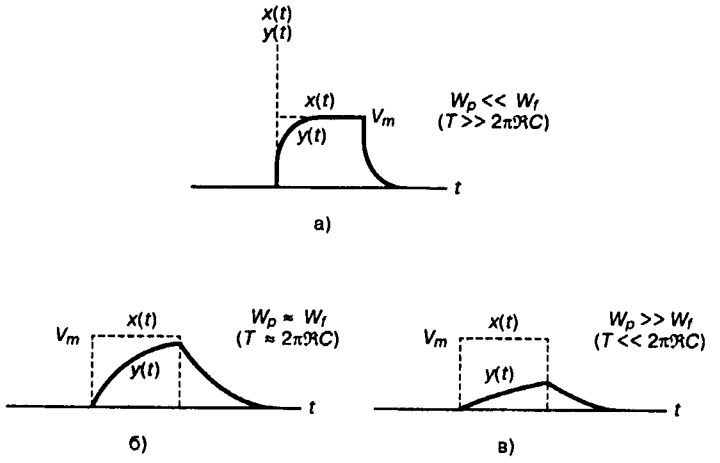


Рис. 1.17. Три примера фильтрации идеального импульса: а) пример 1. Хорошая точность воспроизведения; б) пример 2. Хорошее распознавание; в) пример 3. Плохое распознавание

Амплитудный спектр $|X(f)|$ узкополосного сигнала $x(t)$ с шириной полосы f_m и амплитудный спектр $|X_c(f)|$ двухполосного сигнала $x_c(t)$ с шириной полосы W_{DSB} показаны на рис. 1.18, *б, в*. На графике $|X_c(f)|$ спектральные компоненты, соответствующие положительным частотам узкополосного сигнала, находятся в диапазоне от f_c до $(f_c + f_m)$. Эта часть спектра двухполосного сигнала называется *верхней боковой полосой* (upper sideband — USB). Спектральные компоненты, соответствующие отрицательным частотам узкополосного сигнала, лежат в диапазоне от $(f_c - f_m)$ до f_c. Эта часть спектра двухполосного сигнала называется *нижней боковой полосой* (lower sideband — LSB). Кроме того, в области отрицательных частот находятся зеркальные изображения спектров нижней и верхней боковых полос. *Несущая волна* (или просто *несущая*) иногда еще называется *гетеродином*, или *местным гетеродином*. В общем случае частота несущей значительно больше ширины полосы узкополосного сигнала.

$$f_c \geq f_m$$

С помощью рис. 1.18 можно легко сравнить ширину полосы f_m, требуемую для передачи узкополосного сигнала, с шириной полосы W_{DSB}, достаточной для передачи двухполосного сигнала. Итак, видим следующее.

$$W_{DSB} = 2f_m \tag{1.72}$$

Иными словами, для передачи двухполосной версии сигнала нам необходима вдвое большая полоса, чем для передачи его узкополосного аналога.

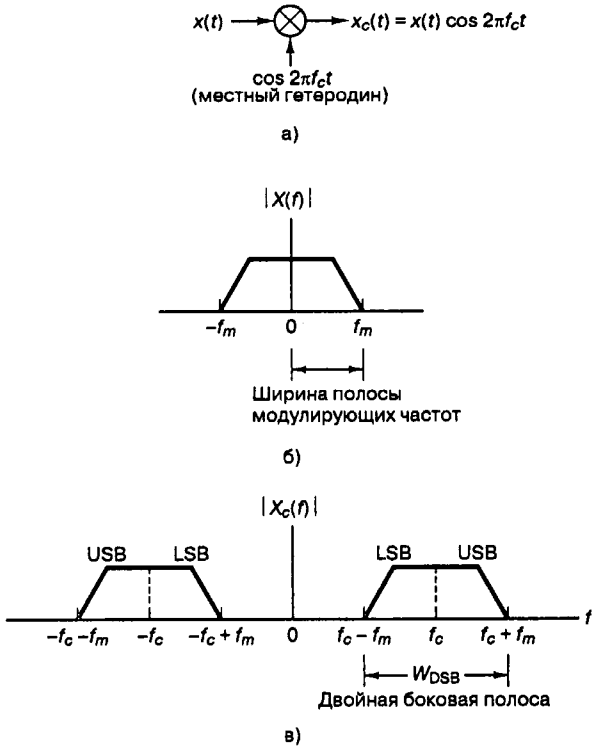


Рис. 1.18. Сравнение узкополосного и двухполосного спектров: а) наложение колебаний; б) узкополосный спектр; в) двухполосный спектр

1.7.2. Дилемма при определении ширины полосы

Множество важных теорем из теории связи и информации опираются на предположение о том, что каналы имеют *строго ограниченную полосу*; это означает, что за пределами определенной полосы мощность сигнала равна нулю. Таким образом, мы сталкиваемся с дилеммой: сигналы со строго ограниченной полосой, как, например, сигнал со спектром $|X_1(f)|$, изображенный на рис. 1.19, *б*, не могут быть реализованы, поскольку они, как показано на рис. 1.19, *а*, подразумевают сигналы бесконечной длительности (обратное преобразование Фурье функции $|X_1(f)|$). Сигналы с ограниченной длительностью, как сигнал $x_2(t)$, показанный на рис. 1.19, *в*, легко реализуются. Но при этом они также непригодны, поскольку их Фурье-образы имеют энергию на относительно высоких частотах, что можно увидеть из спектра сигнала $|X_2(f)|$, показанного на рис. 1.19, *г*. Итак, можно сказать, что для всех спектров с ограниченной полосой сигналы не реализуемы, а для всех реализуемых сигналов абсолютная ширина полосы равна бесконечности. Математическое описание реального сигнала не допускает, чтобы сигнал был строго ограничен по продолжительности и полосе. Значит, математические модели являются абстракциями; поэтому не удивительно, что до настоящего момента не существует единого определения ширины полосы.

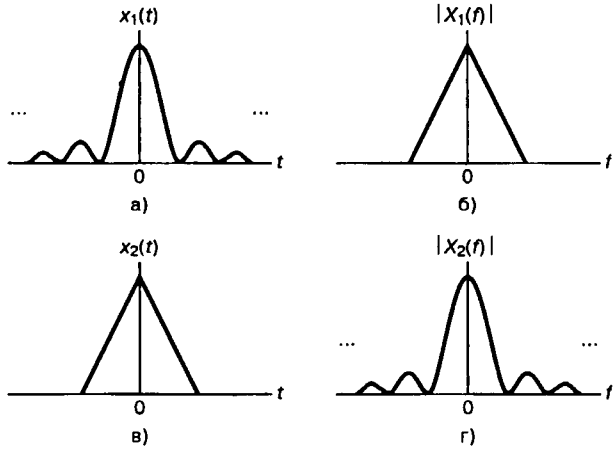


Рис. 1.19. Представление сигнала: а) сигнал со строго ограниченной полосой во временной области; б) в частотной области; в) сигнал со строго ограниченной длительностью во временной области; г) в частотной области

Все критерии определения ширины полосы имеют одно общее свойство: они пытаются найти меру ширины, W, неотрицательной действительной спектральной плотности, определенную для всех частот $|f| < \infty$. На рис 1.20 показаны некоторые наиболее распространенные определения ширины полосы (стоит отметить, что различные критерии не являются взаимозаменяемыми). Однополосная спектральная плотность мощности для отдельного гетеродинного импульса $x_c(t)$ имеет следующее аналитическое выражение.

$$G_x(f) = T\left[\frac{\sin \pi(f - f_c)\, T}{\pi(f - f_c)\, T}\right]^2, \tag{1.73}$$

где f_c — частота несущей, а T — длительность импульса. Эта спектральная плотность мощности (рис. 1.20) также характеризует *случайную последовательность импульсов*; предполагается, что время, по которому производится усреднение, намного больше длительности импульса. График состоит из основного лепестка и меньших симметричных боковых лепестков. Общий вид графика справедлив для большинства форматов цифровых модуляций; в то же время некоторые форматы не имеют ярко выраженных боковых лепестков. Перечислим критерии определения ширины полосы, показанные на рис. 1.20.

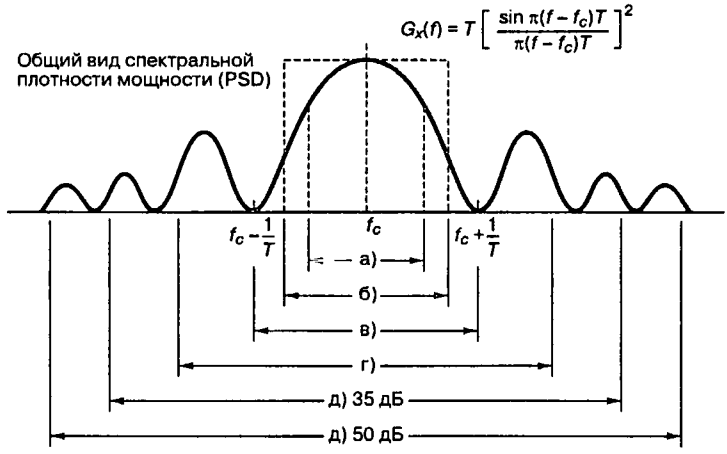


Рис. 1.20. Ширина полосы цифровых данных: а) половинная мощность; б) шумовой эквивалент; в) по первым нулям; г) 99% *мощности; д) ограниченная спектральная плотность мощности по уровню* 35 *дБ и* 50 *дБ*

а) *ширина полосы половинной мощности.* Интервал между частотами, на которых $G_x(f)$ падает до мощности, вдвое (или на 3 дБ) ниже максимального значения.

б) *ширина полосы шумового эквивалента.* Шумовой эквивалент полосы позволяет быстро вычислять мощность шума на выходе усилителя с широкополосным шумом на входе; данное понятие применимо и к ширине полосы сигнала. Ширина полосы шумового эквивалента W_N определяется отношением $W_N = P_x/G_x(f_c)$, где P_x — общая мощность сигнала на всех частотах, а $G_x(f_c)$ — значение $G_x(f)$ в центре полосы (предполагается, что это — максимальное значение по всем частотам).

в) *ширина полосы по первым нулям.* Наиболее популярной мерой ширины полосы в цифровой связи является ширина основного спектрального лепестка, в котором сосредоточена основная мощность сигнала. Этому критерию недостает универсальности, поскольку в некоторых форматах модуляции отсутствуют явно выраженные лепестки.

г) *полоса, вмещающая определенную часть суммарной мощности.* Этот критерий ширины полосы был принят Федеральной комиссией по средствам связи США (Federal Communications Commission — FCC) (см. *FCC Rules and Regulations*, раздел 2.202), и согласно ему полоса ограничивается так, что за ее пределами находится 1% мощности сигнала (0,5% выше верхней границы полосы и 0,5% ниже нижней границы). Таким образом, на определенную полосу приходится 99% мощности сигнала.

д) *спектральная плотность мощности по уровню x дБ.* Еще один популярный метод определения ширины полосы — указать, что за пределами определенной полосы мощность $G_x(f)$ должна снизиться до заданного уровня, меньшего максимального значения (в центре полосы). Типичными уровнями затухания являются 35 и 50 дБ.

е) *абсолютная ширина полосы.* Это интервал между частотами, вне которых спектр равен нулю. Весьма полезная абстракция. Впрочем, для всех реализуемых сигналов абсолютная ширина полосы равна бесконечности.

Пример 1.4. Сигналы со строго ограниченной полосой

Понятие сигнала, который строго ограничен полосой частот, нереализуемо. Докажите это, показав, что сигнал со *строго ограниченной полосой* должен иметь *бесконечную* длительность.

Решение

Пусть $x(t)$ — сигнал с Фурье-образом $X(f)$ и строго ограниченной полосой частот, центрированный на частотах $\pm f_c$ и имеющий ширину $2W$. $X(f)$ можно выразить через передаточную функцию идеального фильтра $H(f)$, показанную на рис. 1.21, как

$$H(f) = X'(f)H(f), \tag{1.74}$$

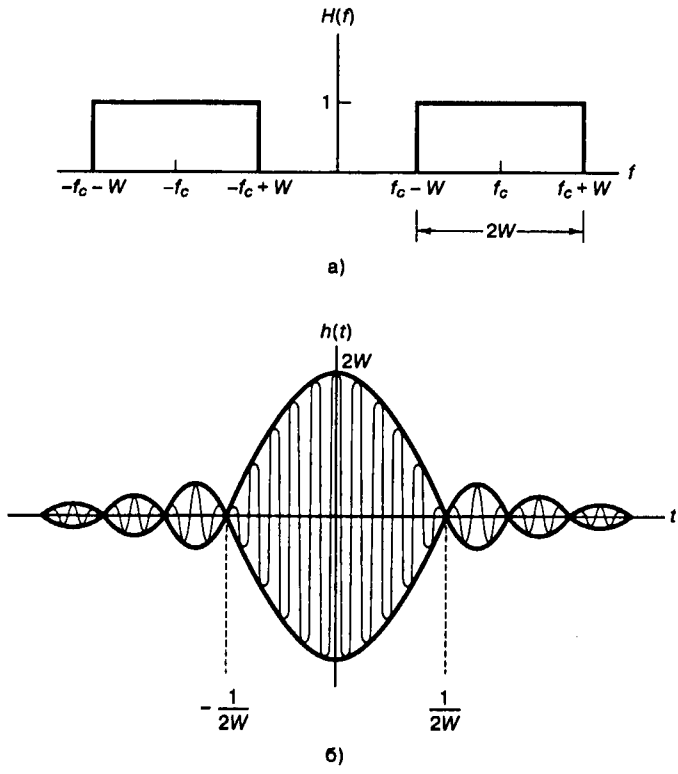


Рис. 1.21. Передаточная функция и импульсная характеристика для сигнала со строго ограниченной полосой: а) идеальный полосовой фильтр; б) идеальная полосовая импульсная характеристика

где $X'(f)$ — Фурье-образ сигнала $x'(t)$, не обязательно имеющий ограниченную ширину полосы, и

$$H(f) = \text{rect}\left(\frac{f - f_c}{2W}\right) + \text{rect}\left(\frac{f + f_c}{2W}\right), \tag{1.75}$$

где

$$\text{rect}\left(\frac{f}{2W}\right) = \begin{cases} 1 & \text{для } -W < f < W \\ 0 & \text{для } |f| > W \end{cases}$$

$X(f)$ можно выразить через $X'(f)$ как

$$X(f) = \begin{cases} X'(f) & \text{для } (f_c - W) \le |f_c| \le (f_c + W) \\ 0 & \text{для остальных } f \end{cases}$$

Умножение в частотной области, как показано в уравнении (1.74), преобразуется в свертку во временной области.

$$x(t) = x'(t) * h(t) \tag{1.76}$$

Здесь $h(t)$ — результат применения обратного преобразования Фурье к функции $H(f)$, который можно записать следующим образом (см. табл. А.1 и А.2).

$$h(t) = 2W (\text{sinc } 2Wt) \cos 2\pi f_c t$$

Вид $h(t)$ показан на рис. 1.21, *б*. Отметим, что $h(t)$ имеет *бесконечную длительность*. Следовательно, сигнал $x(t)$, полученный, как показывает уравнение (1.76), путем свертки $x'(t)$ с $h(t)$, также имеет бесконечную длительность и, следовательно, *не может быть реализован*.

1.8. Резюме

В данной главе намечены цели книги и определена основная терминология. Здесь рассмотрены фундаментальные понятия изменяющихся во времени сигналов, такие как классификация, спектральная плотность и автокорреляция. Кроме того, описаны случайные сигналы, статистически и спектрально охарактеризован белый гауссов шум, для большинства систем связи представляющий собой первичную модель шума. В заключение рассмотрен важный вопрос передачи сигнала через линейные системы и изучены некоторые важные аппроксимации идеального случая. Установлено, что понятие абсолютной ширины полосы является абстракцией и что в реальном мире мы сталкиваемся с необходимостью выбора определения ширины полосы, подходящего для конкретного случая. В последующих главах книги, согласно схеме, приведенной в начале главы, будут рассмотрены все этапы обработки сигналов, введенные в данной главе.

Литература

1. Haykin S. *Communication Systems.* John Wiley & Sons, Inc., New York, 1983.
2. Shanmugam K. S. *Digital and Analog Communication Systems.* John Wiley & Sons, Inc., New York, 1979.
3. Papoulis A. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill Book Company, New York, 1965.
4. Johnson J. B. *Thermal Agitation of Electricity in Conductors. Phys. Rev.*, vol. 32, July 1928, pp. 97–109.
5. Nyguist H. *Thermal Agitation of Electric Charge in Conductors. Phys. Rev.*, vol. 32, July 1928, pp. 110–113.
6. Van Trees H. L. *Detection, Estimation, and Modulation Theory.* Part 1, John Wiley & Sons, New York, 1968.

7. Schwartz M. *Information Transmission, Modulation, and Noise.* McGraw-Hill Book Company, New York, 1970.
8. Millman J. and Taub H. *Pulse, Digital, and Switching Waveforms.* McGraw-Hill Book Company, New York, 1965.

Задачи

1.1. Определите, в каком представлении даны следующие сигналы: в энергетическом или мощностном. Найдите нормированную энергию и нормированную мощность каждого сигнала.

а) $x(t) = A \cos 2\pi f_0 t$ для $-\infty < t < \infty$

б) $x(t) = \begin{cases} A \cos 2\pi f_0 t & \text{для } -T_0/2 \le t \le T_0/2, \text{ где } T_0 = 1/f_0 \\ 0 & \text{для остальных } t \end{cases}$

в) $x(t) = \begin{cases} A \exp(-at) & \text{для } t > 0, a > 0 \\ 0 & \text{для остальных } t \end{cases}$

г) $x(t) = \cos t + 5 \cos 2t$ для $-\infty < t < \infty$

1.2. Определите спектральную плотность энергии квадратного импульса $x(t) = \text{rect}\,(t/T)$, где функция rect (t/T) равна 1 для $-T/2 \le t \le T/2$ и нулю — для остальных t. Вычислите нормированную энергию E_x импульса.

1.3. Выразите среднюю нормированную мощность периодического сигнала через коэффициенты комплексного ряда Фурье.

1.4. Используя усреднение по времени, найдите среднюю нормированную мощность сигнала $x(t) = 10 \cos 10t + 20 \cos 20t$.

1.5. Решите задачу 1.4 посредством суммирования спектральных коэффициентов.

1.6. Определите, какие из перечисленных функций (если такие есть) имеют свойства автокорреляционных функций. Ответ аргументируйте. (*Примечание:* $\mathfrak{F}\{R(\tau)\}$ должна быть неотрицательной функцией. Почему?)

а) $x(\tau) = \begin{cases} 1 & \text{для } -1 \le \tau \le 1 \\ 0 & \text{для остальных } \tau \end{cases}$

б) $x(\tau) = \delta(\tau) + \sin 2\pi f_0 \tau$

в) $x(\tau) = \exp(|\tau|)$

г) $x(\tau) = 1 - |\tau|$ — для $-1 \le \tau \le 1$ и 0 — для остальных

1.7. Определите, какие из перечисленных функций (если такие есть) имеют свойства функций спектральной плотности мощности. Ответ аргументируйте.

а) $X(f) = \delta(f) + \cos^2 2\pi f$

б) $X(f) = 10 + \delta(f - 10)$

в) $X(f) = \exp(-2\pi|f - 10|)$

г) $X(f) = \exp[-2\pi(f^2 - 10)]$

1.8. Выразите автокорреляционную функцию $x(t) = A \cos(2\pi f_0 t + \varphi)$ через ее период $T_0 = 1/f_0$. Найдите среднюю нормированную мощность $x(t)$, используя соотношение $P_x = R(0)$.

1.9. а) Используя результаты задачи 1.8, найдите автокорреляционную функцию $R(\tau)$ сигнала $x(t) = 10 \cos 10t + 20 \cos 20t$.

б) Используя соотношение $P_x = R(0)$, найдите среднюю нормированную мощность сигнала $x(t)$. Сравните ответ с ответами задач 1.4 и 1.5.

1.10. Для функции $x(t) = 1 + \cos 2\pi f_0 t$ вычислите (а) среднее значение $x(t)$; (б) мощность переменной составляющей $x(t)$; (в) среднеквадратическое значение $x(t)$.

1.11. Рассмотрим случайный процесс, описываемый функцией $X(t) = A\cos(2\pi f_0 t + \varphi)$, где A и f_0 — константы, а φ — случайная переменная, равномерно распределенная на промежутке $(0, 2\pi)$. Если $X(t)$ является эргодическим процессом, среднее по времени от $X(t)$ в пределе $t \to \infty$ равно соответствующему среднему по ансамблю от $X(t)$.

а) Используя усреднение по времени целого числа периодов, вычислите приближенно первый и второй моменты $X(t)$.

б) Используя уравнения (1.26) и (1.28), приближенно вычислите средние по ансамблю значения первого и второго моментов $X(t)$. Сравните результаты с ответом на п. а.

1.12. Фурье-образ сигнала $x(t)$ определяется формулой $X(f) = \operatorname{sinc} f$ (функция sinc определена в уравнении (1.39)). Найдите автокорреляционную функцию $R_x(\tau)$ сигнала $x(t)$.

1.13. Используя свойства дельта-функции, вычислите следующие интегралы.

а) $\int_{-\infty}^{\infty} \cos 6t\,\delta(t-3)\,dt$

б) $\int_{-\infty}^{\infty} 10\,\delta(t)(1+t)^{-1}\,dt$

в) $\int_{-\infty}^{\infty} \delta(t+4)(t^2+6t+1)\,dt$

г) $\int_{-\infty}^{\infty} \exp(-t^2)\,\delta(t-2)\,dt$

1.14. Найдите свертку $X_1(f) * X_2(f)$ для спектров, показанных на рис. 31.1.

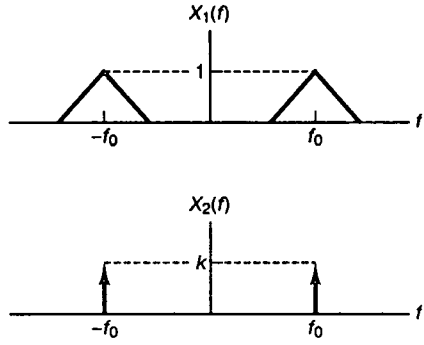


Рис. 31.1

1.15. На рис. 31.2 показана двусторонняя спектральная плотность мощности, $G_x(f) = 10^{-6} f^2$, сигнала $x(t)$.

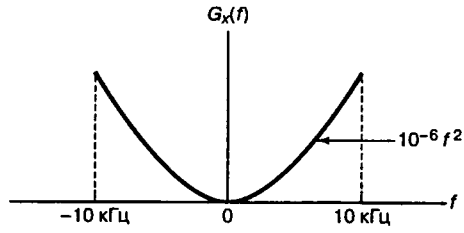


Рис. 31.2

а) Найдите нормированную среднюю мощность $x(t)$ в диапазоне частот от 0 до 10 кГц.

б) Найдите нормированную среднюю мощность $x(t)$ в диапазоне частот от 5 до 6 кГц.

1.16. Как показано в уравнении (1.64,а), децибелы — это логарифмическая мера *отношения мощностей*. Иногда в децибелах выражаются немощностные характеристики (относительно некоторых выделенных единиц). В качестве примера вычислите, сколько децибелов мяса для бифштексов вы приобретете, чтобы в группе из 100 человек каждый получил 2 гамбургера. Предположим, что в качестве эталонной единицы вы и мясник договорились использовать полфунта мяса (вес одного бифштекса).

1.17. Рассмотрим амплитудный отклик фильтра Баттерворта нижних частот в форме, приведенной в уравнении (1.65).

а) Найдите n, при котором $|H(f)|^2$ колеблется в пределах ±1 дБ в диапазоне $|f| \le 0{,}9f_u$.

б) Покажите, что при n, стремящемся к бесконечности, амплитудный отклик приближается к амплитудному отклику идеального фильтра.

1.18. Рассмотрим сеть, приведенную на рис. 1.9, частотная передаточная функция которой равна $H(f)$. На вход подается импульс $\delta(t)$. Покажите, что отклик $y(t)$ на выходе представляет собой результат обратного преобразования Фурье $H(f)$.

1.19. На рис. 31.3 приведен пример *цепи запоминания*, широко используемой в импульсных системах. Определите импульсную характеристику этого канала.

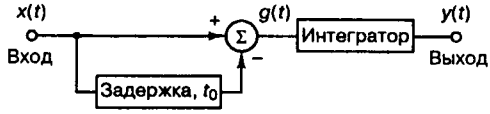


Рис. 31.3

1.20. Для спектра

$$G_x(f) = 10^{-4}\left\{\frac{\sin\left[\pi(f-10^6)10^{-4}\right]}{\pi(f-10^6)10^{-4}}\right\}^2$$

определите ширину полосы сигнала, используя следующие определения ширины полосы:

а) ширина полосы половинной мощности;

б) ширина полосы шумового эквивалента;

в) ширина полосы по первым нулям;

г) полоса, вмещающая 99% мощности (*подсказка:* используйте численные методы);

д) полоса по уровню 35 дБ;

е) абсолютная ширина полосы.

Вопросы для самопроверки

1.1. Как график автокорреляционной функции сигнала характеризует занятость полосы сигнала (см. раздел 1.5.4)?

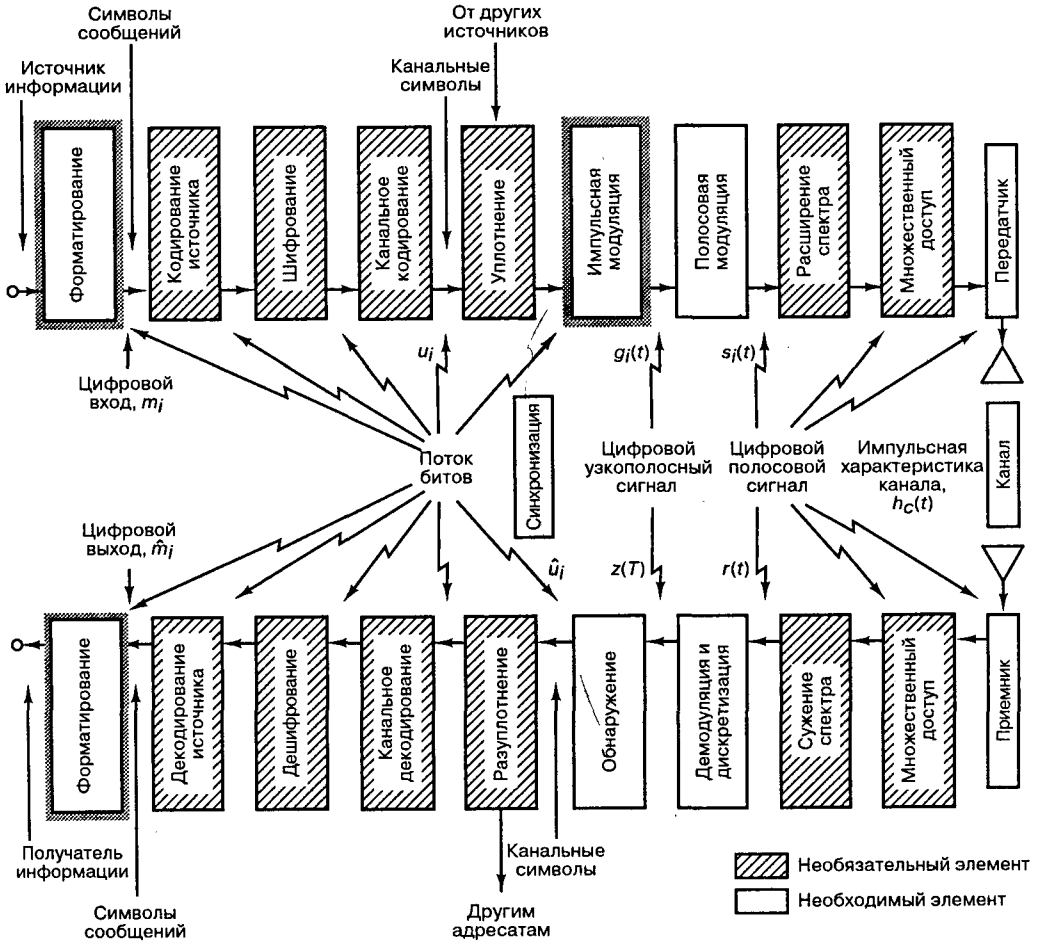
1.2. Какие два требования необходимо удовлетворить для обеспечения передачи без искажения через линейную систему (см. раздел 1.6.3)?

1.3. Дайте определение параметру *групповая задержка* (см. раздел 1.6.3).

1.4. Какая математическая дилемма является причиной существования нескольких определений ширины полосы (см. раздел 1.7.2)?

ГЛАВА 2

Форматирование и узкополосная модуляция



Задачей первого необходимого этапа обработки сигнала, *форматирования* (formatting), является обеспечение совместимости сообщения (или исходного сигнала) со средствами цифровой обработки. *Форматирование с целью передачи* — это преобразование исходной информации в цифровые символы. (В канале приема происходит обратное преобразование.) Если помимо форматирования применяется сжатие данных, процесс называется *кодированием источника* (source coding). Некоторые авторы считают форматирование частным случаем кодирования источника. В данной главе мы рассмотрим форматирование и узкополосную модуляцию, а в главе 13 обсудим кодирование источника как частный случай *эффективного описания* исходной информации.

Обратимся к рис. 2.1, где выделенный блок "Форматирование" перечисляет действия, связанные с преобразованием информации в цифровые сообщения. Считается, что цифровые сообщения имеют логический формат двоичных нулей и единиц и с целью передачи проходят этап импульсной модуляции, в результате чего преобразуются в *узкополосные* (импульсные) сигналы. Затем эти сигналы могут передаваться по каналу передачи данных. Выделенный на рис. 2.1 блок "Узкополосная передача сигналов" содержит перечень импульсно-модулированных сигналов, которые описываются в данной главе. Вообще, термин "узкополосный" (baseband) определяет сигнал, спектр которого начинается от (или около) постоянной составляющей и заканчивается некоторым конечным значением, обычно не более нескольких мегагерц. Обсуждение этой темы мы продолжим в главе 3, где больше внимания будет уделено демодуляции и обнаружению.

2.1. Узкополосные системы

На рис. 1.2 была изображена блочная диаграмма типичной системы цифровой связи. На рис. 2.2 представлен вариант этой диаграммы, в котором выделяются этапы форматирования и передачи *узкополосных* сигналов. Данные, уже имеющие цифровой формат, могут не проходить через этап форматирования. Текстовая информация преобразовывается в двоичные цифры с помощью кодера (coder). Аналоговая информация форматируется с использованием трех отдельных процессов: дискретизации (sampling), квантования (quantization) и кодирования (coding). Во всех случаях после форматирования получается последовательность двоичных цифр.

Цифры необходимо передать через *узкополосный канал*, такой как пара проводников или коаксиальный кабель. При этом никакой канал использовать нельзя, пока двоичные цифры не будут преобразованы в *сигналы*, совместимые с этим каналом. Для узкополосных каналов такими совместимыми сигналами являются импульсы.

На рис. 2.2 преобразование потока битов в последовательность импульсных сигналов происходит в блоке "Импульсная модуляция". На выходе модулятора получим последовательность импульсов, характеристики которых соответствуют характеристикам цифр, поданных на вход. После передачи по каналу импульсные сигналы восстанавливаются (демодулируются) и проходят этап обнаружения; целью последнего этапа, (обратного) форматирования, является восстановление (с определенной степенью точности) исходной информации.

Рис. 2.1. Основные преобразования цифровой связи

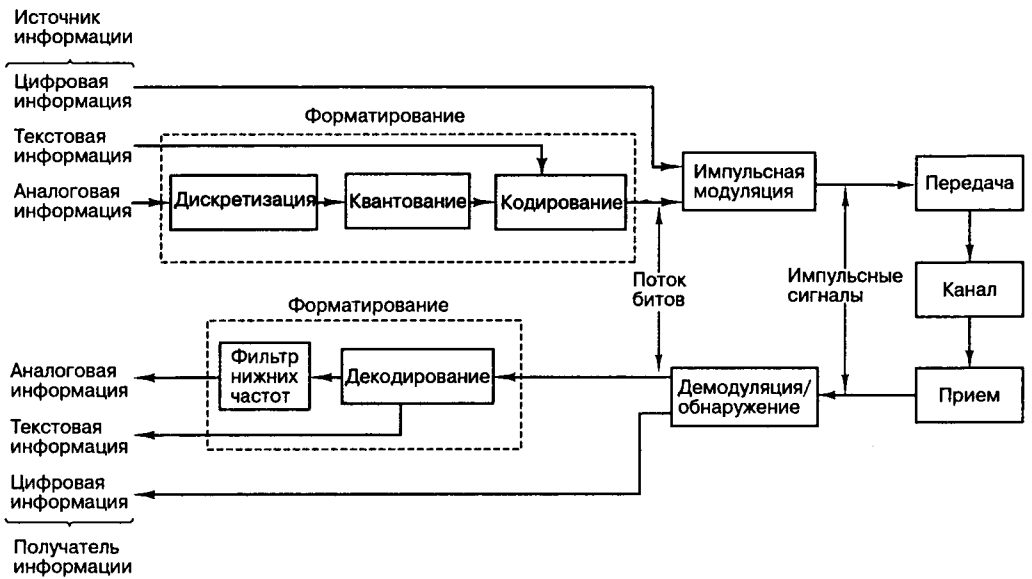


Рис. 2.2. Форматирование и передача узкополосных сигналов

2.2. Форматирование текстовой информации (знаковое кодирование)

Изначально большинство передаваемой информации (исключением является только информационный обмен между двумя компьютерами) имеет текстовую или аналоговую форму. Если информация является буквенно-цифровым текстом, то используется один из нескольких стандартных форматов — методов знакового кодирования: ASCII (American Standard Code for Information Interchange — Американский стандартный код для обмена информацией), EBCDIC (Extended Binary Coded Decimal Interchange Code — расширенный двоичный код обмена информацией), код Бодо, код Холлерита и др. Таким образом, текстовый материал преобразовывается в цифровой формат. На рис. 2.3 показан формат ASCII, а на рис. 2.4 — формат EBCDIC. Двоичные числа определяют порядок последовательной передачи, причем двоичная единица является первой сигнальной посылкой. Знаковое кодирование, следовательно, является этапом преобразования текста в двоичные цифры (биты). Иногда существующие знаковые коды модифицируются для удовлетворения специфических требований. Например, 7-битовый код ASCII (рис. 2.3) может включать дополнительный бит, облегчающий выявление ошибок (см. главу 6). С другой стороны, иногда код укорачивается до 6-битовой версии, кодирующей только 64 знака, а не 128, как 7-битовый код ASCII.

2.3. Сообщения, знаки и символы

Текстовые сообщения состоят из последовательности буквенно-цифровых знаков. При цифровой передаче знаки вначале кодируются в последовательность битов, которая называется *потоком битов*, или *узкополосным сигналом*. После этого формируются группы из k бит, именуемые *символами*, причем число всех символов конечно ($M = 2^k$), а их совокупность называется *алфавитом*. Система, использующая символьный набор размера M, называется *M-арной*. Выбор величины k или M есть важным первоначальным этапом проектирования любой цифровой системы связи. При $k = 1$ система является *бинарной*, размер набора символов равен $M = 2$, а модулятор использует один из двух различных сигналов для представления двоичного значения "один", а другой — для представления двоичного значения "нуль". В этом частном случае символ и бит — это одно и то же. При $k = 2$ система именуется *четверичной*, или *4-уровневой* ($M = 4$). В каждый момент формирования символа модулятор использует один из четырех возможных сигналов для представления символа. Разделение последовательности битов сообщения определяется размером алфавита M. Ниже приведен пример, который поможет лучше понять связь между следующими терминами: "сообщение", "знак", "символ", "бит" и "цифровой сигнал".

Биты				5	0	1	0	1	0	1	0	1
				6	0	0	1	1	0	0	1	1
1	2	3	4	7	0	0	0	0	1	1	1	1
0	0	0	0		NUL	DLE	SP	0	@	P	'	p
1	0	0	0		SOH	DC1	!	1	A	Q	a	q
0	1	0	0		STX	DC2	"	2	B	R	b	r
1	1	0	0		ETX	DC3	#	3	C	S	c	s
0	0	1	0		EOT	DC4	$	4	D	T	d	t
1	0	1	0		ENQ	NAK	%	5	E	U	e	u
0	1	1	0		ACK	SYN	&	6	F	V	f	v
1	1	1	0		BEL	ETB	'	7	G	W	g	w
0	0	0	1		BS	CAN	(	8	H	X	h	x
1	0	0	1		HT	EM	)	9	I	Y	i	y
0	1	0	1		LF	SUB	*	:	J	Z	j	z
1	1	0	1		VT	ESC	+	;	K	[	k	{
0	0	1	1		FF	FS	,	<	L	\	l	|
1	0	1	1		CR	GS	-	=	M	]	m	}
0	1	1	1		SO	RS	.	>	N	^	n	~
1	1	1	1		SI	US	/	?	O	_	o	DEL

NUL Пустой символ, или все нули
SOH Символ начала заголовка
STX Символ начала текста
ETX Символ конца текста
EOT Символ конца передачи
ENQ Символ запроса
ACK Символ подтверждения приема
BEL Символ звуковой сигнализации
BS Символ возврата на позицию
HT Символ горизонтальной табуляции
LF Символ перевода строки
VT Символ вертикальной табуляции
FF Символ перевода страницы
CR Символ возврата каретки
SO Символ расширения кода
SI Символ восстановления кода
DLE Символ переключения
DC1 Символ управления устройством 1
DC2 Символ управления устройством 2
DC3 Символ управления устройством 3
DC4 Символ управления устройством 4
NAK Символ отрицательного подтверждения
SYN Символ синхронизации
ETB Символ конца передачи
CAN Символ аннулирования
EM Символ конца носителя
SUB Символ замены
ESC Символ переключения кода
FS Символ разделения файлов
GS Символ разделения групп
RS Символ разделения записей
US Символ разделения элементов
SP Символ пробела
DEL Удаление

Рис. 2.3. Семибитовый код ASCII

Биты				5	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
				6	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
				7	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
				8	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
1	2	3	4																	
0	0	0	0		NUL	SOH	STX	ETX	PF	HT	LC	DEL			SMM	VT	FF	CR	SO	SI
0	0	0	1		DLE	DC1	DC2	DC3	RES	NL	BS	IL	CAN	EM	CC		IFS	IGS	IRS	IUS
0	0	1	0		DS	SOS	FS		BYP	LF	EOB	PRE			SM			ENQ	ACK	BEL
0	0	1	1				SYN		PN	RS	US	EOT					DC4	NAK		SUB
0	1	0	0		SP										¢		<	(	+	!
0	1	0	1		&										!	$	*	)	;	¬
0	1	1	0		-	/	\									,	%	_	>	?
0	1	1	1												:	#	@	'	=	"
1	0	0	0			a	b	c	d	e	f	g	h	i						
1	0	0	1			j	k	l	m	n	o	p	q	r						
1	0	1	0				s	t	u	v	w	x	y	z						
1	0	1	1																	
1	1	0	0			A	B	C	D	E	F	G	H	I						
1	1	0	1			J	K	L	M	N	O	P	Q	R						
1	1	1	0				S	T	U	V	W	X	Y	Z						
1	1	1	1		0	1	2	3	4	5	6	7	8	9						

PF Символ отмены перфорации
HT Символ горизонтальной табуляции
LC Символ нижнего регистра
DEL Символ удаления
SP Символ пробела
UC Символ верхнего регистра
RES Символ восстановления
NL Символ новой строки

BS Символ возврата на позицию
IL Холостой символ
PN Символ перфорации
EOT Символ конца передачи
BYP Символ обхода
LF Символ перевода строки
EOB Символ конца блока
PRE Символ переключения кода

RS Символ разделения записей
SM Символ начала сообщения
DS Символ выбора цифры
SOS Символ начала значащих цифр
IFS Символ разделения файлов обмена
IGS Символ разделения групп обмена
IRS Символ разделения записей обмена
IUS Символ разделения блоков обмена

Остальные символы те же, что и в ASCII

Рис. 2.4. Кодовая таблица знаков EBCDIC

2.3.1. Пример сообщений, знаков и символов

На рис. 2.5 приведен пример разделения потока битов, определяемого спецификацией системы для различных значений k и M. Текстовое сообщение на рисунке — это слово "THINK". Использование 6-битовой кодировки ASCII (биты 1–6 на рис. 2.3) дает поток битов, состоящий из 30 бит. На рис. 2.5, *а* размер набора символов, M, был выбран равным 8 (каждый символ представляет восьмеричное число). Таким образом, биты группируются по три ($k = \log_2 8$); полученные в результате 10 чисел представляют 10 готовых к передаче восьмеричных символов. Передатчик должен иметь набор из восьми сигналов $s_i(t)$, где $i = 1, ..., 8$, сопоставляемых со всеми возможными символами, причем передача каждого сигнала возможна в течение одного момента формирования символа. В последней строке рис. 2.5, *а* указаны 10 сигналов, передаваемых восьмеричной системой модуляции для представления текстового сообщения "THINK".

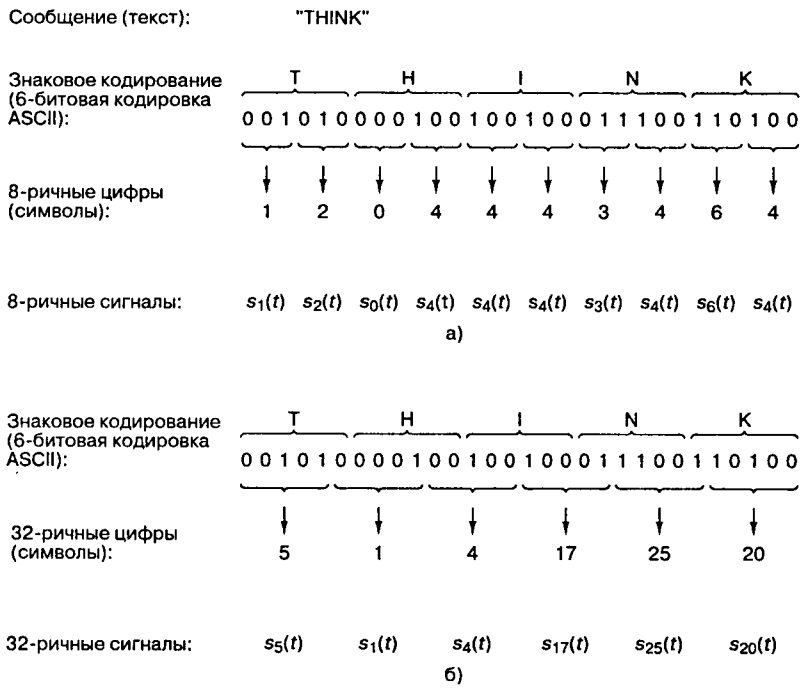


Рис. 2.5. Сообщения, знаки и символы: а) 8-ричный пример; б) 32-ричный пример

На рис. 2.5, *б* размер набора символов, M, был выбран равным 32 (каждый символ представляет 32-ричную цифру). Следовательно, биты берутся по пять, а результирующая группа из шести чисел представляет шесть готовых к передаче 32-ричных символов. Отметим, что границы символов и знаков не обязательно должны совпадать. Первый символ представляет 5/6 первого знака, "T", второй символ — оставшуюся 1/6 знака "T" и 4/6 следующего знака, "H", и т.д. Более эффектное разбиение знаков совсем не обязательно, поскольку система рассматривает знаки как строку символов, которую необходимо передать; только конечный пользователь (или теле-

тайп пользователя) приписывает текстовое значение полученной последовательности битов. В 32-ричном примере передатчик должен содержать набор из 32 сигналов $s_i(t)$, где $i = 1, ..., 32$, сопоставляемых со всеми возможными символами. В последней строке рис. 2.5, *б* указаны шесть сигналов, передаваемых 32-ричной системой модуляции для представления текстового сообщения "THINK".

2.4. Форматирование аналоговой информации

Если информация является аналоговой, ее знаковое кодирование (как в случае текстовой информации) невозможно; вначале информацию следует перевести в цифровой формат. Процесс преобразования аналогового сигнала в форму, совместимую с цифровой системой связи, начинается с дискретизации сигнала; результатом этого процесса является модулированный сигнал, который описывается ниже.

2.4.1. Теорема о дискретном представлении

Аналоговый сигнал и его дискретная версия связаны процессом, который называется *дискретизацией* (sampling process). Этот процесс можно реализовывать по-разному, а наиболее популярной является операция *выборки-хранения* (sample-and-hold). В этом случае коммутирующе-запоминающий механизм (такой, как последовательность транзистора и конденсатора или затвора и диафильма) формирует из поступающего непрерывного сигнала последовательность выборок (sample). Результатом процесса дискретизации является сигнал в *амплитудно-импульсной модуляции* (pulse-amplitude modulation — PAM). Такое название возникло потому, что выходящий сигнал можно описать как последовательность импульсов с амплитудами, определяемыми выборками входящего сигнала. Аналоговый сигнал можно восстановить (с определенной степенью точности) из модулированного сигнала путем прохождения последнего через фильтр нижних частот. Важно знать, насколько точно отфильтрованный модулированный сигнал совпадает с исходным аналоговым сигналом? Ответ на этот вопрос дает *теорема о дискретном представлении* (sampling theorem), которая формулируется следующим образом [1]: сигнал с ограниченной полосой, не имеющий спектральных компонентов с частотами, которые превышают f_m Гц, однозначно определяется значениями, выбранными через равные промежутки времени.

$$T_s \le \frac{1}{2f_m} c \tag{2.1}$$

Это утверждение также известно как *теорема о равномерном дискретном представлении* (uniform sampling theorem). При другой формулировке верхний предел T_s можно выразить через частоту дискретизации (sampling rate), $f_s = 1/T_s$. В этом случае получаем ограничение, именуемое *критерием Найквиста* (Nyquist criterion).

$$f_s \ge 2f_m \tag{2.2}$$

Частота дискретизации $f_s = 2f_m$ также называется *частотой Найквиста* (Nyquist rate). Критерий Найквиста — это теоретическое достаточное условие, которое делает возможным *полное восстановление* аналогового сигнала из последовательности равномерно распределенных дискретных выборок. В следующем разделе демонстрируется справедливость теоремы о дискретном представлении для различных способов взятия выборок.

2.4.1.1. Выборка с использованием единичных импульсов

В данном разделе справедливость теоремы о дискретном представлении демонстрируется с помощью свойства преобразования Фурье, относящегося к свертке в частотной области. Рассмотрим вначале идеальную дискретизацию с помощью последовательности единичных импульсных функций. Предположим, у нас имеется аналоговый сигнал $x(t)$, приведенный на рис. 2.6, *а*, и его Фурье-образ $X(f)$ (рис. 2.6, *б*) равен нулю вне интервала $(-f_m < f < f_m)$. Дискретное представление $x(t)$ можно рассматривать как произведение функции $x(t)$ и последовательности периодических единичных импульсов $x_\delta(t)$, показанной на рис. 2.6, *в* и определяемой следующей формулой.

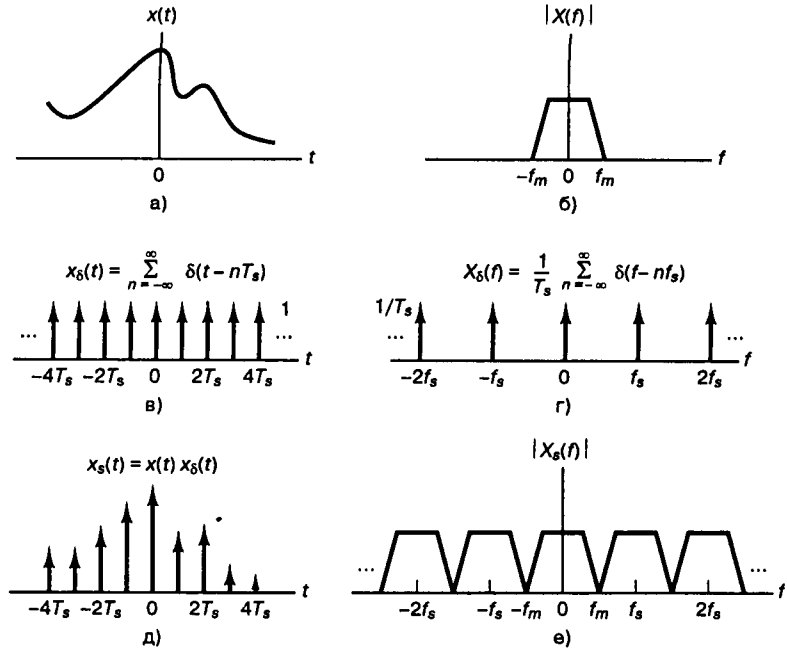


Рис. 2.6. Теорема о дискретном представлении и свертка Фурье-образов

$$x_\delta(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \tag{2.3}$$

Здесь T_s — период дискретизации, а $\delta(t)$ — единичный импульс или дельта-функция Дирака, определенная в разделе 1.2.5. Выберем T_s равным $1/2f_m$, так что будет выполнено минимальное необходимое условие удовлетворения критерия Найквиста.

Выборочное свойство импульсной функции (см. раздел А.4.1) можно описать следующим выражением.

$$x(t)\delta(t - t_0) = x(t_0)\delta(t - t_0) \tag{2.4}$$

Воспользовавшись этим свойством, можно заметить, что $x_s(t)$, дискретный вариант $x(t)$, показанный на рис. 2.6, *д*, описывается следующим выражением.

$$x_s(t) = x(t)x_\delta(t) = \sum_{n=-\infty}^{\infty} x(t)\delta(t - nT_s) = \tag{2.5}$$

$$= \sum_{n=-\infty}^{\infty} x(nT_s)\delta(t - nT_s)$$

Используя свойство преобразования Фурье для свертки в частотной области (см. раздел А.5.3), мы можем преобразовать произведение временных функций $x(t)x_\delta(t)$ в уравнении (2.5) в свертку частотных функций $X(f) * X_\delta(f)$, где

$$X_\delta(f) = \frac{1}{T_s} \sum_{n=-\infty}^{\infty} \delta(f - nf_s) \tag{2.6}$$

является Фурье-образом последовательности импульсов $x_\delta(t)$, а $f_s = 1/T_s$ — частотой дискретизации. Отметим, что Фурье-образ последовательности импульсов — это другая последовательность импульсов; периоды обеих последовательностей обратны друг другу. Последовательность импульсов $x_\delta(t)$ и ее Фурье-образ $X_\delta(f)$ показаны на рис. 2.6, *в*, *г*.

Свертка с импульсной функцией смещает исходную функцию.

$$X(f) * \delta(f - nf_s) = X(f - nf_s) \tag{2.7}$$

Запишем теперь Фурье-образ дискретного сигнала.

$$X_s(f) = X(f) * X_\delta(f) = X(f) * \left[\frac{1}{T_s} \sum_{n=-\infty}^{\infty} \delta(f - nf_s)\right] = \tag{2.8}$$

$$= \frac{1}{T_s} \sum_{n=-\infty}^{\infty} X(f - nf_s)$$

Итак, приходим к заключению, что в пределах исходной полосы спектр $X_s(f)$ дискретного сигнала $x_s(t)$ равен, с точностью до постоянного множителя $(1/T_s)$, спектру исходного сигнала $x(t)$. Кроме того, спектр периодически повторяется по частоте с интервалом f_s Гц. Фильтрующее свойство импульсной функции позволяет легко получить свертку в частотной области последовательности импульсов с другой функцией. Импульсы действуют как стробирующие функции. Значит, свертку можно выполнить графически, накрывая последовательность импульсов $X_\delta(f)$, показанную на рис. 2.6, *г*, образом $|X(f)|$, представленным на рис. 2.6, *б*. Этот процесс повторяет функцию $|X(f)|$ в каждом интервале частот последовательности импульсов, что в конечном итоге дает функцию $|X_s(f)|$, показанную на рис. 2.6, *е*.

После выбора частоты дискретизации (в предыдущем примере $f_s = 2f_m$) каждая спектральная копия отделяется от соседних полосой частот, равной f_s Гц, и аналоговый сигнал полностью восстанавливается из выборок путем фильтрации. В то же время для выполнения этого потребовался бы идеальный фильтр с абсолютно крутыми фронтами. Очевидно, что если $f_s > 2f_m$, копии отдалятся (в частотной области), как показано на рис. 2.7, *а*, и это облегчит операцию фильтрации. На рисунке также показана типичная характеристика фильтра нижних частот, который может использоваться

для выделения спектра немодулированного сигнала. При уменьшении частоты дискретизации до $f_s < 2f_m$ копии начнут перекрываться, как показано на рис. 2.7, *б*, и информация частично будет потеряна. Явление, являющееся результатом недостаточной частоты выборки (выборки, производимой очень редко), называется *наложением* (aliasing). Частота Найквиста $f = 2f_m$ — это предел, ниже которого происходит наложение; чтобы избежать этого нежелательного явления, следует удовлетворять критерий Найквиста $f_s \geq 2f_m$.

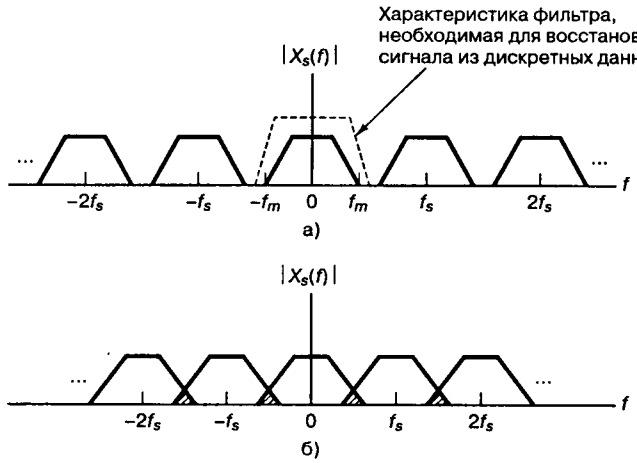


Рис. 2.7. Спектры для различных частот дискретизации: а) дискретный спектр ($f_s > 2f_m$); б) дискретный спектр ($f_s < 2f_m$)

С практической точки зрения ни сигналы, представляющие технический интерес, ни реализуемые узкополосные фильтры не имеют строго ограниченной полосы. Сигналы с идеально ограниченной полосой не существуют в природе (см. раздел 1.7.2); следовательно, реализуемые сигналы, даже если мы можем считать, что они имеют ограниченную полосу, в действительности всегда включают некоторое наложение. Эти сигналы и фильтры могут, впрочем, рассматриваться как ограниченные. Под последним мы подразумеваем, что можно определить полосу, вне которой спектральные компоненты затухают настолько, что ими можно пренебречь.

2.4.1.2. Естественная дискретизация

В данном разделе справедливость теоремы о дискретном представлении демонстрируется с помощью свойства преобразования Фурье, заключающегося в сдвиге частоты. Хотя мгновенная выборка и является удобной моделью, все же более практичный способ дискретизации аналогового сигнала $x(t)$ с ограниченной полосой частот (рис. 2.8, *а,б*) состоит в его умножении на серию импульсов или коммутирующий сигнал $x_p(t)$ (рис. 2.8, *в*). Каждый импульс серии $x_p(t)$ имеет ширину T и амплитуду $1/T$. Умножение на $x_p(t)$ можно рассматривать как включение и выключение коммутатора. Как и ранее, частота дискретизации обозначается через f_s, а величина, обратная к ней (время между выборками), — через T_s. Получаемая последовательность дискретных данных, $x_s(t)$, показана на рис. 2.8, *д*; она выражается следующей формулой.

$$x_s(t) = x(t)x_p(t) \tag{2.9}$$

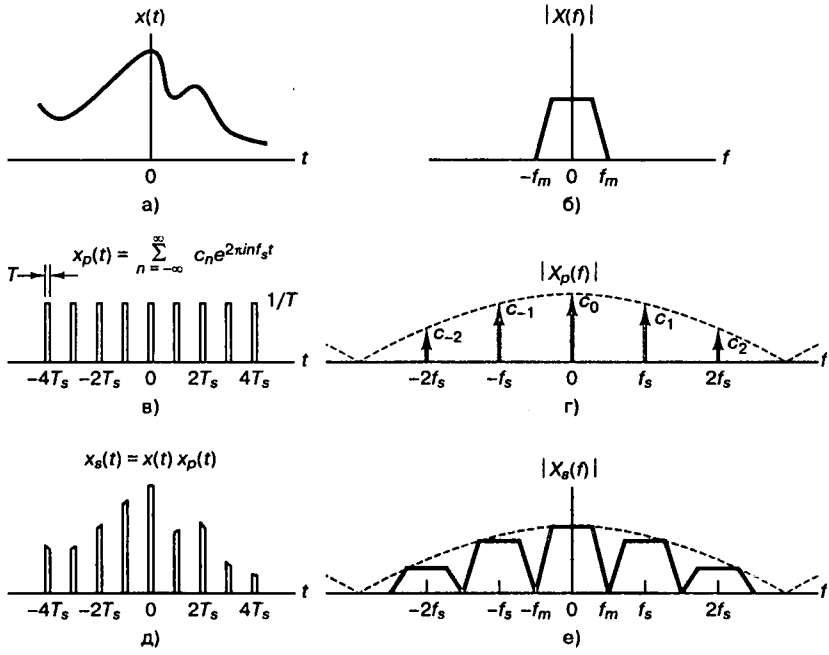


Рис. 2.8. Теорема о дискретном представлении и сдвиг частоты Фурье-образа

В данном случае мы имеем дело с так называемой *естественной дискретизацией* (natural sampling), поскольку вершина каждого импульса $x_s(t)$ в течение интервала его передачи имеет форму соответствующего аналогового сегмента. С помощью уравнения (A.13) периодическую серию импульсов можно выразить как ряд Фурье:

$$x_p(t) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n f_s t} \tag{2.10}$$

где частота дискретизации, $f_s = 1/T_s$, выбрана равной $2f_m$, так что выполнено минимальное необходимое условие критерия Найквиста. Из уравнения (A.24) $c_n = (1/T_s)\,\mathrm{sinc}\,(nT/T_s)$, где T — ширина импульса, $1/T$ — его амплитуда, а

$$\mathrm{sinc}\ y = \frac{\sin \pi y}{\pi y}.$$

Огибающая спектра амплитуд серии импульсов, показанная на рис. 2.8, *г* пунктиром, имеет вид функции sinc. Объединяя выражения (2.9) и (2.10), получаем следующее.

$$x_s(t) = x(t) \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n f_s t} \tag{2.11}$$

Образ $X_s(f)$ дискретного сигнала находится следующим образом.

$$X_s(f) = \mathfrak{F}\left\{ x(t) \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n f_s t} \right\} \tag{2.12}$$

Для линейных систем операции суммирования и преобразования Фурье можно менять местами. Следовательно, можно записать следующее.

$$X_s(f) = \sum_{n=-\infty}^{\infty} \mathfrak{F}\left\{x(t)c_n e^{2\pi i n f_s t}\right\} \tag{2.13}$$

Используя свойство *трансляции частоты* преобразования Фурье (см. раздел А.3.2), получаем следующее выражение для $X_s(f)$.

$$X_s(f) = \sum_{n=-\infty}^{\infty} c_n X(f - nf_s) \tag{2.14}$$

Подобно дискретизации с использованием единичных импульсов формула (2.14) и рис. 2.8, *е* показывают, что $X_s(f)$ — это копия $X(f)$, периодически повторяющаяся по частоте с интервалом f_s Гц. Впрочем, при естественной дискретизации видим, что $X_s(f)$ взвешена на коэффициенты ряда Фурье серии импульсов, тогда как при дискретизации единичными импульсами имеем импульсы постоянной формы. Отметим, что *в пределе*, при стремящейся к нулю ширине импульса T, c_n стремится к $1/T_s$ для всех n (см. пример ниже) и уравнение (2.14) переходит в уравнение (2.8).

Пример 2.1. Сравнение дискретизации единичными импульсами и естественной дискретизации

Рассмотрим данный сигнал $x(t)$ и его Фурье-образ $X(f)$. Пусть $X_{s1}(f)$ — спектр сигнала $x_{s1}(t)$, являющегося результатом дискретизации $x(t)$ с помощью серии единичных импульсов $x_\delta(t)$, а $X_{s2}(f)$ — спектр сигнала $x_{s2}(t)$, являющегося результатом дискретизации $x(t)$ с помощью серии импульсов $x_p(t)$, имеющих ширину T, амплитуду $1/T$ и период T_s. Покажите, что в пределе $T \to 0$ $X_{s1}(f) = X_{s2}(f)$.

Решение

Из уравнения (2.8)

$$X_{s1}(f) = \frac{1}{T_s} \sum_{n=-\infty}^{\infty} X(f - nf_s)$$

и из уравнения (2.14)

$$X_{s2}(f) = \sum_{n=-\infty}^{\infty} c_n X(f - nf_s)$$

При $T \to 0$ амплитуда импульса стремится к бесконечности (площадь импульса постоянна) и $x_p(t) \to x_\delta(t)$. С помощью уравнения (А.14) коэффициенты c_n можно записать как следующий предел.

$$c_n = \lim_{T \to 0} \frac{1}{T_s} \int_{-T_s/2}^{T_s/2} x_p(t) e^{-2\pi i n f_s t} dt =$$

$$= \frac{1}{T_s} \int_{-T_s/2}^{T_s/2} x_\delta(t) e^{-2\pi i n f_s t} dt$$

Следовательно, в пределах интегрирования (от $-T_s/2$ до $T_s/2$) единственный ненулевой вклад в интеграл дает значение $x_\delta(t) = \delta(t)$; в данном случае можно записать следующее.

$$c_n = \frac{1}{T_s}\int_{-T_s/2}^{T_s/2}\delta(t)e^{-2\pi inf_s t}dt = \frac{1}{T}$$

Получаем, что в пределе для всех n $X_{s1}(f) = X_{s2}(f)$.

2.4.1.3. Метод "выборка-хранение"

Простейшим, а поэтому и наиболее популярным методом дискретизации является *выборка-хранение*. Описать этот метод можно с помощью свертки серии дискретных импульсов, $[x(t)x_\delta(t)]$, показанной на рис. 2.6, *д*, с прямоугольным импульсом $p(t)$, имеющим единичную амплитуду и ширину T_s. Эта свертка дает дискретную последовательность импульсов с плоским верхом.

$$\begin{aligned} x_s(t) &= p(t) * [x(t)x_\delta(t)] = \\ &= p(t) * \left[x(t) * \sum_{n=-\infty}^{\infty}\delta(t - nT_s) \right] \end{aligned} \tag{2.15}$$

Фурье-образ, $X_s(f)$, временной свертки в уравнении (2.15) равен произведению в частотной области Фурье-образа $P(f)$ прямоугольного импульса и периодического спектра импульсно-дискретных данных, показанного на рис. 2.6, *е*.

$$\begin{aligned} X_s(f) &= P(f)\Im\left\{ x(t)\sum_{n=-\infty}^{\infty}\delta(t - nT_s) \right\} = \\ &= P(f)\left\{ X(f) * \left[\frac{1}{T_s}\sum_{n=-\infty}^{\infty}\delta(f - nf_s) \right] \right\} = \\ &= P(f)\frac{1}{T_s}\sum_{n=-\infty}^{\infty}X(f - nf_s) \end{aligned} \tag{2.16}$$

Здесь $P(f)$ имеет вид $T_s \operatorname{sinc} fT_s$. Результатом умножения является спектр, подобный спектру примера естественной дискретизации (рис. 2.8, *е*). Наиболее явный результат операции хранения — значительное затухание высокочастотных спектральных копий (сравните рис. 2.8, *е* и 2.6, *е*), что весьма желательно. Как правило, для завершения процесса фильтрации требуется дополнительная аналоговая фильтрация, позволяющая подавить остаточные спектральные компоненты, кратные частоте дискретизации. Вторичным результатом операции хранения является неоднородное усиление (или подавление) спектра нужной полосы частот за счет функции $P(f)$ (см. формулу 2.16). После фильтрации это подавление можно компенсировать путем применения функции, обратной к $P(f)$.

2.4.2. Наложение

На рис. 2.9 представлено увеличенное изображение рис. 2.7, *б*, на котором дана положительная половина спектра немодулированного сигнала и одна копия сиг-

нала. Этот рисунок иллюстрирует наложение в частотной области. Перекрывающаяся область, показанная на рис. 2.9, *б*, содержит ту часть спектра, которая налагается вследствие *недостаточной частоты выборки*. Накладывающиеся спектральные компоненты представляют собой неоднозначную информацию, находящуюся в полосе частот $(f_s - f_m, f_m)$. Из рис. 2.10 видно, что повышение частоты дискретизации f'_s позволяет устранить наложение путем разделения спектральных копий; результирующий спектр, показанный на рис. 2.10, *б*, соответствует случаю, приведенному на рис. 2.7, *а*. На рис. 2.11 и 2.12 продемонстрированы два способа борьбы с наложением, в которых используются *фильтры защиты от наложения спектров* (antialiasing filter). На рис. 2.11 аналоговый сигнал *предварительно фильтруется*, так что новая максимальная частота f'_m уменьшается до $f_s/2$ или даже сильнее. Таким образом, поскольку $f_s > 2f'_m$, на рис. 2.11, *б* уже отсутствуют перекрывающиеся компоненты. Такой метод устранения наложения до дискретизации очень хорошо себя зарекомендовал в области проектирования цифровых систем. При хорошо известной структуре сигнала наложение может устраняться и после дискретизации, для чего дискретные данные пропускаются через фильтр нижних частот [2]. На рис. 2.12, *а,б* накладывающиеся компоненты удаляются *после* дискретизации; частота среза фильтра f''_m удаляет перекрывающиеся компоненты; частота f''_m должна быть меньше $(f_s - f_m)$. Отметим, что методы фильтрации, применяемые для удаления части спектра, в которой присутствует наложение, на рис. 2.11 и 2.12 *приведут к потере* некоторой информации. По этой причине частота дискретизации, ширина полосы среза и тип фильтра, выбираемые для конкретного сигнала, не являются независимыми параметрами.

Реализуемые фильтры требуют ненулевой ширины полосы для перехода между полосой пропускания и областью затухания. Эта область называется *полосой перехода*. Для минимизации частоты дискретизации системы желательно было бы, чтобы *фильтры защиты от наложения спектров* имели узкую полосу перехода. В то же время при сужении полосы перехода резко возрастает сложность фильтров и их стоимость, так что необходимо принять компромиссное решение относительно цены более узкой полосы перехода и цены высокой частоты дискретизации. Во многих системах оптимальной шириной полосы перехода является 10–20% от ширины полосы сигнала. Рассчитав частоту дискретизации Найквиста для 20%-ной ширины перехода фильтра защиты от наложения спектров, получим *инженерную версию* критерия Найквиста.

$$f_s \geq 2{,}2f_m \tag{2.17}$$

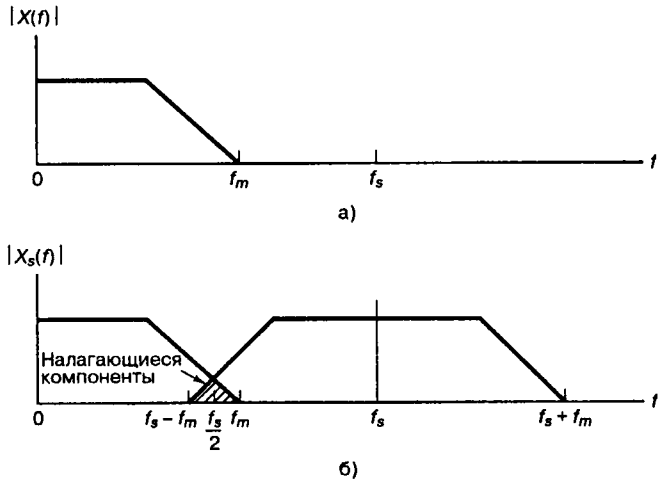


Рис. 2.9. Наложение в частотной области: а) спектр непрерывного сигнала; б) спектр дискретного сигнала

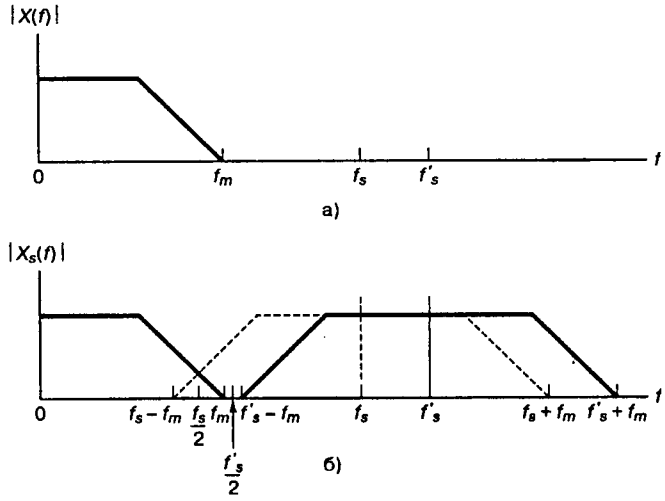


Рис. 2.10. Большая частота дискретизации позволяет избежать наложения: а) спектр непрерывного сигнала; б) спектр дискретного сигнала

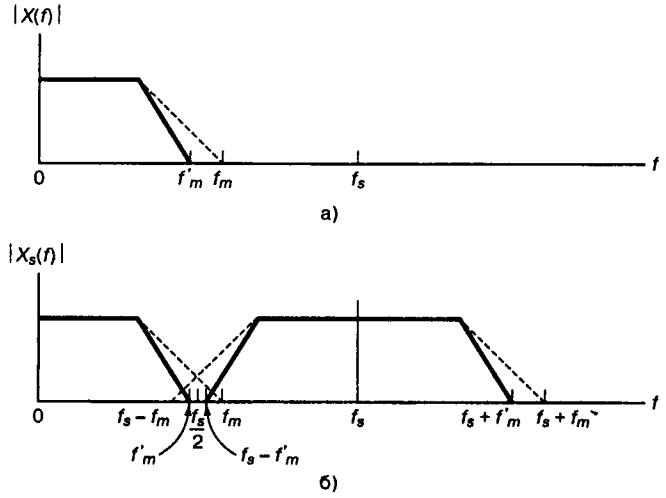


Рис. 2.11. Фильтры с более острым отсеканием позволяют устранить наложение: а) спектр непрерывного сигнала; б) спектр дискретного сигнала

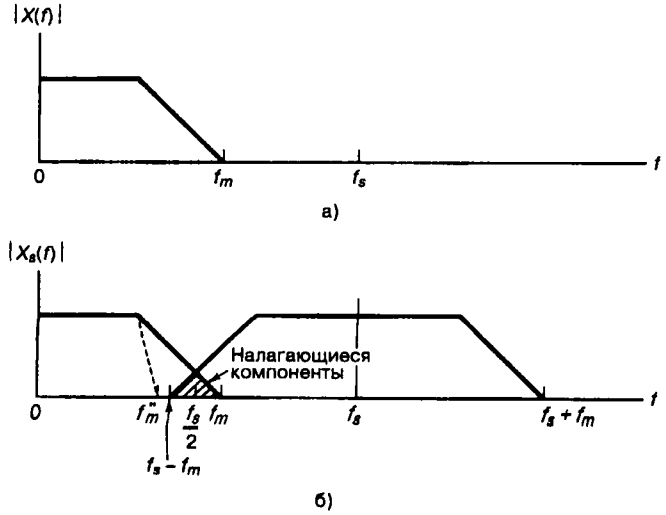


Рис. 2.12. Фильтрация после дискретизации устраняет часть спектра, в которой имеется наложение: а) спектр непрерывного сигнала; б) спектр дискретного сигнала

На рис. 2.13 показано, как выглядит наложение во временной области. Точками показаны выборки сигнала (сплошная синусоида). Отметим, что вследствие недостаточной частоты выборки через точки выборки можно проложить другую синусоиду (пунктир).

Пример 2.2. Частота дискретизации для музыкальной системы высокого качества

Требуется с высоким качеством оцифровать музыкальный источник с шириной полосы 20 кГц. Для этого нужно определить частоту дискретизации. Используя инженерную версию критерия Найквиста, формулу (2.17), получаем, что частота дискретизации должна превышать 44,0 тысячи

выборок в секунду. Для сравнения, стандартная частота дискретизации для аудиопроигрывателя компакт-дисков составляет 44,1 тысячи выборок в секунду, а стандартная частота дискретизации аудиодисков студийного качества равна 48,0 тысяч выборок в секунду.

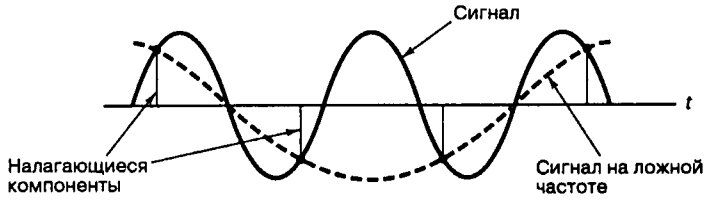


Рис. 2.13. Наложенные частоты, возникшие вследствие дискретизации с частотой, меньшей частоты Найквиста

2.4.3. Зачем нужна выборка с запасом

Выборка с запасом (oversampling) — это наиболее экономичное решение задачи преобразования аналогового сигнала в цифровой или цифрового в аналоговый. Это объясняется тем, что обработка сигнала выполняется на высокопроизводительном аналоговом оборудовании, что обычно дороже использования для этой же задачи цифрового оборудования обработки сигналов. Рассмотрим преобразование аналоговых сигналов в цифровые. Если это выполняется без выборки с запасом, то процесс дискретизации описывается тремя простыми этапами.

Выборка без запаса

1. Сигнал пропускается через высокопроизводительный аналоговый фильтр нижних частот для ограничения его полосы.
2. Отфильтрованный сигнал дискретизируется с частотой Найквиста с целью создания сигнала с (приблизительно) ограниченной полосой. Как описывалось в разделе 1.7.2, сигнал со строго ограниченной полосой относится к разряду нереализуемых.
3. Выборки квантуются устройством преобразования аналоговых сигналов в цифровые, отображающим выборки, которые могут принимать значения из непрерывного диапазона, в конечный набор дискретных уровней.

Если же выборку производить с запасом, то процесс будет состоять из пяти этапов.

Выборка с запасом

1. Сигнал пропускается через менее производительный (более дешевый) аналоговый фильтр нижних частот (предварительная фильтрация) для ограничения его полосы.
2. Предварительно отфильтрованный сигнал выбирается с частотой в несколько раз выше частоты Найквиста для создания сигнала с ограниченной полосой.
3. Выборки преобразовываются преобразователем аналоговых сигналов в цифровые, отображающим выборки, которые могут принимать значения из непрерывного диапазона, в конечный набор дискретных уровней.
4. Цифровые выборки обрабатываются высокопроизводительным цифровым фильтром для сужения полосы цифровых выборок.
5. Частота дискретизации на выходе цифрового фильтра уменьшается пропорционально сужению полосы, полученному при использовании этого цифрового фильтра.

Преимущества выборки с запасом подробно рассматриваются в двух следующих разделах.

2.4.3.1. Аналоговая фильтрация, дискретизация и преобразование аналоговых сигналов в цифровые

Полоса пропускания аналогового фильтра, ограничивающая ширину полосы входящего сигнала, равна ширине полосы сигнала плюс область спада (stop band). Наличие области перехода приводит к увеличению ширины полосы сигнала на выходе на некоторую величину f_t. Частоту Найквиста f_s для отфильтрованного выхода, обычно равную $2f_m$ (удвоенной максимальной частоте дискретного сигнала), теперь необходимо увеличить до $2f_m + f_t$. Ширина полосы спада фильтра является служебными издержками процесса дискретизации. Этот дополнительный спектральный интервал не представляет полосы полезного сигнала, а нужен для защиты полосы сигнала путем резервирования спектральной области для накладывающегося спектра, возникающего в процессе дискретизации. Наложение возникает вследствие того, что реальный сигнал не может быть строго ограниченным. Типичные полосы спада дают 10–20%-ное увеличение частоты дискретизации по сравнению с частотой, определяемой критерием Найквиста. Примером таких служебных издержек может служить цифровая аудиосистема проигрывания компакт-дисков, где двусторонняя полоса равна 40 кГц, а частота дискретизации — 44,1 кГц, или система проигрывания цифровых аудиокассет (digital audio type — DAT), в которой ширина двусторонней полосы также равна 40 кГц, а частота дискретизации — 48,0 кГц.

Естественным желанием является использование для создания аналоговых фильтров с узкой полосой перехода и максимально низкой из возможных частот дискретизации. В то же время аналоговые фильтры имеют две нежелательные особенности. Во-первых, они могут приводить к искажению (нелинейное изменение фазы с частотой), вызванному малыми областями перехода. Во-вторых, цена системы может оказаться высокой, поскольку узкие области перехода подразумевают применение фильтров высоких порядков (см. раздел 1.6.3.2), требующих большого числа высококачественных составляющих. Проблема состоит в том, что для уменьшения стоимости хранения данных хотелось бы работать с устройством дискретизации с максимально низкой частотой. Для достижения этой цели можно создать изощренный аналоговый фильтр с узкой областью перехода. Однако такой фильтр не только дорог, но и искажает сам сигнал, хотя задачей фильтра как раз является защита сигнала (от нежелаемого наложения).

В данном случае выборка с запасом наиболее приемлема — при наличии проблемы, решить которую мы не можем, превращаем ее в проблему, поддающуюся решению. Мы используем дешевый, менее сложный предварительный аналоговый фильтр для ограничения полосы входящего сигнала. Этот аналоговый фильтр можно упростить за счет выбора более широкой переходной области. При этом увеличивается ширина спектра, из-за чего нам нужно увеличить требуемую частоту дискретизации. Обычно начинают с выбора частоты дискретизации, в 4 раза превышающей исходную, после чего разрабатывают аналоговый фильтр, ширина полосы которого соответствует этой увеличенной частоте дискретизации. Например, вместо дискретизации сигнала компакт-диска на частоте 44,1 кГц при ширине области перехода 4,1 кГц, реализованной с использованием сложнейшего эллиптического фильтра 10-го порядка (подразумевается, что фильтр включает 10 избирательных элементов, таких как конденсаторы и индуктивности), мы выбираем выборку с запасом. В этом случае устройство дискретизации может работать на частоте 176,4 кГц с областью перехода 136,4 кГц, реализованное простым эллиптическим фильтром 4-го порядка (имеющим всего 4 избирательных элемента).

2.4.3.2. Цифровая фильтрация и повторная выборка

Итак, у нас есть дискретные данные с большей, чем требуется, частотой дискретизации, и эти данные пропускаются через недорогой высокопроизводительный цифровой фильтр для выполнения фильтрации, необходимой для предотвращения наложения. Цифровой фильтр может реализовать узкую область перехода без искажения, свойственного аналоговым фильтрам, а его эксплуатация недорогая. После того как цифровая фильтрация уменьшила ширину полосу перехода, мы снижаем частоту дискретизации сигнала (повторная выборка). В результате в единую структуру объединяются качественные методы цифровой обработки, фильтрация и повторная выборка.

Рассмотрим теперь вопрос дальнейшего улучшения качества процесса сбора данных. Предварительный аналоговый фильтр приводит к некоторому искажению амплитуды и фазы. Поскольку заранее известно, каково это искажение, цифровой фильтр проектируется не только для защиты (совместно с аналоговым фильтром) от наложения, но и для компенсации усиления и искажения фазы, вносимых аналоговым фильтром. Суммарный результат может, по желанию, улучшаться до любого предела. Таким образом, получаем сигнал более высокого качества (менее искаженный) по более низкой цене. Аппаратура цифровой обработки сигналов, представляющая собой развитие компьютерной индустрии, характеризуется значительным ежегодным снижением цен, чего нельзя сказать об аналоговой аппаратуре.

Подобным образом выборка с запасом используется в процессе преобразования цифрового сигнала в аналоговый (digital-to-analog conversion — DAC). Аналоговый фильтр, через который пропускается преобразованный сигнал, будет искажать сигнал, если последний будет иметь узкую полосу перехода. Но полоса перехода уже не будет узкой, если данные, полученные после преобразования DAC, были оцифрованы с помощью выборки с запасом.

2.4.4. Сопряжение сигнала с цифровой системой

Рассмотрим четыре способа описания аналоговой исходной информации. Возможные варианты показаны на рис. 2.14. Сигнал, изображенный на рис. 2.14, *а*, будем называть *исходным аналоговым*. На рис. 2.14, *б* представлена дискретная версия исходного сигнала, обычно именуемая *данными, оцифрованными естественным способом*, или *данными с амплитудно-импульсной модуляцией* (pulse amplitude modulation — PAM). Думаете, дискретные данные на рис. 2.14, *б* совместимы с цифровой системой? Нет, поскольку амплитуда каждой естественной выборки все еще может принимать бесконечное множество возможных значений, а цифровая система работает с конечным набором значений. Даже если дискретные сигналы имеют плоские вершины, возможные значения составляют бесконечное множество, поскольку они отражают все возможные значения непрерывного аналогового сигнала. На рис. 2.14, *в* показано представление исходного сигнала дискретными импульсами. Здесь импульсы имеют плоскую вершину, *и* возможные значения амплитуд импульсов ограничены конечным множеством. Каждый импульс характеризуется уровнем, причем все уровни предопределены и составляют конечное множество; каждый уровень может представляться символом конечного алфавита. Импульсы на рис. 2.14, *в* называются *квантованными выборками*; такой формат является естественным выбором для сопряжения с цифровой системой. Формат, показанный на рис. 2.14, *г*, может быть получен на выходе схемы выборки-хранения. При квантовании дискретных значений в конечное множество, данные в таком формате совместимы с цифровой системой. После квантования аналоговый сигнал по-прежнему может восста-

навливаться, но уже не абсолютно точно; повысить точность восстановления аналогового сигнала можно за счет увеличения числа уровней квантования (требуется увеличение ширины полосы системы). Искажение сигнала вследствие квантования будет рассмотрено далее в этой главе (и в главе 13).

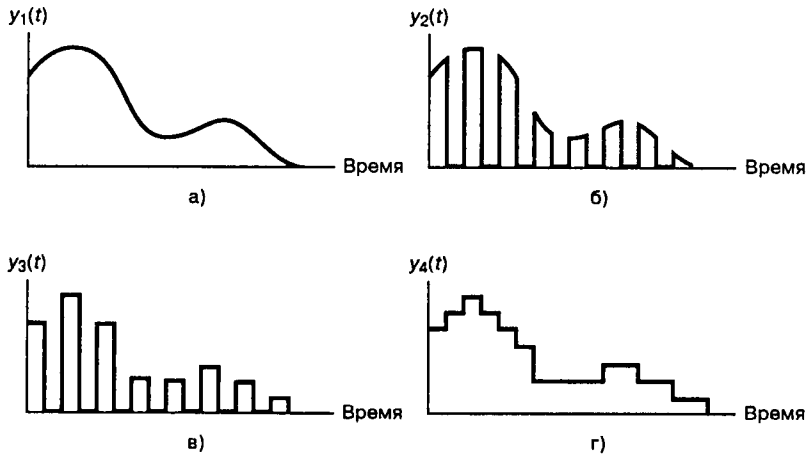


Рис. 2.14. Исходные данные в системе координат "время-амплитуда": а) исходный аналоговый сигнал; б) данные в естественной дискретизации; в) квантованные выборки; г) выборка-хранение

2.5. Источники искажения

Аналоговый сигнал, восстановленный из дискретных, квантованных и переданных импульсов, будет искажен. Основные источники искажения связаны с (1) влиянием дискретизации и квантования и (2) воздействием канала. Ниже эти вопросы рассматриваются подробно.

2.5.1. Влияние дискретизации и квантования

2.5.1.1. Шум квантования

Искажение, присущее квантованию, — это ошибка округления или усечения. Процесс кодирования сигнала PAM в квантованный сигнал PAM включает отбрасывание некоторой исходной аналоговой информации. Это искажение, вызванное необходимостью аппроксимации аналогового сигнала квантованными выборками, называется *шумом квантования*; величина этого шума обратно пропорциональна числу уровней, задействованных в процессе квантования. (Отношение сигнал/шум для квантованных импульсов рассматривается в разделах 2.5.3 и 13.2.)

2.5.1.2. Насыщение устройства квантования

Устройство квантования (преобразования аналоговых сигналов в цифровые) для аппроксимации значений из непрерывного диапазона на входе значениями из конечного множества на выходе выделяет L уровней. Диапазон входных значений, для которых разница между входом и выходом незначительна, называется *рабочим диапазоном* преобразователя. Если входящее значение не принадлежит этому диапазону, значения на вхо-

де и выходе отличаются сильнее, и мы говорим, что преобразователь работает *в режиме насыщения*. Ошибки насыщения значительнее и менее желательны, чем шум квантования. В общем случае насыщение устраняется путем автоматической регулировки усиления (automatic gain control — AGC), которая эффективно расширяет рабочий диапазон преобразователя. (Подробнее о насыщении устройства квантования в главе 13.)

2.5.1.3. Синхронизация случайного смещения

Наш анализ теоремы о дискретном представлении предсказывал точное восстановление сигнала на основе равномерно размещенных выборок. При наличии случайного смещения положения выборки, дискретизация уже не является равномерной. Если местоположения выборок точно известны, точное восстановление все еще возможно, но смещение — это обычно случайный процесс, так что заранее предсказать положения выборок нельзя. Воздействие смещения равносильно частотной модуляции узкополосного сигнала. Если смещение является случайным, вносится низкоуровневый широкополосный спектральный вклад, характеристики которого весьма подобны свойствам шума квантования. Если смещение является периодическим, как, например, при считывании данных с магнитофона, то в данных появятся низкоуровневые спектральные линии. Управлять синхронизацией случайного смещения можно посредством развязки по питанию и использования кварцевых генераторов.

2.5.2. Воздействие канала

2.5.2.1. Шум канала

Тепловой шум, а также помехи со стороны других пользователей и коммутационного оборудования канала могут приводить к ошибкам в обнаружении импульсов, представляющих оцифрованные выборки. Ошибки, индуцируемые каналом, могут достаточно быстро ухудшить качество восстанавливаемого сигнала. Быстрое ухудшение качества выходного сигнала за счет ошибок, индуцированных каналом, называется *пороговым эффектом* (threshold effect). Если шум канала мал, то проблем с обнаружением сигнала не возникнет. Следовательно, небольшой шум не разрушает восстанавливаемые сигналы. В этом случае при восстановлении единственным шумом является шум квантования. С другой стороны, если шум канала достаточно велик, чтобы повлиять на нашу способность к обнаружению сигналов, в результате полученная ошибка обнаружения приводит к ошибкам восстановления. *Пороговым* данный эффект называется потому, что при небольших изменениях уровня шума канала поведение сигнала может измениться довольно сильно.

2.5.2.2. Межсимвольная интерференция

Канал всегда имеет ограниченную полосу пропускания. Канал с ограниченной полосой всегда искажает или расширяет импульсный сигнал, проходящий через него (см. раздел 1.6.4). Если ширина полосы канала значительно больше ширины полосы импульса, импульс искажается незначительно. Если же ширина полосы канала приблизительно равна ширине полосы сигнала, то искажение будет превышать интервал передачи символа и приведет к наложению импульсов сигнала. Этот эффект называется *межсимвольной интерференцией* (intersymbol interference — ISI). Как и любой другой источник интерференции, ISI приводит к ухудшению качества передачи (повышению уровня ошибок); к тому же эта форма интерференции особенно болезненна, поскольку повышение мощности сигнала для преодоления интерференции не всегда улучшает достоверность передачи. (Подробнее о методах борьбы с межсимвольной интерференцией см. в разделах 3.3 и 3.4.)

2.5.3. Отношение сигнал/шум для квантованных импульсов

Рассмотрим рис. 2.15, на котором изображено L-уровневое устройство квантования аналогового сигнала с полным диапазоном напряжений, равным $V_{pp} = V_p - (-V_p) = 2V_p$ В. Как показано на рисунке, квантованные импульсы могут иметь положительные и отрицательные значения. Шаг между уровнями квантования, называемый *интервалом квантования*, составляет q вольт. Если уровни квантования равномерно распределены по всему диапазону, устройство квантования именуется *равномерным*, или *линейным*. Каждое дискретное значение аналогового сигнала аппроксимируется квантованным импульсом: аппроксимация дает ошибку, не превышающую $q/2$ в положительном направлении или $-q/2$ в отрицательном. Таким образом, ухудшение сигнала вследствие квантования ограничено половиной квантового интервала, $\pm q/2$ вольт.

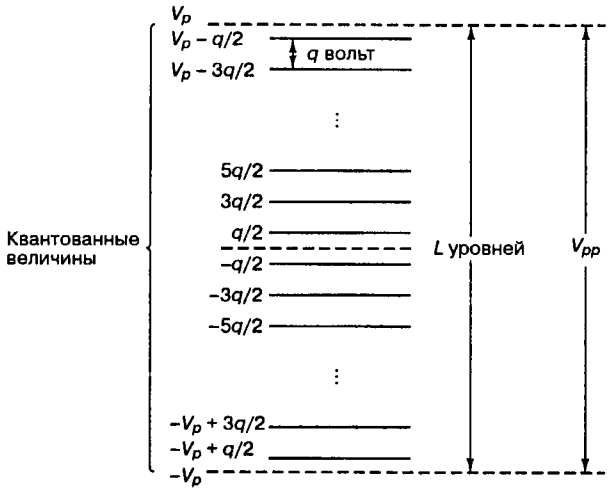


Рис. 2.15. Уровни квантования

Хорошим критерием качества равномерного устройства квантования является его дисперсия (среднеквадратическая ошибка при подразумеваемом нулевом среднем). Если считать, что ошибка квантования, e, равномерно распределена в пределах интервала квантования шириной q (т. е. аналоговый входящий сигнал принимает все возможные значения с равной вероятностью), то дисперсия ошибок для устройства квантования составляет

$$\sigma^2 = \int_{-q/2}^{+q/2} e^2 p(e)\, de = \tag{2.18,а}$$

$$= \int_{-q/2}^{+q/2} e^2 \frac{1}{q}\, de = \frac{q^2}{12}, \tag{2.18,б}$$

где $p(e) = 1/q$ — (равномерно распределенная) плотность вероятности возникновения ошибки квантования. Дисперсия, σ^2, соответствует *средней мощности шума квантова-*

ния. Пиковую мощность аналогового сигнала (нормированную на 1 Ом) можно выразить как

$$V_p^2 = \left(\frac{V_{pp}}{2}\right)^2 = \left(\frac{Lq}{2}\right)^2 = \frac{L^2q^2}{4}, \tag{2.19}$$

где L — число уровней квантования. Объединение выражений (2.18) и (2.19) дает отношение *пиковой* мощности сигнала к *средней* мощности квантового шума $(S/N)_q$ (предполагается отсутствие ошибок, вызванных межсимвольной интерференцией или шумом канала).

$$\left(\frac{S}{N}\right)_q = \frac{L^2q^2/4}{q^2/12} = 3L^2 \tag{2.20}$$

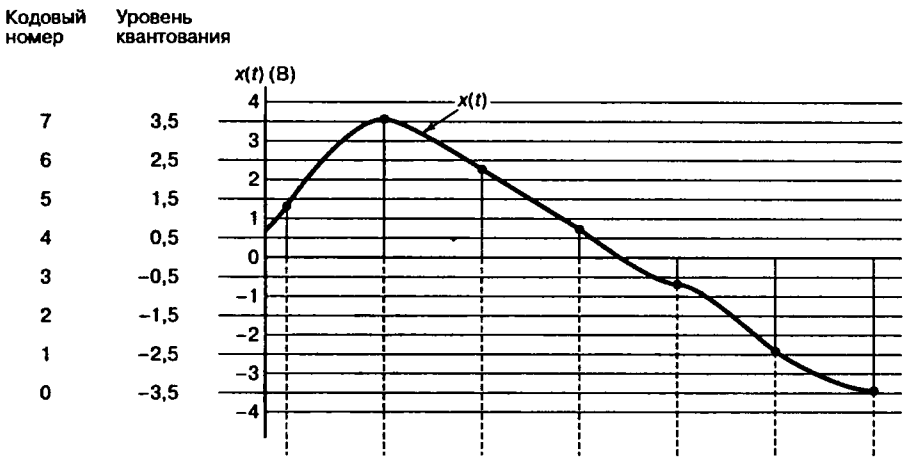
Очевидно, что отношение $(S/N)_q$ квадратично растет с числом уровней квантования. В пределе ($L \to \infty$) сигнал переходит в формат PAM (без квантования) и отношение сигнал/шум становится бесконечным; другими словами, при бесконечном числе уровней квантования имеем нулевой шум квантования.

2.6. Импульсно-кодовая модуляция

Импульсно-кодовая модуляция (pulse-code modulation — PCM) — это название, данное классу узкополосных сигналов, полученных из сигналов PAM путем кодирования каждой квантованной выборки *цифровым словом* [3]. Исходная информация дискретизируется и квантуется в один из L уровней; после этого каждая квантованная выборка проходит цифровое кодирование для превращения в l-битовое ($l = \log_2 L$) кодовое слово. Для узкополосной передачи биты кодового слова преобразовываются в импульсные сигналы. Рассмотрим рис. 2.16, на котором представлена бинарная импульсно-кодовая модуляция. Предположим, что амплитуды аналогового сигнала $x(t)$ ограничены диапазоном от –4 до +4 В. Шаг между уровнями квантования составляет 1 В. Следовательно, используется 8 квантовых уровней; они расположены на –3,5, –2,5, ..., +3,5 В. Уровню –3,5 В присвоим кодовый номер 0, уровню –2,5 — 1 и так до уровня 3,5 В, которому присвоим кодовый номер 7. Каждый кодовый номер имеет представление в двоичной арифметике — от 000 для кодового номера 0 до 111 для кодового номера 7. Почему уровни напряжения выбраны именно так, а не с использованием набора последовательных чисел 1, 2, 3, ...? На выбор уровней напряжения влияют два ограничения. Во-первых, интервалы квантования между уровнями должны быть одинаковыми; и, во-вторых, удобно, чтобы уровни были симметричны относительно нуля.

На оси ординат (рис. 2.16) отложены уровни квантования и их кодовые номера. Каждая выборка аналогового сигнала аппроксимируется ближайшим уровнем квантования. Под аналоговым сигналом $x(t)$ изображены четыре его представления: значения выборок в естественной дискретизации, значения квантованных выборок, кодовые номера и последовательность PCM.

Отметим, что в примере на рис. 2.16 каждая выборка соотнесена с одним из восьми уровней или трехбитовой последовательностью PCM.



Значения, полученные при естественной дискретизации	1,3	3,6	2,3	0,7	−0,7	−2,4	−3,4
Значения, полученные при квантовании	1,5	3,5	2,5	0,5	−0,5	−2,5	−3,5
Кодовый номер	5	7	6	4	3	1	0
Последовательность РСМ	101	111	110	100	011	001	000

Рис. 2.16. Естественные выборки, квантованные выборки и импульсно-кодовая модуляция. (Перепечатано с разрешения авторов из книги Taub and Schilling. Principles of Communications Systems. McGraw-Hill Book Company, New York, 1971, Fig. 6.5-1, p. 205.)

Предположим, что аналоговый сигнал представляет собой музыкальный фрагмент, который выбирается с частотой Найквиста. Допустим также, что при прослушивании музыки в цифровой форме качество звучания ужасное. Что нужно делать для улучшения точности воспроизведения? Напомним, что процесс квантования замещает реальный сигнал его аппроксимацией (т.е. вводит шум квантования). Следовательно, увеличение числа уровней приведет к уменьшению шума квантования. Какими будут последствия, если удвоить число уровней (теперь их будет 16)? В этом случае каждая аналоговая выборка будет представлена четырехбитовой последовательностью PCM. Будет ли это чего-либо стоить? В системе связи реального времени сообщения должны доставляться без задержки. Следовательно, время передачи должно быть одинаковым для всех выборок, вне зависимости от того, сколько битов представляет выборку. Значит, если на выборку приходится больше битов, то они должны перемещаться быстрее; другими словами, они должны заменяться "более узкими" битами. Это приводит к повышению скорости передачи данных, и мы платим увеличением полосы передачи. Сказанное объясняет, как можно получить более точное воспроизведение за счет более широкой полосы передачи. В то же время следует помнить о существовании областей связи, в которых задержка допустима. Рассмотрим, например, передачу планетарных изображений с космического аппарата. Проект "Galileo", начатый в 1989 году, как раз выполнял такую миссию; задача состояла в фотографировании и передаче изображений Юпитера. Аппарат "Galileo" прибыл к своему месту назначения (к Юпитеру) в 1995 году. Путешествие заняло несколько лет; следовательно, любая задержка сигнала на несколько минут (часов или даже дней), естественно, не будет представлять проблемы. В таких случаях за большее число уровней квантования и большую точность воспроизведения не обязательно платить шириной полосы; можно обойтись временным запаздыванием.

На рис. 2.1 термин "PCM" встречается в двух местах. Во-первых, в блоке форматирования, поскольку преобразование аналоговых сигналов в цифровые включает дискретизацию, квантование и, в конечном итоге, посредством преобразования квантованных сигналов PAM в сигналы PCM дает двоичные цифры. Здесь цифры PCM — это просто двоичные числа. Во-вторых, этот термин встречается на рис. 2.1 в разделе "Узкополосная передача сигналов". Здесь перечислены различные сигналы PCM (коды канала), которые могут использоваться для переноса цифр PCM. Отметим, таким образом, что отличие модуляции PCM и сигнала PCM состоит в том, что первая представляет собой последовательность битов, а второй — передачу этой последовательности с помощью сигналов.

2.7. Квантование с постоянным и переменным шагом

2.7.1. Статистика амплитуд при передаче речи

Передача речи — это очень важная и специализированная область цифровой связи. Человеческая речь характеризуется уникальными статистическими свойствами, одно из которых проиллюстрировано на рис. 2.17. На оси абсцисс отложены амплитуды сигнала, нормированные на среднеквадратическое значение величины таких амплитуд в типичном канале связи, а на оси ординат — вероятность. Для большинства каналов речевой связи доминируют очень низкие тона; 50% времени напряжение, характеризующее энергию обнаруженной речи, составляет менее четверти среднеквадратиче-

ского значения. Значения с большими амплитудами встречаются относительно редко; только 15% времени напряжение превышает среднеквадратическое значение. Из уравнения (2.18,б) видно, что шум квантования зависит от шага (размера интервала квантования). Если шаг квантования постоянен, квантование является *равномерным (квантованием с постоянным шагом)*. При передаче речи подобная система будет неэкономной; многие уровни квантования будут использоваться довольно редко. В системе, использующей равномерное квантование, шум квантования будет одинаковым для всех амплитуд сигнала. Следовательно, при таком квантовании отношение сигнал/шум (signal-to-noise ratio — SNR) будет хуже для сигналов низких уровней, чем для сигналов высоких уровней. *Неравномерное квантование* может обеспечить лучшее квантование слабых сигналов и грубое квантование сильных сигналов. Значит, в этом случае шум квантования может быть пропорциональным сигналу. Результатом является повышение общего отношения сигнал/шум — уменьшение шума для доминирующих слабых сигналов за счет повышения шума для редко встречающихся сильных сигналов. На рис. 2.18 сравнивается квантование слабого и сильного сигналов при равномерном и неравномерном квантовании. Ступенчатые сигналы представляют собой аппроксимации аналоговых сигналов (после введения искажения вследствие квантования). Улучшение отношения SNR для слабого сигнала, которое дает неравномерное квантование, должно быть очевидным. Неравномерное квантование может использоваться при фиксации отношения SNR для всех сигналов входного диапазона. Для сигналов речевого диапазона, динамический диапазон типичного входного сигнала составляет 40 дБ, где значение в децибелах определяется через отношение мощности P_1 к мощности P_2.

$$\text{значение в децибелах} = 10 \lg \frac{P_2}{P_1} \tag{2.21}$$

В устройстве с равномерным квантованием слабые сигналы будут иметь на 40 дБ худшее отношение SNR, чем сильные сигналы. В стандартной телефонной связи для обработки большого диапазона возможных входных уровней сигналов используется не обычное устройство с равномерным квантованием, а устройство с логарифмическим сжатием. При этом отношение сигнал/шум на выходе не зависит от распределения уровней сигнала на входе.

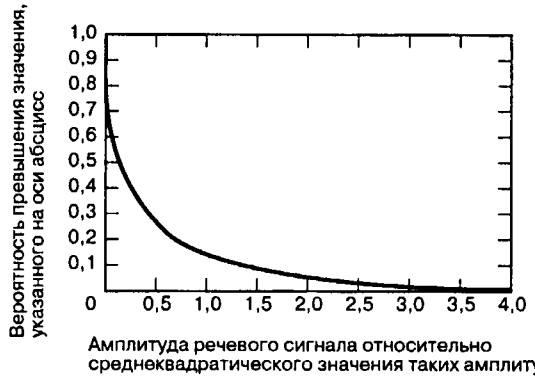


Рис. 2.17. Статистическое распределение амплитуд речи одного лица

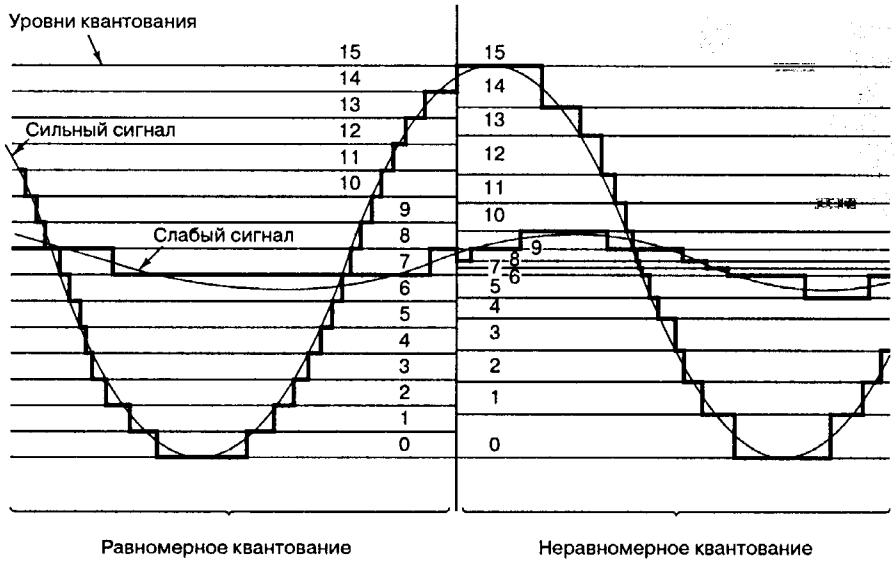


Рис. 2.18. Равномерное и неравномерное квантование сигналов

2.7.2. Неравномерное квантование

Одним из способов получения неравномерного квантования является использование устройства с неравномерным квантованием с характеристикой, показанной на рис. 2.19, *а*. Гораздо чаще неравномерное квантование реализуется следующим образом: вначале исходный сигнал деформируется с помощью устройства, имеющего логарифмическую характеристику сжатия, показанную на рис. 2.19, *б*, а потом используется устройство квантования с равномерным шагом. Для сигналов малой амплитуды характеристика сжатия имеет более крутой фронт, чем для сигналов большой амплитуды. Следовательно, изменение данного сигнала при малых амплитудах затронет большее число равномерно размещенных уровней квантования, чем то же изменение при больших амплитудах. Характеристика сжатия эффективно меняет распределение амплитуд входного сигнала, так что на выходе системы сжатия уже не существует превосходства сигналов *малых* амплитуд. После сжатия деформированный сигнал подается на вход равномерного (линейного) устройства квантования с характеристикой, показанной на рис. 2.19, *в*. После приема сигнал пропускается через устройство с характеристикой, обратной к показанной на рис. 2.19, *б* и называемой *расширением*, так что общая передача не является деформированной. Описанная пара этапов обработки сигнала (сжатие и расширение) в совокупности обычно именуется *компандированием*.

2.7.3. Характеристики компандирования

В ранних системах PCM функции сжатия были гладкими логарифмическими. Большинство современных систем использует кусочно-линейную аппроксимацию функции логарифмического сжатия. В Северной Америке характеристика устройства сжатия описывается следующим законом.

$$y = y_{max} \frac{\ln[1 + \mu(|x|/x_{max})]}{\ln(1+\mu)} \operatorname{sgn} x, \tag{2.22}$$

где

$$\operatorname{sgn} x = \begin{cases} +1 & \text{при } x \geq 0 \\ -1 & \text{при } x < 0 \end{cases}$$

μ — положительная константа, x и y — напряжения на входе и выходе, а x_{max} и y_{max} — максимальные положительные амплитуды напряжений на входе и выходе. Характеристика устройства сжатия показана на рис. 2.20, *а* для нескольких значений μ. В Северной Америке стандартным значением для μ является 255. Отметим, что $\mu = 0$ соответствует линейному усилению (равномерному квантованию).

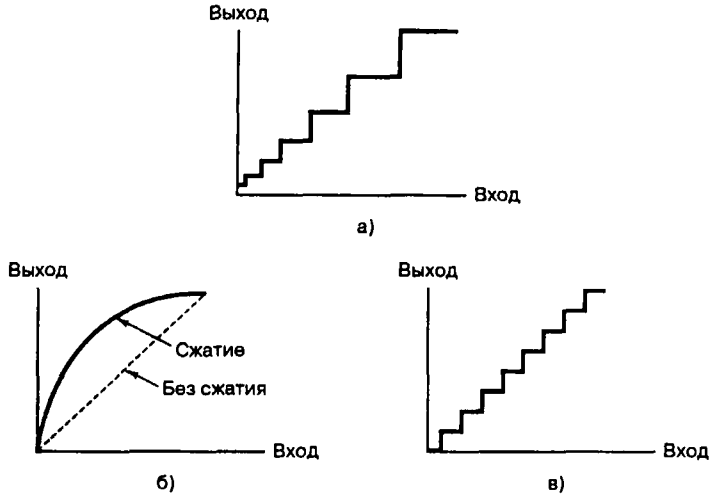


Рис. 2.19. Примеры характеристик: а) характеристика неравномерного устройства квантования; б) характеристика сжатия; в) характеристика равномерного устройства квантования

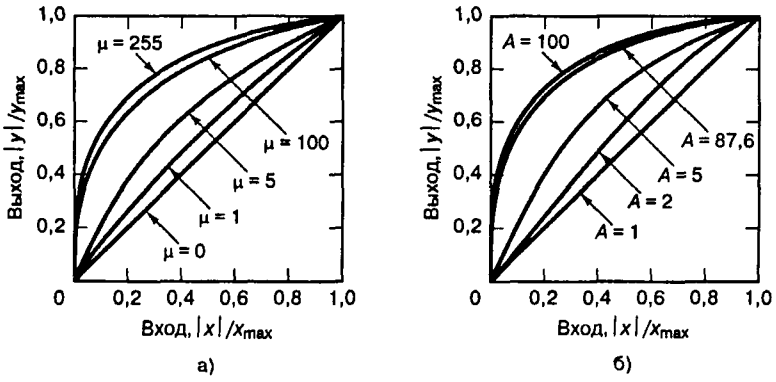


Рис. 2.20. Характеристики устройств сжатия: а) для различных значений μ; б) для различных значений A

В Европе для описания характеристики устройства сжатия используется несколько иной закон.

$$y = \begin{cases} y_{\max} \dfrac{A(|x|/x_{\max})}{1+\ln A} \operatorname{sgn} x & 0 < \dfrac{|x|}{x_{\max}} \le \dfrac{1}{A} \\ y_{\max} \dfrac{1+\ln[A(|x|/x_{\max})]}{1+\ln A} \operatorname{sgn} x & \dfrac{1}{A} < \dfrac{|x|}{x_{\max}} < 1 \end{cases} \tag{2.23}$$

Здесь A — положительная константа, а x и y такие же, как и в формуле (2.22). На рис. 2.20, *б* изображены характеристики устройств сжатия для нескольких значений A. Стандартным значением для A является величина 87,6. (Обсуждение темы равномерного и неравномерного квантования продолжается в главе 13, раздел 13.2.)

2.8. Узкополосная передача

2.8.1. Представление двоичных цифр в форме сигналов

В разделе 2.6 показывалось, как аналоговые сигналы преобразовываются в двоичные цифры посредством использования PCM. В результате этого не получается ничего "физического", только цифры. Цифры — это просто абстракция, способ описания информации, содержащейся в сообщении. Следовательно, нам необходимо иметь что-то физическое, что будет представлять цифры или "являться носителем" цифр.

Чтобы передать двоичные цифры по узкополосному каналу, будем представлять их электрическими импульсами. Подобное представление изображено на рис. 2.21. На рис. 2.21, *а* показаны разделенные во времени интервалы передачи кодовых слов, причем каждое кодовое слово является 4-битовым представлением квантованной выборки. На рис. 2.21, *б* каждая двоичная единица представляется импульсом, а каждый двоичный нуль — отсутствием импульса. Таким образом, последовательность электрических импульсов, представленная на рис. 2.21, *б*, может использоваться для передачи информации двоичного потока PCM, а значит информации, закодированной в квантованных выборках сообщения.

Задача приемника — определить в каждый момент приема бита, имеется ли импульс в канале передачи. В разделе 2.9 будет показано, что вероятность точного определения наличия импульса является функцией энергии принятого импульса (или площади под графиком импульса). Следовательно, ширину импульса T' (рис. 2.21, *б*) выгодно делать как можно больше. Если увеличить ширину импульса до максимально возможного значения (равного времени передачи бита T), то получится сигнал, показанный на рис. 2.21, *в*. Вместо того чтобы описывать этот сигнал как последовательность импульсов и их отсутствий (униполярное представление), мы можем описать его как последовательность переходов между двумя ненулевыми уровнями (биполярное представление). Если сигнал находится на верхнем уровне напряжения, он представляет двоичную единицу, а если на нижнем — двоичный нуль.

2.8.2. Типы сигналов PCM

При применении импульсной модуляции к *двоичному* символу получаем двоичный сигнал, называемый *сигналом с импульсно-кодовой модуляцией* (pulse-code modulation —

PCM). Существует несколько типов сигналов PCM; они изображены на рис. 2.22 и будут описаны ниже. В приложениях телефонной связи эти сигналы часто именуются *кодами канала* (line code).

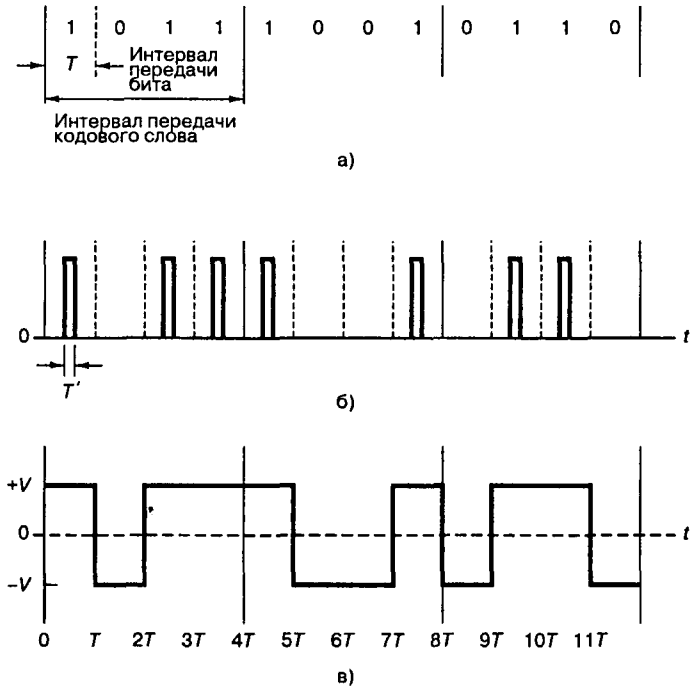


Рис. 2.21. Пример представления двоичных цифр в форме сигналов: а) последовательность PCM; б) импульсное представление последовательности PCM; в) импульсный сигнал (переход между двумя уровнями)

При применении импульсной модуляции к *недвоичному* символу получаем сигнал, называемый *M-арным импульсно-модулированным*; существует несколько типов таких сигналов. Описываются они в разделе 2.8.5, особое внимание уделяется амплитудно-импульсной модуляции (pulse-amplitude modulation — PAM). На рис. 2.1 в выделенном блоке "Узкополосная передача сигналов" показана базовая классификация сигналов PCM и *M*-арных импульсных сигналов. Сигналы PCM делятся на четыре группы.

1. Без возврата к нулю (nonreturn-to-zero — NRZ)
2. С возвратом к нулю (return-to-zero — RZ)
3. Фазовое кодирование
4. Многоуровневое бинарное кодирование

Самыми используемыми сигналами PCM являются, пожалуй, сигналы в кодировках NRZ. Группа кодировок NRZ включает следующие подгруппы: NRZ-L (L = level — уровень), NRZ-M (M = mark — метка) и NRZ-S (S = space — пауза). Кодировка NRZ-L (nonreturn-to-zero level — без возврата к нулевому уровню) широко используется в цифровых логических схемах. Двоичная единица в этом случае представляется одним уровнем напряжения, а двоичный нуль — другим.

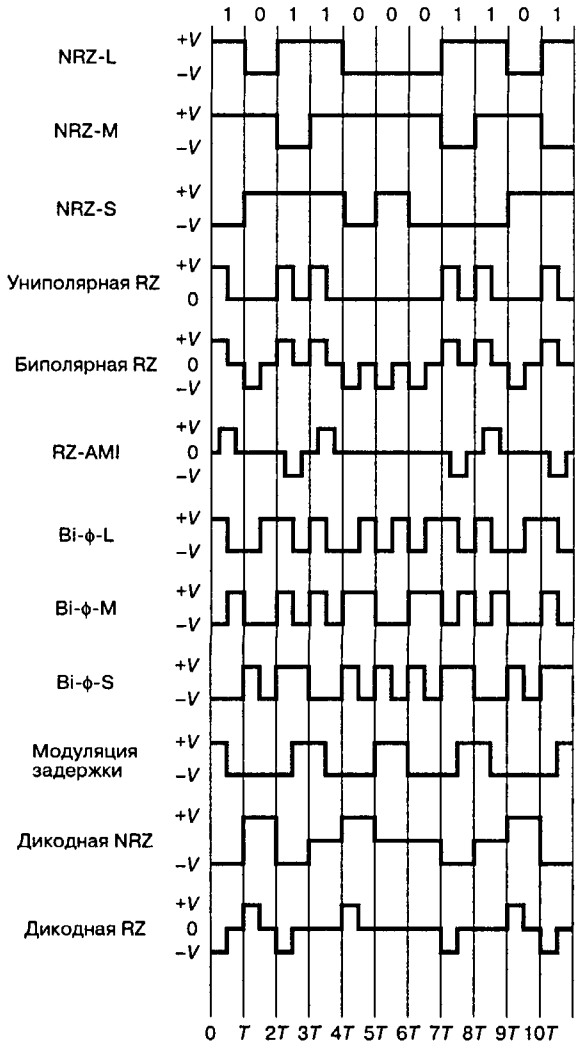


Рис. 2.22. Различные сигналы PCM

Изменение уровня происходит всякий раз при переходе в последовательности передаваемых битов от нуля к единице или от единицы к нулю. При использовании кодировки NRZ-M двоичная единица, или *метка* (mark), представляется изменением уровня, а нуль, или *пауза* (space), — отсутствием изменения уровня. Такая кодировка часто называется *дифференциальной*. Применяется кодировка NRZ-M преимущественно при записи на магнитную ленту. Кодировка NRZ-S является обратной к кодировке NRZ-M: двоичная единица представляется отсутствием изменения уровня, а двоичный нуль — изменением уровня.

Группа кодировок RZ включает униполярную кодировку RZ, биполярную кодировку RZ и кодировку RZ-AMI. Эти коды применяются при узкополосной передаче данных и магнитной записи. В униполярной кодировке RZ единица представляется наличием импульса, длительность которого составляет половину ширины бита, а

нуль — его отсутствием. В биполярной кодировке RZ единицы и нули представляются импульсами противоположных уровней, длительность каждого из которых также составляет половину ширины бита. В каждом интервале передачи бита присутствует импульс. Кодировка RZ-AMI (AMI = alternate mark inversion — с чередованием полярности) — это схема передачи сигналов, используемая в телефонных системах. Единицы представляются наличием импульсов равных амплитуд с чередующимися полярностями, а нули — отсутствием импульсов.

Группа фазового кодирования включает следующие кодировки: bi-ϕ-L (bi-phase-level — двухфазный уровень), более известная как *манчестерское кодирование* (Manchester encoding); bi-ϕ-M (bi-phase-mark); bi-ϕ-S (bi-phase-space); и *модуляция задержки* (delay modulation — DM), или *кодировка Миллера*. Схемы фазовых кодировок используются в системах магнитной записи и оптической связи, а также в некоторых спутниковых телеметрических каналах передачи данных. В кодировке bi-ϕ-L единица представляется импульсом, длительностью в половину ширины бита, расположенным в первой половине интервала передачи бита, а нуль — таким же импульсом, но расположенным во второй половине интервала передачи бита. В кодировке bi-ϕ-M в начале каждого интервала передачи бита происходит переход. Единица представляется вторым переходом в середине интервала, нуль — единственным переходом в начале интервала передачи бита. В кодировке bi-ϕ-S в начале каждого интервала также происходит переход. Единица представляется этим единственным переходом, а для представления нуля необходим второй переход в середине интервала. При модуляции задержки [4] единица представляется переходом в середине интервала передачи бита, а нуль — отсутствием иных переходов, если за ним не следует другой нуль. В последнем случае переход помещается в конец интервала передачи первого нуля. Приведенные объяснения станут понятнее, если обратиться к рис. 2.22.

Многие двоичные сигналы для кодировки двоичных данных используют три уровня, а не два. К этой группе относятся сигналы в кодировках RZ и RZ-AMI. Кроме того, сюда входят схемы, называемые *дикодной* (dicode) и *двубинарной кодировкой* (duobinary). При дикодной кодировке NRZ переходы в передаваемой информации от единицы к нулю и от нуля к единице меняют полярность импульсов; при отсутствии переходов передается сигнал нулевого уровня. При дикодной кодировке RZ переходы от единицы к нулю и от нуля к единице вызывают изменение полярности, длительностью в половину интервала импульса; при отсутствии переходов передается сигнал нулевого уровня. Подробнее трехуровневые двубинарные схемы передачи сигналов рассмотрены в разделе 2.9.

Может возникнуть вопрос, почему так много различных сигналов PCM? Неужели так много уникальных приложений требуют разнообразных кодировок для представления двоичных цифр? Причина такого разнообразия заключается в отличии производительности, которая характеризует каждую кодировку [5]. При выборе кодировки PCM внимание следует обращать на следующие параметры.

1. *Постоянная составляющая.* Удаление из спектра мощностей постоянной составляющей позволяет системе работать на переменном токе. Системы магнитной записи или системы, использующие трансформаторную связь, слабо чувствительны к гармоникам очень низких частот. Следовательно, существует вероятность потери низкочастотной информации.
2. *Автосинхронизация.* Каждой системе цифровой связи требуется символьная или битовая синхронизация. Некоторые кодировки PCM имеют встроенные функции синхронизации, помогающие восстанавливать синхронизирующий сигнал. Например, манчестерская кодировка включает переходы в середине каждого интер-

вала передачи бита, вне зависимости от передаваемого знака. Этот гарантированный переход и может использоваться в качестве синхронизирующего сигнала.

3. *Выявление ошибок.* Некоторые схемы, такие как двубинарная кодировка, предлагают средство выявления информационных ошибок без введения в последовательность данных дополнительных битов выявления ошибок.
4. *Сжатие полосы.* Такие схемы, как, например, многоуровневые кодировки, повышают эффективность использования полосы, разрешая уменьшение полосы, требуемой для получения заданной скорости передачи данных; следовательно, на единицу полосы приходится больший объем передаваемой информации.
5. *Дифференциальное кодирование.* Этот метод позволяет инвертировать полярность сигналов в дифференциальной кодировке, не затрагивая при этом процесс обнаружения данных. Это большой плюс в системах связи, в которых иногда происходит инвертирование сигналов. (Дифференциальная кодировка подробно рассмотрена в главе 4, раздел 4.5.2.)
6. *Помехоустойчивость.* Различные типы сигналов PCM могут различаться по вероятности появления ошибочных битов при данном отношении сигнал/шум. Некоторые схемы более устойчивы к шумам, чем другие. Например, сигналы в кодировке NRZ имеют лучшую достоверность передачи, чем сигналы в униполярной кодировке RZ.

2.8.3. Спектральные параметры сигналов PCM

Наиболее распространенными критериями, используемыми при сравнении кодировок PCM и выборе подходящего типа сигнала из многих доступных, являются спектральные характеристики, возможности битовой синхронизации и выявления ошибок, устойчивость к интерференции и помехам, а также цена и сложность реализации. Спектральные характеристики некоторых распространенных кодировок PCM показаны на рис. 2.23. Здесь изображена зависимость спектральной плотности мощности (измеряется в Вт/Гц) от нормированной ширины полосы, WT, где W — ширина полосы, а T — длительность импульса. Произведение WT часто называют *базой* сигнала. Поскольку скорость передачи импульсов или сигналов R_s обратна T, нормированную ширину полосы можно также выразить как W/R_s. Из последнего выражения видно, что нормированная ширина полосы измеряется в герц/(импульс/с) или в герц/(символ/с). Это относительная мера ширины полосы; она описывает, насколько эффективно используется полоса пропускания при интересующей нас кодировке. Считается, что любой тип кодировки, требующий менее 1,0 Гц для передачи одного символа в секунду, эффективно использует полосу. Примеры: модулирование задержки и двубинарная кодировка (см. раздел 2.9). Для сравнения, любая кодировка, требующая более 1,0 Гц полосы для передачи одного символа в секунду, неэффективно использует полосу. Пример: двухфазная (манчестерская) кодировка. На рис. 2.23 можно также видеть распределение энергии сигналов в различных кодировках по спектру. Например, двубинарная кодировка и схема NRZ имеют значительное число спектральных компонентов около постоянной составляющей и на низких частотах, тогда как двухфазная кодировка вообще не содержит энергии на частоте постоянной составляющей.

Важным параметром измерения *эффективности использования полосы* является отношение R/W (измеряется в бит/с/герц). Эта мера характеризует скорость передачи данных, а не скорость передачи сигналов. Для данной схемы передачи сигналов отношение R/W описывает, какой объем данных может быть передан из расчета на каждый герц доступной полосы. (Подробнее об эффективности использования полосы в главе 9.)

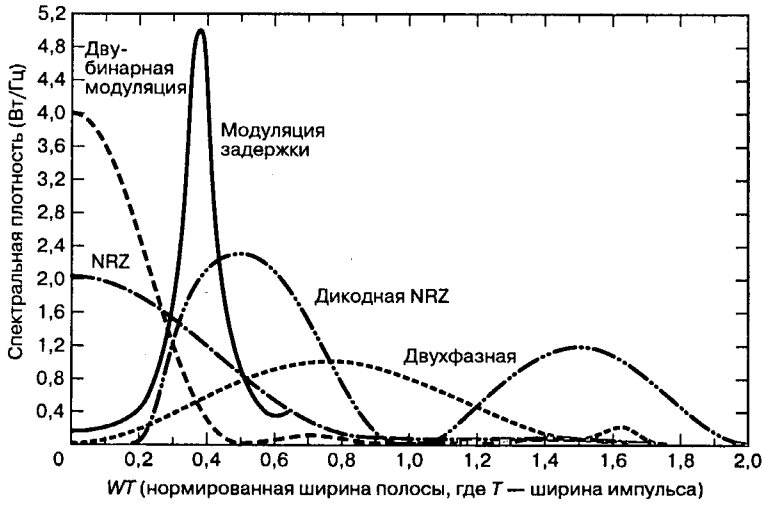


Рис. 2.23. Спектральные плотности различных кодировок PCM

2.8.4. Число бит на слово PCM и число бит на символ

До настоящего момента для разбиения битов на группы с целью формирования символов для обработки и передачи сигналов использовалось двоичное разделение ($M = 2^k$). Рассмотрим теперь аналогичное приложение, где также применима концепция $M = 2^k$. Опишем процесс форматирования аналоговой информации в двоичный поток посредством дискретизации, квантования и кодирования. Каждая аналоговая выборка преобразовывается в слово PCM, состоящее из группы битов. Размер слова PCM можно выразить через число квантовых уровней, разрешенных для каждой выборки; это равно числу значений, которое может принимать слово PCM. Квантование также можно описать числом битов, требуемых для определения этого набора уровней. Связь между числом уровней на выборку и количеством битов, необходимых для представления этих уровней, аналогична связи между размером набора символов сообщения и числом битов, необходимых для представления символа ($M = 2^k$). Чтобы различать эти два случая, изменим форму записи для сигналов PCM. Вместо $M = 2^k$ будем писать $L = 2^l$, где L — число квантовых уровней в слове PCM, а l — число битов, необходимых для представления этих уровней.

2.8.4.1. Размер слова PCM

Сколько бит нужно выделить каждой аналоговой выборке? Для цифровых телефонных каналов каждая выборка речевого сигнала кодируется с использованием 8 бит, что дает 2^8, или 256 уровней на выборку. Выбор числа уровней (или числа бит на выборку) зависит от того, какое искажение, вызванное квантованием, мы можем допустить при использовании формата PCM. Вообще, полезно вывести общую формулу, выражающую соотношение между требуемым числом бит на аналоговую выборку (размер слова PCM) и допустимым искажением, вызванным квантованием. Итак, пусть величина ошибки вследствие квантования, $|e|$, определяется как часть p удвоенной амплитуды напряжения аналогового сигнала.

$$|e| \le pV_{pp} \tag{2.24}$$

Поскольку ошибка квантования не может быть больше $q/2$, где q — интервал квантования, можем записать

$$|e|_{\max} = \frac{q}{2} = \frac{V_{pp}}{2(L-1)} \approx \frac{V_{pp}}{2L}, \tag{2.25}$$

где L — число уровней квантования. Для большинства приложений число уровней достаточно велико, так что $(L-1)$ можно заменить L, что и было сделано выше. Следовательно, из формул (2.24) и (2.25) можем записать следующее.

$$\frac{V_{pp}}{2L} \le pV_{pp} \tag{2.26}$$

$$2^l = L \ge \frac{1}{2p} \text{ уровней} \tag{2.27}$$

и

$$l \ge \log_2 \frac{1}{2p} \text{ бит} \tag{2.28}$$

Важно отметить, что мы не путаем число бит на слово PCM, обозначенное через l в уравнении (2.28), и число бит k, используемое в описании M-уровневой передачи данных. (Несколько ниже приводится пример 2.3, который поможет понять, чем отличаются эти два понятия.)

2.8.5. *M*-арные импульсно-модулированные сигналы

Существует три основных способа модулирования информации в последовательность импульсов: можно варьировать амплитуду, положение или длительность импульсов, что дает, соответственно, следующие схемы: *амплитудно-импульсная модуляция* (pulse-amplitude modulation — PAM), *фазово-импульсная модуляция* (pulse-position modulation — PPM) и *широтно-импульсная модуляция* (pulse-duration modulation — PDM или pulse-width modulation — PWM). Если информационные выборки без квантования модулируются в импульсы, получаемая импульсная модуляция называется *аналоговой*. Если информационные выборки вначале квантуются, превращаясь в символы M-арного алфавита, а затем модулируются импульсами, получаемая импульсная модуляция является цифровой, и мы будем называть ее *M-арной импульсной модуляцией*. При M-арной амплитудно-импульсной модуляции каждому из M возможных значений символов присваивается один из разрешенных уровней амплитуды. Ранее сигналы PCM описывались как двоичные, имеющие два значения амплитуды (например, кодировки NRZ, RZ). Отметим, что такие сигналы PCM, требующие всего двух уровней, представляют собой частный случай ($M = 2$) M-арной кодировки PAM. В данной книге сигналы PCM выделены (см. разделы 2.1 и 2.8.2) и рассмотрены особо, поскольку они являются наиболее популярными схемами импульсной модуляции.

M-арная фазово-импульсная модуляция (PPM) сигнала осуществляется через задержку (или упреждение) появления импульса на время, соответствующее значению информационных символов. M-арная широтно-импульсная модуляция (PDM) осуществляется посредством измерения ширины импульса на величину,

соответствующую значению символа. Для кодировок PPM и PDM амплитуда импульса фиксируется. Стоит отметить, что узкополосные модуляции с использованием импульсов имеют аналоги среди полосовых модуляций. Кодировка PAM подобна амплитудной модуляции, тогда как кодировки PPM и PDM подобны, соответственно, фазовой и частотной модуляциям. В данном разделе мы рассмотрим только M-арные сигналы PAM и сопоставим их с сигналами PCM.

Полоса пропускания, необходимая для двоичных цифровых сигналов, таких как сигналы в кодировке PCM, может быть очень большой. Как сузить требуемую полосу? Одна из возможностей — использовать *многоуровневую передачу сигналов*. Рассмотрим двоичный поток со скоростью передачи данных R бит/секунду. Чтобы не передавать импульсные сигналы для каждого отдельного бита, можно вначале разделить данные на k-битовые группы, после чего использовать для передачи $(M = 2^k)$-уровневые импульсы. При такой многоуровневой передаче сигналов, или M-арной амплитудно-импульсной модуляции, каждый импульсный сигнал может теперь представлять k-битовый символ в потоке символов, перемещающемся со скоростью R/k символов в секунду (в k раз медленнее, чем поток битов). Следовательно, при данной скорости передачи данных для уменьшения числа символов, передаваемых в секунду, может использоваться многоуровневая ($M > 2$) передача сигналов; другими словами, при уменьшении требований к ширине полосы передачи может применяться не двоичная кодировка PCM, а M-уровневая кодировка PAM. Чем мы платим за такое сужение полосы, и платим ли мы вообще чем-либо? Разумеется, ничто не достается даром, и это будет рассмотрено ниже.

Рассмотрим задачу, которую должен выполнять приемник. Он должен различать все возможные уровни каждого импульса. Одинаково ли легко приемник различает восемь возможных уровней импульса, приведенного на рис. 2.24, *а*, и два возможных уровня каждого двоичного импульса на рис. 2.24, *б*? Передача восьмиуровневого (по сравнению с двухуровневым) импульса требует большей энергии для эквивалентной эффективности обнаружения. (Достоверность обнаружения сигнала определяется отношением E_b/N_0 в приемнике.) При равной средней мощности двоичных и восьмеричных импульсов первые обнаружить проще, поскольку детектор приемника при принятии решения о принадлежности сигнала к одному из двух уровней располагает большей энергией сигнала на каждый уровень, чем при принятии решения относительно принадлежности сигнала к одному из 8 уровней. Чем расплачивается разработчик системы, если решает использовать более удобную в обнаружении двоичную кодировку PCM, а не восьмиуровневую кодировку PAM? Плата состоит в трехкратном увеличении ширины полосы для данной скорости передачи данных, по сравнению с восьмеричными импульсами, поскольку каждый восьмеричный импульс должен заменяться тремя двоичными (ширина каждого из которых втрое меньше ширины восьмеричного импульса). Может возникнуть вопрос, почему бы ни использовать двоичные импульсы той же длительности, что и восьмеричные, и разрешить запаздывание информации? В некоторых случаях это приемлемо, но для систем связи реального времени такое увеличение задержки допустить нельзя — шестичасовые новости *должны* приниматься в 6 часов. (В главе 9 будет подробно рассмотрен компромисс между мощностью сигнала и шириной полосы передачи.)

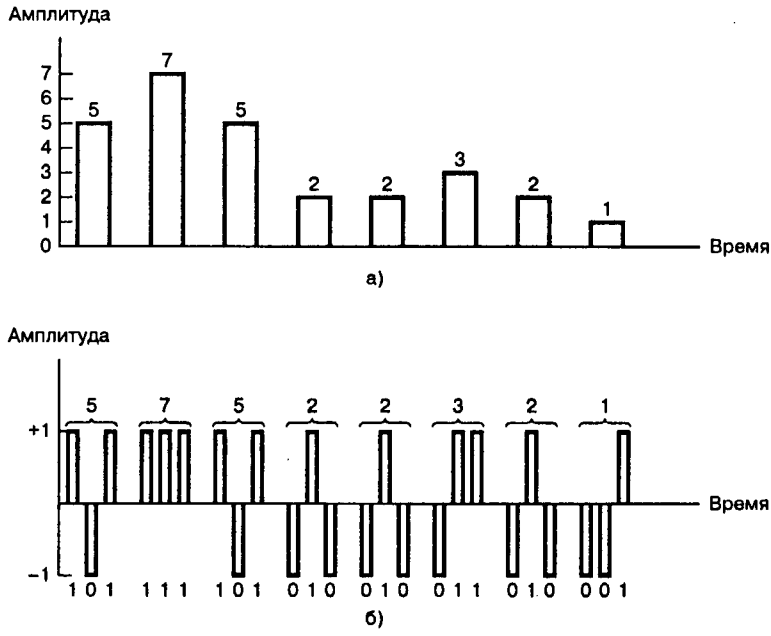


Рис. 2.24. Передача сигналов с использованием импульсно-кодовой модуляции: а) восьмиуровневая передача; б) двухуровневая передача

Пример 2.3. Уровни квантования и многоуровневая передача сигналов

Информацию в форме аналоговых сигналов с максимальной частотой f_m =3 кГц необходимо передать через систему с M-уровневой кодировкой PAM, где общее число уровней импульсов M = 16. Искажение, вызванное квантованием, не должно превышать ± 1% удвоенной амплитуды аналогового сигнала.

а) Чему равно минимальное число бит в выборке или слове PCM, которое можно использовать при оцифровывании аналогового сигнала?

б) Чему равны минимальная требуемая частота дискретизации и получаемая при этом скорость передачи битов?

в) Чему равна скорость передачи импульсов в кодировке PAM (или символов)?

г) Если ширина полосы передачи (включая фильтрацию) равна 12 кГц, чему будет равно эффективное использование полосы для этой системы?

В этом примере мы имеем дело с двумя типами *уровней*: несколькими уровнями квантования, необходимыми для удовлетворения требований ограничения искажения, и 16 уровнями импульсов в кодировке PAM.

Решение

а) С помощью формулы (2.28) вычисляем следующее.

$$l \geq \log_2 \frac{1}{0{,}02} = \log_2 50 \approx 5{,}6$$

Следовательно, $l = 6$ уровней удовлетворяют требованиям, относящимся к искажению.

б) Используя критерий Найквиста, получаем минимальную частоту дискретизации $f_s = 2f_m$ = 6000 выборок/секунду. Из п. а получаем, что каждая выборка — это 6-битовое слово в кодировке PCM. Следовательно, скорость передачи битов $R = lf_s$ = 36 000 бит/с.

в) Поскольку нужно использовать многоуровневые импульсы с $M = 2^k = 16$ уровнями, то $k = \log_2 16 = 4$ бит/символ. Следовательно, поток битов разбивается на группы по 4 бита с целью формирования новых 16-уровневых цифр PAM, и полученная скорость передачи символов R_s равна $R/k = 36\,000/4 = 9\,000$ символов/с.

г) Эффективность использования полосы — это отношение пропускной способности к ширине полосы в герцах, R/W. Поскольку $R = 36\,000$ бит/с, а $W = 12$ кГц, получаем $R/W = 3$ бит/с/Гц.

2.9. Корреляционное кодирование

В 1963 году Адам Лендер (Adam Lender) [6, 7] показал, что с нулевой межсимвольной интерференцией можно передавать $2W$ символов/с, используя теоретическую минимальную полосу в W герц, без применения фильтров с высокой добротностью. Он использовал так называемый метод *двубинарной передачи сигналов* (duobinary signaling), также известный как *корреляционное кодирование* (correlative coding) и *передача сигналов с частичным откликом* (partial response signaling). Основной идеей, лежащей в основе двубинарного метода, является введение некоторого управляемого объема межсимвольной интерференции в поток данных, вместо того чтобы пытаться устранить ее полностью. Введя корреляционную интерференцию между импульсами и изменив процедуру обнаружения, Лендер, по сути, "уравновесил" интерференцию в детекторе и, следовательно, получил идеальное заполнение в 2 символа/с/Гц, что ранее считалось неосуществимым.

2.9.1. Двубинарная передача сигналов

Чтобы понять, как двубинарная передача сигналов вводит контролируемую межсимвольную интерференцию, рассмотрим модель процесса. Операцию двубинарного кодирования можно рассматривать как реализацию схемы, показанной на рис. 2.25. Предположим, что последовательность двоичных символов $\{x_k\}$ необходимо передать на скорости R символов/с через систему, имеющую идеальный прямоугольный спектр ширины $W = R/2 = 1/2T$ Гц. Вы можете спросить: чем этот квадратный спектр на рис. 2.25 отличается от нереализуемой характеристики Найквиста? Он имеет ту же идеальную характеристику, но дело в том, что мы не пытаемся реализовать идеальный прямоугольный фильтр. На рис. 2.25 изображена эквивалентная модель, используемая для разработки фильтра, который легче аппроксимировать. До подачи на идеальный фильтр импульсы, как показано на рисунке, проходят через простой цифровой фильтр. Цифровой фильтр вносит задержку, длительностью в одну цифру; к каждому поступающему импульсу фильтр добавляет значение предыдущего импульса. Другими словами, с выхода цифрового фильтра поступает сумма двух импульсов. Каждый импульс последовательности $\{y_k\}$, получаемой на выходе цифрового фильтра, можно выразить следующим образом.

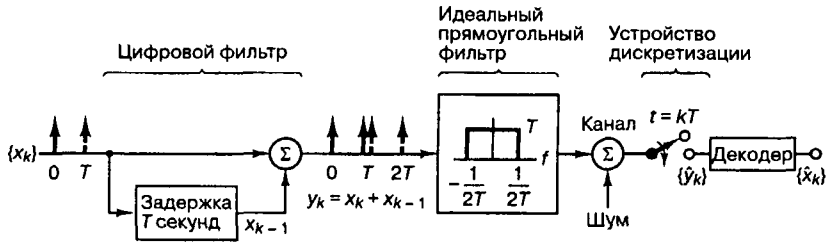


Рис. 2.25. Двубинарная передача сигналов

$$y_k = x_k + x_{k-1} \tag{2.29}$$

Следовательно, амплитуды импульсов $\{y_k\}$ не являются независимыми; каждое значение y_k использует предыдущее значение выходного сигнала. Межсимвольная интерференция, вносимая в каждую цифру y_k, проявляется только от предыдущей цифры x_{k-1}. Эту корреляцию между амплитудами импульсов $\{y_k\}$ можно рассматривать как управляемую межсимвольную интерференцию, введенную двубинарным кодированием. Управляемая интерференция составляет суть этого нового метода, поскольку в детекторе она может удаляться так же легко, как была введена. Последовательность $\{y_k\}$ проходит через идеальный фильтр Найквиста, который не вводит новой межсимвольной интерференции. В устройстве квантования приемника, показанном на рис. 2.25, мы надеемся (при отсутствии помех) точно восстановить последовательность $\{y_k\}$. Выходную последовательность $\{y_k\}$, подверженную воздействию шума, обозначим через $\{y'_k\}$. Удаление управляемой интерференции с помощью двубинарного декодера даст восстановленную оценку $\{x_k\}$, которую мы будем обозначать через $\{x'_k\}$.

2.9.2. Двубинарное декодирование

Если двоичная цифра x_k равна ±1, то, используя формулу (2.29), видим, что y_k может принимать одно из трех значений: +2, 0 или –2. Двубинарный код дает трехуровневый выход: в общем случае, для M-уровневой кодировки передача сигналов с частичным откликом дает на выходе $2M - 1$ уровней. Процедура декодирования включает процесс, обратный процедуре кодирования, который именуется вычитанием x_{k-1} решений из y_k цифр. Рассмотрим следующий пример кодирования/декодирования.

Пример 2.4. Двубинарное кодирование и декодирование

Воспользуемся формулой (2.29) для демонстрации двубинарного кодирования и декодирования следующей последовательности: $\{x_k\}$ = 0 0 1 0 1 1 0. Первый бит последовательности будем считать начальной цифрой, а не частью информационной последовательности.

Решение

Последовательность двоичных цифр $\{x_k\}$	0	0	1	0	1	1	0
Биполярные амплитуды $\{x_k\}$	−1	−1	+1	−1	+1	+1	−1
Правило кодирования: $y_k = x_k + x_{k-1}$		−2	0	0	0	2	0
Правило декодирования	Если $y'_k = 2$, то $x'_k = +1$ (или двоичная единица) Если $y'_k = -2$, то $x'_k = -1$ (или двоичный нуль) Если $y'_k = 0$, взять число, противоположное предыдущему						
Декодированная биполярная последовательность $\{x'_k\}$	−1	+1	−1	+1	+1	−1	
Декодированная бинарная последовательность $\{x'_k\}$	0	1	0	1	1	0	

Правило принятия решения просто реализует вычитание каждого решения x'_{k-1} из каждого y'_k. Одним из недостатков этого метода обнаружения является то, что при появлении ошибка имеет тенденцию к распространению, вызывая дальнейшие ошибки (причина в том, что текущее решение зависит от предыдущих). Избежать этого позволяет метод *предварительного кодирования*.

2.9.3. Предварительное кодирование

Предварительное кодирование выполняется посредством первоначального дифференциального кодирования бинарной последовательности $\{x_k\}$ в новую бинарную последовательность $\{w_k\}$, для чего используется выражение

$$w_k = x_k \oplus w_{k-1}, \tag{2.30}$$

где символ "$\oplus$" представляет сложение двоичных цифр по модулю 2 (эквивалентно операции *исключающего ИЛИ*). Сложение по модулю 2 имеет следующие правила.

$$0 \oplus 0 = 0$$
$$0 \oplus 1 = 1$$
$$1 \oplus 0 = 1$$
$$1 \oplus 1 = 0$$

Затем двоичная последовательность $\{w_k\}$ преобразовывается в последовательность биполярных импульсов, и операция кодирования проходит так же, как было показано в примере 2.4. В то же время, как показано ниже, в примере 2.5 при выполнении предварительного кодирования процесс обнаружения отличается от обнаружения в обычной двубинарной схеме. Схема предварительного кодирования показана на рис. 2.26; стоит обратить внимание на то, что сложение по модулю 2, дающее предварительно кодированную последовательность $\{w_k\}$, выполняется над *двоичными* цифрами, а цифровая фильтрация, результатом которой является последовательность $\{y_k\}$, — над *биполярными* импульсами.

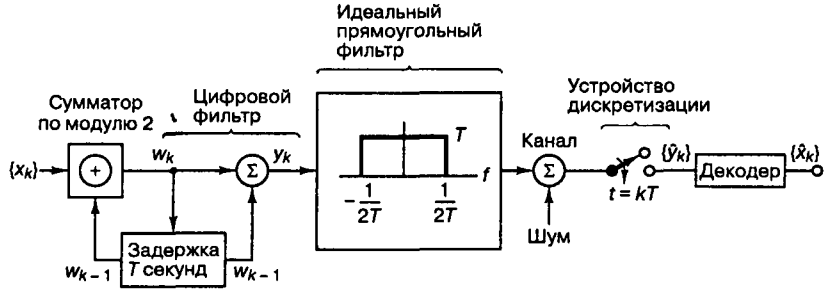


Рис. 2.26. Передача сигналов с предварительным кодированием

Пример 2.5. Двубинарное предварительное кодирование

Проиллюстрируем правила двубинарного кодирования и декодирования при использовании предварительной дифференциальной кодировки, определенной формулой (2.30). Будем использовать ту же последовательность $\{x_k\}$, что и в примере 2.4.

Решение

Последовательность двоичных цифр $\{x_k\}$	0	0	1	0	1	1	0
Предварительно кодированная последовательность $w_k = x_k \oplus w_{k-1}$	0	0	1	1	0	1	1
Биполярная последовательность $\{w_k\}$	−1	−1	+1	+1	−1	+1	+1
Правило кодирования: $y_k = w_k + w_{k-1}$		−2	0	+2	0	0	+2
Правило декодирования:	Если $y'_k = \pm 2$, то x'_k = двоичный нуль						
	Если $y'_k = 0$, то x'_k = двоичная единица						
Декодированная бинарная последовательность $\{x_k\}$	0	1	0	1	1	0	

Предварительное дифференциальное кодирование позволяет декодировать последовательность $\{y'_k\}$ путем принятия решения по каждой принятой выборке отдельно, не обращаясь к предыдущим, которые могут быть ошибочными. Преимущество заключается в том, что при возникновении из-за помех ошибочной цифры ошибка не будет распространяться на другие цифры. Отметим, что первый бит двоичной последовательности $\{w_k\}$, подвергаемой дифференциальному кодированию, выбирается произвольно. Если бы начальный бит последовательности $\{w_k\}$ был выбран равным 1, а не 0, результат декодирования был бы таким же.

2.9.4. Эквивалентная двубинарная передаточная функция

В разделе 2.9.1 двубинарная передаточная функция реализовывалась как цифровой фильтр, вводящий задержку длительностью в одну цифру, за которым следовала идеальная прямоугольная передаточная функция. Рассмотрим эквивалентную модель. Фурье-образ задержки можно записать как $e^{-2\pi ifT}$ (см. раздел А.3.1); следовательно, первый цифровой фильтр на рис. 2.25 можно описать следующей частотной характеристикой.

$$H_1(f) = 1 + e^{-2\pi ifT} \tag{2.31}$$

Передаточная функция идеального прямоугольного фильтра имеет следующий вид.

$$H_2(f) = \begin{cases} T & \text{при } |f| < \dfrac{1}{2T} \\ 0 & \text{для других } |f| \end{cases} \tag{2.32}$$

Таким образом, общая эквивалентная передаточная функция цифрового и идеального прямоугольного фильтров дается выражением

$$\begin{aligned} H_e(f) &= H_1(f)H_2(f) \quad \text{при } |f| < \frac{1}{2T} \\ &= (1 + e^{2\pi ifT})T = \\ &= T(e^{\pi ifT} + e^{-\pi ifT})e^{-\pi ifT}, \end{aligned} \tag{2.33}$$

так что

$$|H_e(f)| = \begin{cases} 2T\cos \pi fT & \text{при } |f| < \dfrac{1}{2T} \\ 0 & \text{для других } |f| \end{cases} \tag{2.34}$$

Таким образом, $H_e(f)$, составная передаточная функция последовательности цифрового и прямоугольного фильтров, последовательно выравнивает край полосы пропускания, как показано на рис. 2.27, *а*. Передаточную функцию можно аппроксимировать, используя для этого реализуемый аналоговый фильтр; отдельный цифровой фильтр не нужен. Двубинарный эквивалент $H_e(f)$ называется *косинусоидальным фильтром* [8]. Этот фильтр не следует путать с *фильтром с характеристикой типа приподнятого косинуса* (описанным в главе 3, раздел 3.3.1.) Соответствующая импульсная характеристика $h_e(t)$ получается, если взять Фурье-образ функции $H_e(f)$, описанной в формуле (2.33).

$$h_e(t) = \operatorname{sinc}\left(\frac{t}{T}\right) + \operatorname{sinc}\left(\frac{t-T}{T}\right) \tag{2.35}$$

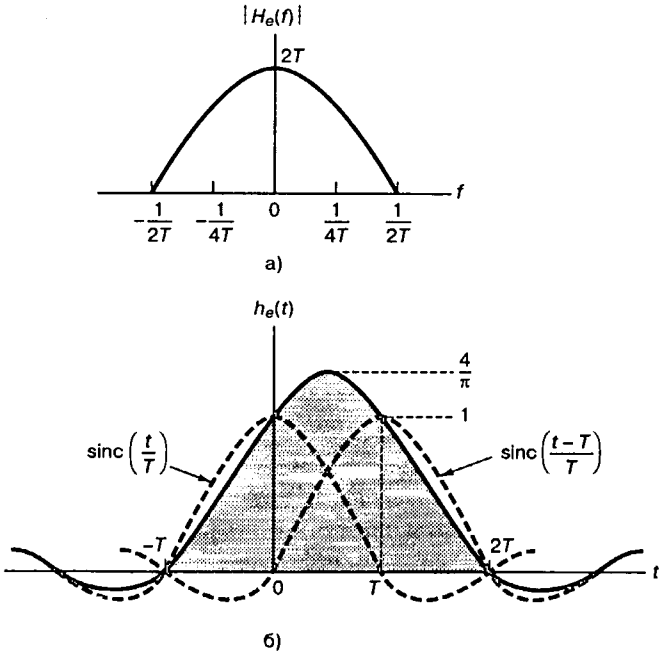


Рис. 2.27. Двубинарная передаточная функция и форма импульса: а) косинусоидальный фильтр; б) импульсная характеристика косинусоидального фильтра

Эта функция изображена на рис. 2.27, *б*. Для импульса $\delta(t)$, поданного на вход схемы, изображенной на рис. 2.25, на выход поступит сигнал $h_e(t)$ соответствующей полярности. Отметим, что в каждом T-секундном интервале имеется всего две ненулевые выборки, которые вносят вклад в управляемую межсимвольную интерференцию с соседними битами. Внесенная межсимвольная интерференция устраняется путем использования процедуры декодирования, описанной в разделе 2.9.2. Хотя косинусоидальный фильтр не является причинным, а следовательно нереализуем, его можно легко аппроксимировать. Реализацию двубинарного метода с предварительным кодированием, описанного в разделе 2.9.3, можно выполнить следующим образом. Вначале двоичная последовательность $\{x_k\}$ с использованием дифференциального кодирования превращается в последовательность $\{w_k\}$ (см. пример 2.5). Затем последовательность импульсов $\{w_k\}$ фильтруется схемой с эквивалентной косинусоидальной характеристикой, описанной в формуле (2.34).

2.9.5. Сравнение бинарного и двубинарного методов передачи сигналов

Двубинарный метод вводит корреляцию между амплитудами импульсов, тогда как критерий Найквиста предполагает независимость амплитуд передаваемых импульсов. Выше показывалось, что двубинарная передача сигналов может использовать введенную корреляцию для получения передачи без межсимвольной интерференции, требуя при этом меньшую полосу, чем пришлось бы использовать в ином случае. Можно ли получить это преимущество без сопутствующих недостатков? К сожалению, нет. Практически всегда при принятии конструкторского решения требуется искать приемлемый компромисс. Выше демонстрировалось, что двубинарное кодирование требует трех уровней, а не двух, как при обычном бинарном кодировании. Вспомним раз-

дел 2.8.5, где мы сравнивали производительность и требуемую мощность сигнала при выборе между восьмиуровневой кодировкой PAM и двухуровневой PCM. При фиксированной мощности сигнала принятие правильного решения обратно пропорционально числу уровней сигнала, которые необходимо различать. Следовательно, не должно удивлять то, что, хотя двубинарная передача сигналов позволяет получить нулевую межсимвольную интерференцию при минимальной ширине полосы, такая схема требует большей мощности, чем бинарная передача сигналов для получения равносильного сопротивления шуму. Для данной вероятности появления ошибочного бита (P_B) двубинарная схема передачи сигналов требует приблизительно на 2,5 дБ большего отношения сигнал/шум, чем бинарная схема, используя при этом всего лишь $1/(1 + r)$ полосы, требуемой бинарной схемой [7], где r — сглаживание фильтра.

2.9.6. Полибинарная передача сигналов

Двубинарная передача сигналов может быть расширена на большее, чем три, количество уровней, что приводит к большей эффективности использования полосы; называются подобные системы *полибинарными* [7, 9]. Предположим, что бинарное сообщение с двумя сигнальными уровнями преобразовывается в сигнал с j уровнями, последовательно пронумерованными от нуля до $(j - 1)$. Преобразование двубинарного сигнала в полибинарный проходит в два этапа. Вначале исходная последовательность $\{x_k\}$, состоящая из двоичных нулей и единиц, преобразовывается в другую бинарную последовательность $\{y_k\}$. Текущее двоичное число последовательности $\{y_k\}$ формируется путем сложения по модулю 2 $(j - 2)$ непосредственно предшествующих цифр последовательности $\{y_k\}$ и текущего числа x_k. Например, пусть

$$y_k = x_k \oplus y_{k-1} \oplus y_{k-2} \oplus y_{k-3}, \tag{2.36}$$

где x_k представляет входящие двоичные цифры, а y_k — k-ю кодируемую цифру. Поскольку выражение включает $(j - 2) = 3$ бит, предшествующих y_k, имеем $j = 5$ сигнальных уровней. Далее двоичная последовательность $\{y_k\}$ преобразовывается в серию полибинарных импульсов $\{z_k\}$, для чего текущий бит последовательности $\{y_k\}$ *алгебраически* складывается с $(j - 2)$ предыдущими битами последовательности $\{y_k\}$. Следовательно, z_k по модулю 2 равно x_k; и двоичные элементы один и нуль отображаются в импульсы с четными и нечетными значениями последовательности $\{z_k\}$. Отметим, что каждая цифра $\{z_k\}$ может обнаруживаться независимо, несмотря на сильную корреляцию между битами. Главным преимуществом подобной схемы передачи сигналов является перераспределение спектральной плотности исходной последовательности $\{x_k\}$ в пользу низких частот, что, в свою очередь, повышает эффективность использования системы.

2.10. Резюме

В данной главе рассмотрен первый важный этап преобразований, выполняемых в любой системе цифровой связи, — преобразование исходной информации (текстовой и аналоговой) в форму, совместимую с цифровой системой. Здесь описаны различные аспекты дискретизации, квантования (с постоянным и переменным шагом) и импульсно-кодовой модуляции (pulse code modulation — PCM). Рассмотрен также выбор кодировки для передачи через канал узкополосных сигналов. Кроме того, описано

введение контролируемого объема межсимвольной интерференции для улучшения эффективности использования полосы за счет повышения мощности.

Литература

1. Black H. S. *Modulation Theory.* D. Van Nostrand Company, Princeton, N. J., 1953.
2. Oppenheim A. V. *Application of Digital Signal Processing.* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1978.
3. Stiltz H., ed. *Aerospace Telemetry.* Vol. 1, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1961, p. 179.
4. Hecht M. and Guida A. *Delay Modulation.* Proc. IEEE, vol. 57, n. 7, July, 1969, pp. 1314–1316.
5. Deffebach H. L. and Frost W. O. *A Survey of Digital Baseband Signaling Techniques.* NASA Technical Memorandum NASATM X-64615, June, 30, 1971.
6. Lender A. *The Duobinary Technique for High Speed Data Transmission.* IEEE Trans. Commun. Electron., vol. 82, May, 1963, pp. 214–218.
7. Lender A. *Correlative (Partial Response) Techniques and Applications to Digital Radio Systems*; in K. Feher. *Digital Communications: Microwave Applications.* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1981, Chap. 7.
8. Couch L. W., II. *Digital and Analog Communication Systems.* Macmillan Publishing Company, New York, 1982.
9. Lender A. *Correlative Digital Communication Techniques.* IEEE Trans. Commun. Technol., December, 1964, pp. 128–135.

Задачи

2.1. Необходимо передать слово "HOW" с использованием восьмеричной системы.

а) Закодируйте слово "HOW" в последовательность битов, используя 7-битовый код ASCII, причем с целью выявления ошибок каждый знак дополняется восьмым битом. Значение этого бита выбирается так, чтобы число единиц среди всех 8 бит было четным. Сколько всего битов содержит сообщение?

б) Разделите поток битов на $k = 3$-битовые сегменты. Представьте каждый из 3-битовых сегментов восьмеричным числом (символом). Сколько восьмеричных символов имеется в сообщении?

в) Если бы в системе использовалась 16-уровневая модуляция, сколько символов понадобилось бы для представления слова "HOW"?

г) Если бы в системе применялась 256-уровневая модуляция, сколько символов понадобилось бы для представления слова "HOW"?

2.2. Нужно передавать данные со скоростью 800 знаков/с, причем каждый символ представляется соответствующим 7-битовым кодовым словом ASCII, за которым следует восьмой бит выявления ошибок, как в задаче 2.1. Используется многоуровневая ($M = 16$) кодировка PAM.

а) Чему равна эффективная скорость передачи битов?

б) Чему равна скорость передачи символов?

2.3. Необходимо передать 100-знаковое сообщение за 2 с, используя 7-битовую кодировку ASCII и восьмой бит выявления ошибок, как в задаче 2.1. Используется многоуровневая ($M = 32$) кодировка PAM.

а) Вычислите эффективную скорость передачи битов и передачи символов.

б) Повторите п. а для 16-уровневой кодировки PAM, восьмиуровневой кодировки PAM, четырехуровневой кодировки PAM и бинарной кодировки PCM.

2.4. Дан аналоговый сигнал, который выбирался с частотой Найквиста f_s посредством естественной дискретизации. Докажите, что сигнал, пропорциональный исходному сигналу, может быть восстановлен из выборок с использованием метода, показанного на рис. 32.1. Параметр mf_s — это частота локального осциллятора, причем m — целое.

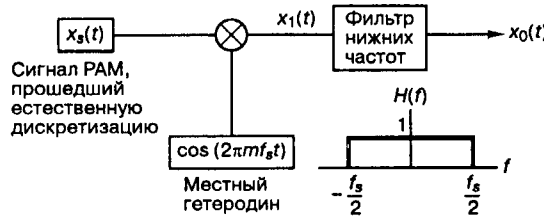


Рис. 32.1

2.5. Аналоговый сигнал выбирается с частотой Найквиста $1/T_s$ и квантуется с использованием L уровней квантования. Затем полученный цифровой сигнал передается по некоторому каналу.

а) Покажите, что длительность T одного бита передаваемого двоично-кодированного сигнала должна удовлетворять условию $T \leq T_s/(\log_2 L)$.

б) Когда имеет место равенство?

2.6. Определите число уровней квантования при следующем количестве битов на выборку данного кода PCM:

а) 5;

б) 8;

в) x.

2.7. Определите максимальную частоту дискретизации, необходимую для выборки и точного восстановления сигнала $[x(t) = \sin(6280t)]/(6280t)$.

2.8. Рассмотрим аудиосигнал, спектральные компоненты которого ограничены полосой частот от 300 до 3 300 Гц. Предположим, что для создания сигнала PCM используется частота дискретизации 8 000 выборок/с. Предположим также, что отношение пиковой мощности сигнала к средней мощности шума квантования должно быть равным 30 дБ.

а) Чему равно минимальное число уровней квантования с равномерным шагом и минимальное число битов на выборку?

б) Вычислите ширину полосы системы (определяемую как ширину основного спектрального лепестка сигнала), необходимую для обнаружения подобного сигнала PCM.

2.9. Сигнал $x(t) = 10\cos(1000t + \pi/3) + 20\cos(2000t + \pi/6)$ равномерно выбирается для цифровой передачи.

а) Чему равен максимальный разрешенный интервал между выборками, обеспечивающий безупречное воспроизведение сигнала?

б) Если необходимо воспроизвести 1 час подобного сигнала, сколько необходимо запомнить выборок?

2.10. а) Сигнал, ограниченный полосой 50 кГц, выбирается каждые 10 мкс. Покажите графически, что эти выборки единственным образом определяют сигнал. (Для простоты используйте синусоидальный сигнал. Избегайте выборок в точках, где сигнал равен нулю.)

б) Предположим, что выборки производятся не каждые 10 мкс, а каждые 30 мкс. Покажите графически, что подобные выборки могут определять сигнал, отличный от исходного.

2.11. Используйте метод свертки для иллюстрации следствия недостаточной выборки $x(t) = \cos 2\pi f_0 t$ при частоте дискретизации $f_s = 3/2f_0$.

2.12. Наложение не происходит, если частота дискретизации больше удвоенной ширины полосы сигнала. В то же время сигналов со строго ограниченной полосой не существует. Таким образом, наложение присутствует всегда.

а) Предположим, что фильтрованный сигнал имеет спектр, который описывается фильтром Баттерворта шестого порядка и верхней частотой среза $f_u = 1000$ Гц. Ка-

кая частота дискретизации необходима для снижения наложения до точки –50 дБ в спектре мощностей.

б) Повторите п. а) для фильтра Баттерворта двенадцатого порядка.

2.13. а) Изобразите схематично характеристику сжатия для $\mu = 10$, для системы, диапазон входящих напряжений которой принадлежит интервалу от –5 до +5 В.

б) Нарисуйте соответствующую характеристику расширения.

в) Изобразите характеристику 16-уровневого устройства квантования с неравномерным шагом, соответствующую характеристике сжатия при $\mu = 10$.

2.14. Необходимо передать информацию в форме аналогового сигнала, максимальная частота которого $f_m = 4000$ Гц, используя для этого 16-уровневую систему амплитудно-импульсной модуляции. Искажение, вызванное квантованием, не должно превышать $\pm 1\%$ удвоенной амплитуды аналогового сигнала.

а) Чему равно минимальное число бит на выборку или на слово PCM, которое может использоваться в этой системе?

б) Чему равна минимальная требуемая частота дискретизации и получаемая в результате скорость передачи битов?

в) Чему равна скорость передачи шестнадцатеричных символов PAM?

2.15. Сигнал в диапазоне частот 300–3300 Гц имеет удвоенную амплитуду 10 В. Он выбирается с частотой 8000 выборок/с, а выборки квантуются в 64 равномерно расположенных уровня. Вычислите и сравните ширину полос и отношения пиковой мощности сигнала к среднеквадратическому шуму квантования, если квантованные выборки передаются или как бинарные, или как четырехуровневые импульсы. Считайте, что ширина полосы системы определяется основным спектральным лепестком сигнала.

2.16. В цифровой аудиосистеме проигрывания компакт-дисков аналоговый сигнал оцифровывается так, что отношение пиковой мощности сигнала к пиковой мощности шума квантования не менее 96 дБ. Частота дискретизации — 44,1 тысяча выборок в секунду.

а) Сколько необходимо уровней квантования аналогового сигнала, чтобы $(S/N_q)_{пик} =$ 96 дБ?

б) Какое число бит на выборку необходимо при таком числе уровней?

в) Чему равна скорость передачи данных в бит/с?

2.17. Вычислите разницу в требуемой мощности между двумя сигналами PCM в кодировках NRZ и RZ, предполагая, что обе схемы имеют одинаковые скорости передачи и вероятности появления ошибочного бита. Предполагается, что сигналы равновероятны и разница между уровнями высокого и низкого напряжений одинакова для обеих схем. Можно ли отдать предпочтение какой-либо из схем, если рассматривать их с точки зрения требуемой мощности? Если да, то какие имеются недостатки у этой схемы?

2.18. В 1962 году компания AT&T первой предложила цифровую телефонную передачу, названную службой T1. Каждый кадр T1 разбивается на 24 канала (интервала времени). Каждый интервал содержит 8 бит (одна речевая выборка) и один бит для выравнивания. Кадр выбирается с частотой Найквиста 8000 выборок/с, а ширина полосы, используемая для передачи составного сигнала, равна 386 кГц. Определите для этой схемы эффективность использования полосы (в бит/с/Гц).

2.19. а) Предположим, требуется система цифровой передачи, в которой искажение, вызванное квантованием, не превышало бы $\pm 2\%$ удвоенного напряжения аналогового сигнала. Если ширина полосы аудиосигнала и разрешенная полоса передачи равны по 4000 Гц, а выборка ведется с частотой Найквиста, какая необходима эффективность использования полосы (в бит/с/Гц).

б) Повторите п. а для ширины полосы аудиосигнала 20 кГц (большая точность воспроизведения) при той же доступной полосе 4000 Гц.

Вопросы для самопроверки

2.1. Назовите общие особенности и отличия терминов "форматирование" и "кодирование источника" (см. введение).

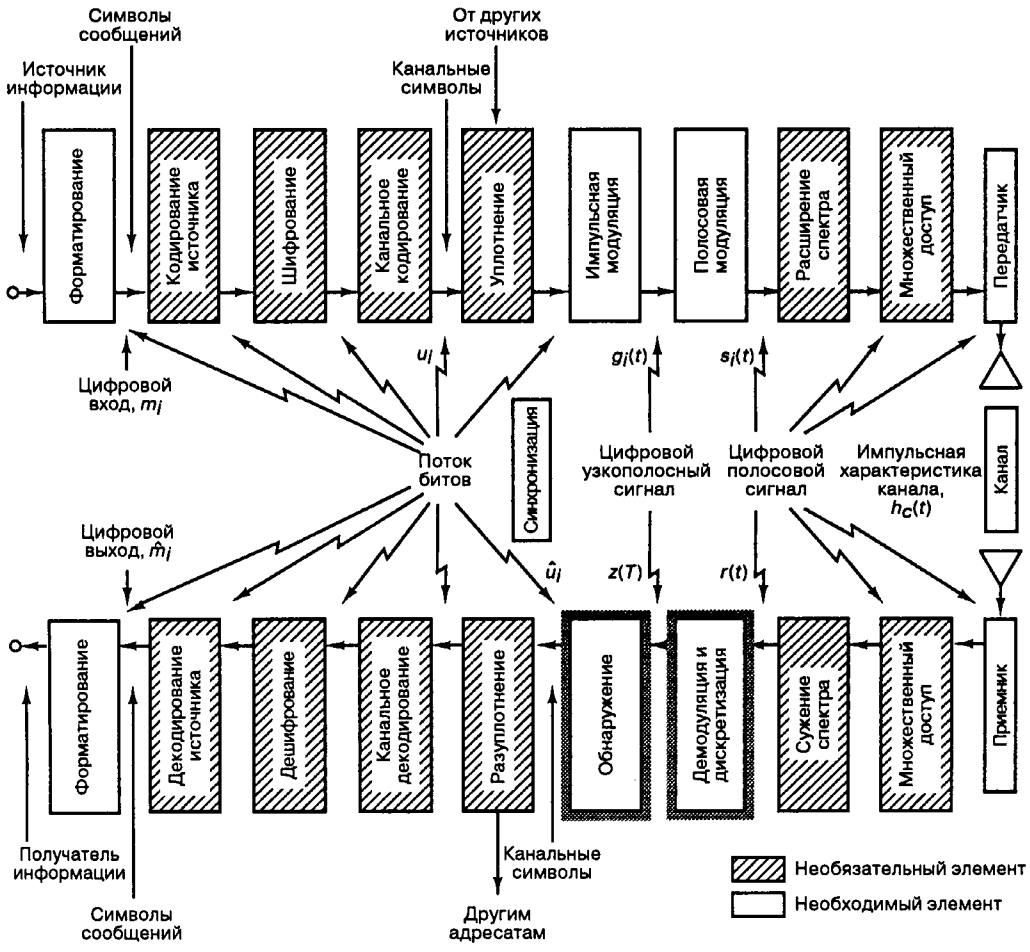
2.2. Почему в процессе *форматирования* информации зачастую желательна *выборка с запасом* (см. раздел 2.4.3)?

2.3. Как при использовании импульсно-кодовой модуляции (pulse-code modulation — PCM) для оцифровывания аналоговой информации можно увеличить один из следующих параметров за счет других: *точность воспроизведения*, *ширина полосы* и *задержка* (см. раздел 2.6)?

2.4. Почему зачастую предпочтительнее использовать единицы нормированной ширины полосы WT, а не самой ширины полосы (см. раздел 2.8.3)?

ГЛАВА 3

Узкополосная демодуляция/обнаружение



При узкополосной передаче принимаемые сигналы уже имеют форму импульсов. Может возникнуть вопрос, зачем же тогда для восстановления импульсных сигналов нужен демодулятор? Ответ связан с тем, что форма принимаемых импульсов, как правило, отличается от идеальной, когда длительность каждого импульса точно равна одному интервалу передачи символа. Фильтрация в передатчике и канале обычно приводит к тому, что принятая последовательность импульсов искажается межсимвольной интерференцией (intersymbol interference — ISI) и появляется в виде аморфного "смазанного" сигнала, не совсем готового к дискретизации и обнаружению. Задачей демодулятора (принимающего фильтра) является восстановление исходного импульса с максимально возможным отношением сигнал/шум (signal-to-noise ratio — SNR) без какой-либо межсимвольной интерференции. Для достижения этого используется метод выравнивания (equalization), рассмотренный в данной главе. Стоит отметить, что не для всех типов каналов связи процесс выравнивания является обязательным. Но все же нужно заметить, что выравнивание включает в себя набор специальных методов обработки сигнала, позволяющих компенсировать введенную каналом интерференцию, поэтому этот этап является важным для всех систем.

Полосовая модель процесса обнаружения, описанная в главе 4, практически идентична узкополосной модели, рассмотренной в данной главе. Дело в том, что принятый полосовой сигнал вначале преобразуется в узкополосный, после чего наступает этап окончательного обнаружения. Для линейных систем математические методы обнаружения не зависят от смещения частоты. Фактически *теорему эквивалентности* можно определить следующим образом: выполнение полосовой линейной обработки сигнала с последующим частотным преобразованием сигнала (превращением полосового сигнала в узкополосный) дает те же результаты, что и наложение сигнала с последующей узкополосной линейной обработкой сигнала. Термин "наложение сигнала" (heterodyning) обозначает *преобразование* частоты или процесс *смешивания*, вызывающий смещение спектра сигнала. Как следствие теоремы эквивалентности, любая линейная модель обработки сигналов может использоваться на узкополосных сигналах (что предпочтительнее с точки зрения простоты) с теми же результатами, что и на полосовых сигналах. Это означает, что производительность большинства цифровых систем связи часто можно описать и проанализировать, считая канал передачи узкополосным.

3.1. Сигналы и шум

3.1.1. Рост вероятности ошибки в системах связи

Задача детектора — максимально безошибочно угадать принятый сигнал, насколько это возможно при данном ухудшении качества сигнала в процессе передачи. Существует две причины роста вероятности ошибки. Первая — это последствия фильтрации в передатчике, канале и приемнике, рассмотренные в разделе 3.3. В этом разделе показано, что неидеальная передаточная функция системы приводит к "размыванию" символов, или *межсимвольной интерференции* (intersymbol interference — ISI).

Вторая причина роста вероятности ошибки — электрические помехи, порождаемые различными источниками, такими как галактика и атмосфера, импульсные помехи, комбинационные помехи, а также интерференция с сигналами от других источников. (Этот вопрос подробно рассмотрен в главе 5.) При надлежащих мерах предосторожности можно устранить большую часть помех и уменьшить последствия интерференции.

В то же время существуют помехи, устранить которые нельзя; это — помехи, вызываемые тепловым движением электронов в любой проводящей среде. Это движение порождает в усилителях и каналах связи *тепловой шум*, который аддитивно накладывается на сигнал. Использование квантовой механики позволило разработать хорошо известную статистику теплового шума [1].

Основная статистическая характеристика теплового шума заключается в том, что его амплитуды распределены по нормальному или гауссову закону распределения, рассмотренному в разделе 1.5.5 (рис. 1.7). На этом рисунке показано, что наиболее вероятные амплитуды шума — амплитуды с небольшими положительными или отрицательными значениями. Теоретически шум может быть бесконечно большим, но на практике очень большие амплитуды шума крайне редки. Основная спектральная характеристика теплового шума в системе связи заключается в том, что его двусторонняя спектральная плотность мощности $G_n(f) = N_0/2$ является плоской для всех частот, представляющих практический интерес. Другими словами, в тепловом шуме в среднем на низкочастотные флуктуации приходится столько же мощности на герц, сколько и на высокочастотные флуктуации — вплоть до частоты порядка 10^{12} герц. Если мощность шума характеризуется постоянной спектральной плотностью мощности, шум называется *белым*. Поскольку тепловой шум присутствует во всех системах связи и для многих систем является доминирующим источником помех, характеристики теплового шума[1] часто используются для моделирования шума при обнаружении и проектировании приемников. Всякий раз, когда канал связи определен как канал AWGN (при отсутствии указаний на другие параметры, ухудшающие качество передачи), мы, по сути, говорим, что ухудшение качества сигнала связано исключительно с неустранимым тепловым шумом.

3.1.2. Демодуляция и обнаружение

В течение данного интервала передачи сигнала, T, бинарная узкополосная система передает один из двух возможных сигналов, обозначаемых как $g_1(t)$ и $g_2(t)$. Подобным образом бинарная полосовая система передает один из двух возможных сигналов, обозначаемых как $s_1(t)$ и $s_2(t)$. Поскольку общая трактовка демодуляции и обнаружения, по сути, совпадает для узкополосных и полосовых систем, будем использовать запись $s_i(t)$ для обозначения передаваемого сигнала, вне зависимости от того, является система узкополосной или полосовой. Это позволяет совместить многие аспекты демодуляции/обнаружения в узкополосных системах, рассмотренные в данной главе, с соответствующими описаниями для полосовых систем, рассмотренных в главе 4. Итак, для любого канала двоичный сигнал, переданный в течение интервала $(0, T)$, представляется следующим образом.

$$s_i(t) = \begin{cases} s_1(t) & 0 \le t \le T \quad \text{для символа 1} \\ s_2(t) & 0 \le t \le T \quad \text{для символа 0} \end{cases}$$

Принятый сигнал $r(t)$ искажается вследствие воздействия шума $n(t)$ и, возможно, неидеальной импульсной характеристики канала $h_c(t)$ и описывается следующей формулой (1.1).

---

[1]Эти характеристики (аддитивный, белый, гауссов) определили принятое название шума — AWGN (additive white Gaussian noise).

$$r(t) = s_i(t) * h_c(t) + n(t) \tag{3.1}$$

В нашем случае $n(t)$ предполагается процессом AWGN с нулевым средним, а знак "*" обозначает операцию свертки. Для бинарной передачи по идеальному, свободному от искажений каналу, где свертка с функцией $h_c(t)$ не ухудшает качество сигнала (поскольку для идеального случая $h_c(t)$ — импульсная функция), вид $r(t)$ можно упростить.

$$r(t) = s_i(t) + n(t) \qquad i = 1, 2, \qquad 0 \le t \le T \tag{3.2}$$

Типичные функции демодуляции и обнаружения цифрового приемника показаны на рис. 3.1. Некоторые авторы используют термины "демодуляция" и "обнаружение" как синонимы. В данной книге они имеют различные значения. *Демодуляцию* (demodulation) мы определим как восстановление сигнала (в неискаженный узкополосный импульс), а *обнаружение* (detection) — как процесс принятия решения относительно цифрового значения этого сигнала. При *отсутствии* кодов коррекции ошибок на выход детектора поступают аппроксимации символов (или битов) сообщений m_i' (также называемые *жестким решением*). При использовании кодов коррекции ошибок на выход детектора поступают аппроксимации канальных символов (или кодированных битов) u_i', имеющие вид *жесткого* или *мягкого решения* (см. раздел 7.3.2). Для краткости термин "обнаружение" иногда применяется для обозначения совокупности всех этапов обработки сигнала, выполняемых в приемнике, вплоть до этапа принятия решения. Блок *преобразования с понижением частоты*, показанный на рис. 3.1 в разделе демодуляции, отвечает за трансляцию полосовых сигналов, работающих на определенных радиочастотах. Эта функция может реализовываться различными способами. Она может выполняться на входе приемника, в демодуляторе, распределяться между этими двумя устройствами или вообще не реализовываться.

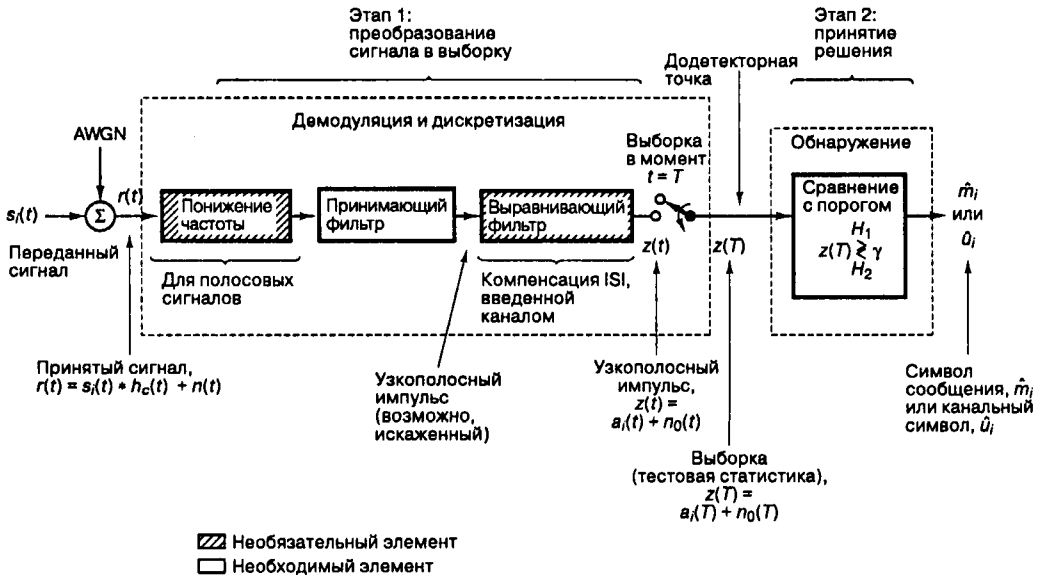


Рис. 3.1. Два основных этапа в процессе демодуляции/обнаружения цифровых сигналов

В блоке *демодуляции и дискретизации* (рис. 3.1) изображен *принимающий фильтр* (по сути, демодулятор), выполняющий восстановление сигнала в качестве подготовки к следующему необходимому этапу — обнаружению. Фильтрация в передатчике и кана-

ле обычно приводит к искажению принятой последовательности импульсов, вызванному межсимвольной интерференцией, а значит, эти импульсы не совсем готовы к дискретизации и обнаружению. Задачей принимающего фильтра является восстановление узкополосного импульса с максимально возможным отношением сигнал/шум (signal-to-noise ratio — SNR) и без межсимвольной интерференции. Оптимальный принимающий фильтр, выполняющий такую задачу, называется *согласованным* (matched), или *коррелятором* (correlator) и описывается в разделах 3.2.2 и 3.2.3. За принимающим фильтром может находиться *выравнивающий фильтр* (equalizing filter), или эквалайзер (equalizer); он необходим только в тех системах, в которых сигнал может искажаться вследствие межсимвольной интерференции, введенной каналом. Принимающий и выравнивающий фильтры показаны как два отдельных блока, что подчеркивает различие их функций. Впрочем, в большинстве случаев при использовании эквалайзера для выполнения обеих функций (а следовательно, и для компенсации искажения, внесенного передатчиком и каналом) может разрабатываться единый фильтр. Такой составной фильтр иногда называется просто *выравнивающим* или *принимающим и выравнивающим.*

На рис. 3.1 выделены два этапа процесса демодуляции/обнаружения. Этап 1, преобразование сигнала в выборку, выполняется демодулятором и следующим за ним устройством дискретизации. В конце каждого интервала передачи символа T на выход устройства дискретизации, *додетекторную точку*, поступает выборка $z(T)$, иногда называемая тестовой статистикой. Значение напряжения выборки $z(T)$ прямо пропорционально энергии принятого символа и обратно пропорционально шуму. На этапе 2 принимается решение относительно цифрового значения выборки (выполняется обнаружение). Предполагается, что шум является случайным гауссовым процессом, а принимающий фильтр демодулятора — линейным. Линейная операция со случайным гауссовым процессом дает другой случайный гауссов процесс [2]. Следовательно, на выходе фильтра шум также является гауссовым. Значит, выход этапа 1 можно описать выражением

$$z(T) = a_i(T) + n_0(T) \quad i = 1, 2, \tag{3.3}$$

где $a_i(T)$ — желаемый компонент сигнала, а $n_0(T)$ — шум. Для упрощения записи выражение (3.3) будем иногда представлять в виде $z = a_i + n_0$. Шумовой компонент n_0 — это случайная гауссова переменная с нулевым средним, поэтому $z(T)$ — случайная гауссова переменная со средним a_1 или a_2, в зависимости от того, передавался двоичный нуль или двоичная единица. Как описывалось в разделе 1.5.5, плотность вероятности случайного гауссового шума n_0 можно выразить как

$$p(n_0) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{n_0}{\sigma_0}\right)^2\right], \tag{3.4}$$

где σ_0^2 — дисперсия шума. Используя выражения (3.3) и (3.4), можно выразить плотности условных вероятностей $p(z|s_1)$ и $p(z|s_2)$.

$$p(z|s_1) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - a_1}{\sigma_0}\right)^2\right] \tag{3.5}$$

и

$$p(z|s_2) = \frac{1}{\sigma_0\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{z-a_2}{\sigma_0}\right)^2\right] \tag{3.6}$$

Эти плотности условных вероятностей показаны на рис. 3.2. Плотность $p(z|s_1)$, изображенная справа, называется *правдоподобием* s_1 и показывает плотность вероятности случайной переменной $z(T)$ при условии передачи символа s_1. Подобным образом функция $p(z|s_2)$ (слева) является *правдоподобием* s_2 и показывает плотность вероятности $z(T)$ при условии передачи символа s_2. Ось абсцисс, $z(T)$, представляет полный диапазон возможных значений выборки, взятой в течение этапа 1, изображенного на рис. 3.1.

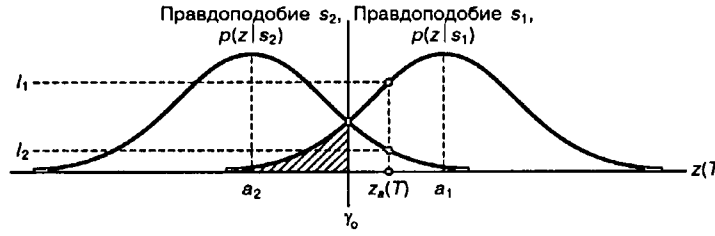


Рис. 3.2. Плотности условных вероятностей: $p(z|s_1)$ и $p(z|s_2)$

После того как принятый сигнал преобразован в выборку, действительная форма сигнала уже не имеет значения; сигналы всех типов, преобразованные в одинаковое значение $z(T)$, идентичны для схемы обнаружения. Далее будет показано, что оптимальный принимающий фильтр (согласованный фильтр) на этапе 1 (рис. 3.1) отображает все сигналы с равными энергиями в одну и ту же точку $z(T)$. Следовательно, важным параметром процесса обнаружения является *энергия* (а не форма) принятого сигнала, именно поэтому анализ обнаружения для узкополосных сигналов не отличается от анализа для полосовых сигналов. Поскольку $z(T)$ является сигналом напряжения, пропорциональным энергии принятого символа, то чем больше амплитуда $z(T)$, тем более достоверным будет процесс принятия решения относительно цифрового значения сигнала. На этапе 2 обнаружение выполняется посредством выбора гипотезы, являющейся следствием порогового измерения

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} \gamma, \tag{3.7}$$

где H_1 и H_2 — две возможные (бинарные) гипотезы. Приведенная запись указывает, что гипотеза H_1 выбирается при $z(T) > \gamma$, а H_2 — при $z(T) < \gamma$. Если $z(T) = \gamma$, решение может быть любым. Выбор H_1 равносилен тому, что передан был сигнал $s_1(t)$, а значит, результатом обнаружения является двоичная единица. Подобным образом выбор H_2 равносилен передаче сигнала $s_2(t)$, а значит, результатом обнаружения является двоичный нуль.

3.1.3. Векторное представление сигналов и шума

Рассмотрим геометрическое или векторное представление, приемлемое как для узкополосных, так и полосовых сигналов. Определим N-мерное *ортогональное пространство* как пространство, определяемое набором N линейно независимых функций $\{\phi_j(t)\}$,

именуемых *базисными*. Любая функция этого пространства может выражаться через линейную комбинацию базисных функций, которые должны удовлетворять условию

$$\int_0^T \psi_j(t)\psi_k(t)\,dt = K_j\delta_{jk} \quad 0 \le t \le T \qquad j, k = 1, \ldots, N, \tag{3.8,а}$$

где оператор

$$\delta_{jk} = \begin{cases} 1 & \text{для } j = k \\ 0 & \text{для } j \ne k \end{cases} \tag{3.8,б}$$

называется дельта-функцией Кронекера и определяется формулой (3.8,б). При ненулевых константах K_j пространство именуется *ортогональным*. Если базисные функции нормированы так, что все $K_j = 1$, пространство называется *ортонормированным*. Основное условие ортогональности можно сформулировать следующим образом: каждая функция $\psi_j(t)$ набора базисных функций должна быть независимой от остальных функций набора. Каждая функция $\psi_j(t)$ не должна интерферировать с другими функциями в процессе обнаружения. С геометрической точки зрения все функции $\psi_j(t)$ взаимно перпендикулярны. Пример подобного пространства с $N = 3$ показан на рис. 3.3, где взаимно перпендикулярные оси обозначены $\psi_1(t)$, $\psi_2(t)$ и $\psi_3(t)$. Если $\psi_j(t)$ соответствует действительному компоненту напряжения или тока сигнала, нормированному на сопротивление 1 Ом, то, используя формулы (1.5) и (3.8), получаем следующее выражение для нормированной энергии в джоулях, переносимой сигналом $\psi_j(t)$ за T секунд.

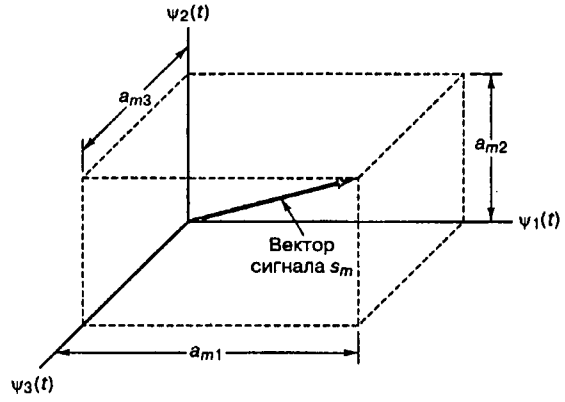


Рис. 3.3. Векторное представление сигнала $s_m(t)$

$$E_j = \int_0^T \psi_j^2(t)\,dt = K_j \tag{3.9}$$

Одной из причин нашего внимания к *ортогональному сигнальному пространству* является то, что в нем проще всего определяется Евклидова мера расстояния, используемая в процессе обнаружения. Стоит отметить, что даже если волны, переносящие сигналы, не формируют подобного пространства, они могут преобразовываться в ли-

нейную комбинацию ортогональных сигналов. Можно показать [3], что *произвольный* конечный набор сигналов $\{s_i(t)\}$ $(t = 1, ..., M)$, где каждый элемент множества физически реализуем и имеет длительность T, можно выразить как линейную комбинацию N ортогональных сигналов $\psi_1(t)$, $\psi_2(t)$, ..., $\psi_N(t)$, где $N \le M$, так, что

$$
\begin{aligned}
s_1(t) &= a_{11}\psi_1(t) + a_{12}\psi_2(t) + \ldots + a_{1N}\psi_N(t) \\
s_2(t) &= a_{21}\psi_1(t) + a_{22}\psi_2(t) + \ldots + a_{2N}\psi_N(t) \\
&\ldots \\
s_M(t) &= a_{M1}\psi_1(t) + a_{M2}\psi_2(t) + \ldots + a_{MN}\psi_N(t)
\end{aligned}
$$

Эти соотношения можно записать в более компактной форме.

$$
s_i(t) = \sum_{j=1}^{N} a_{ij}\psi_j(t) \quad i = 1, ..., M \tag{3.10}
$$

$$
N \le M,
$$

где

$$
a_{ij} = \frac{1}{K_j}\int_0^T s_i(t)\psi_j(t)\,dt \quad i = 1, ..., M \quad 0 \le t \le T \tag{3.11}
$$

$$
j = 1, ..., N
$$

a_{ij} — это коэффициент при $\psi_j(t)$ разложения сигнала $s_i(t)$ по базисным функциям. Вид базиса $\{\psi_j(t)\}$ не задается; эти сигналы выбираются с точки зрения удобства и зависят от формы волн передачи сигналов. Набор таких волн $\{s_i(t)\}$ можно рассматривать как набор векторов $\{\mathbf{s}_i\} = \{a_{i1}, a_{i2}, ..., a_{iN}\}$. Если, например, $N = 3$, то сигналу

$$
s_m(t) = a_{m1}\psi_1(t) + a_{m2}\psi_2(t) + a_{m3}\psi_3(t),
$$

соответствует вектор $\mathbf{s}_m$, который можно изобразить как точку в трехмерном Евклидовом пространстве с координатами (a_{m1}, a_{m2}, a_{m3}), как показано на рис. 3.3. Взаимная ориентация векторов сигналов описывает связь между сигналами (относительно их фаз или частот), а амплитуда каждого вектора набора $\{s_i\}$ является мерой энергии сигнала, перенесенной в течение времени передачи символа. Вообще, после выбора набора из N ортогональных функций, каждый из переданных сигналов $s_i(t)$ полностью определяется вектором его коэффициентов.

$$
\mathbf{s}_i = (a_{i1}, a_{i2}, ..., a_{iN}) \quad i = 1, ..., M \tag{3.12}
$$

В дальнейшем для отображения сигналов в векторной форме будем использовать запись $\{\mathbf{s}\}$ или $\{s(t)\}$. На рис. 3.4 в векторной форме (которая в данном случае является очень удобной) показан процесс обнаружения. Векторы $\mathbf{s}_j$ и $\mathbf{s}_k$ представляют *сигналы-прототипы,* или *опорные сигналы,* принадлежащие набору из M сигналов, $\{s_i(t)\}$. Приемник априори знает местонахождение в пространстве сигналов всех векторов-прототипов, принадлежащих M-мерному множеству. В процессе передачи каждый сигнал подвергается воздействию шумов, так что в действительности принимается искаженная версия исходного сигнала (например, $\mathbf{s}_j + \mathbf{n}$ или $\mathbf{s}_k + \mathbf{n}$), где $\mathbf{n}$ — вектор помех. Будем считать, что помехи являются аддитивными и имеют гауссово распределение; следовательно, результирующее распределение возможных принимаемых сигна-

лов — это кластер или облако точек вокруг $\mathbf{s}_j$ и $\mathbf{s}_k$. Кластер сгущается к центру и разрежается с увеличением расстояния от прототипа. Стрелочка с пометкой "**r**" представляет вектор сигнала, который поступает в приемник в течение определенного интервала передачи символа. Задача приемника — определить, на какой из прототипов M-мерного множества сигнал "похож" больше. Мерой "сходства" может быть *расстояние.* Приемник или детектор должен решить, какой из прототипов сигнального пространства *ближе* к принятому вектору **r**. Анализ всех схем демодуляции или обнаружения включает использование понятия *расстояние* между принятым сигналом и набором возможных переданных сигналов. Детектор должен следовать одному простому правилу: определять **r** к тому же классу, к которому принадлежит его ближайший сосед (ближайший вектор-прототип).

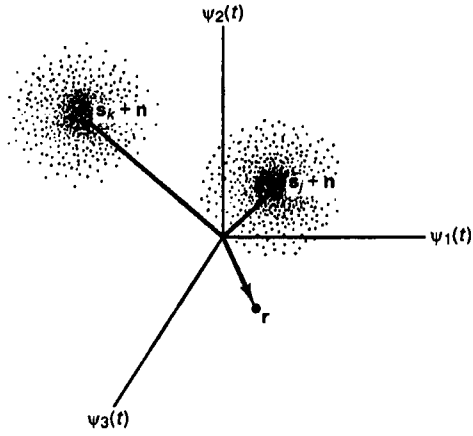


Рис. 3.4. Сигналы и шум в трехмерном векторном пространстве

3.1.3.1. Энергия сигнала

С помощью формул (1.5), (3.10) и (3.8) нормированную энергию E_i, связанную с сигналом $s_i(t)$ в течение периода передачи символа T, можно выразить через ортогональные компоненты $s_i(t)$.

$$E_i = \int_0^T s_i^2(t)\,dt = \int_0^T \left[\sum_j a_{ij}\psi_j(t)\right]^2 dt = \tag{3.13}$$

$$= \int_0^T \sum_j a_{ij}\psi_j(t) \sum_k a_{ik}\psi_k(t)\,dt = \tag{3.14}$$

$$= \sum_j \sum_k a_{ij}a_{ik} \int_0^T \psi_j(t)\psi_k(t)\,dt = \tag{3.15}$$

$$= \sum_j \sum_k a_{ij}a_{ik}K_j\delta_{jk} = \tag{3.16}$$

$$= \sum_{j=1}^{N} a_{ij}^2 K_j \quad i = 1, \ldots, M \tag{3.17}$$

Уравнение (3.17) — это частный случай теоремы Парсеваля, связывающей интеграл от квадрата сигнала $s_i(t)$ с суммой квадратов коэффициентов ортогонального разложения $s_i(t)$. При использовании ортонормированных функций (т.е. при $K_j = 1$) нормированная энергия за промежуток времени T дается следующим выражением.

$$E_i = \sum_{j=1}^{N} a_{ij}^2 \tag{3.18}$$

Если все сигналы $s_i(t)$ имеют одинаковую энергию, формулу (3.18) можно записать следующим образом.

$$E = \sum_{j=1}^{N} a_{ij}^2 \text{ для всех } i \tag{3.19}$$

3.1.3.2. Обобщенное преобразование Фурье

Преобразование, описанное формулами (3.8), (3.10) и (3.11), называется *обобщенным преобразованием Фурье*. При обычном преобразовании Фурье множество $\{\psi_j(t)\}$ включает синусоиды и косинусоиды, а в случае обобщенного преобразования оно не ограничено какой-либо конкретной формой; это множество должно лишь удовлетворять условию ортогональности, записанному в форме уравнения (3.8). Обобщенное преобразование Фурье позволяет представить любой произвольный интегрируемый набор сигналов (или шумов) в виде линейной комбинации ортогональных сигналов [3]. Следовательно, в подобном ортогональном пространстве в качестве критерия принятия решения для обнаружения *любого* набора сигналов при шуме AWGN вполне оправдано использование расстояния (Евклидового расстояния). Вообще, важнейшее применение этого ортогонального преобразования связано с действительной передачей и приемом сигналов. Передача неортогонального набора сигналов в общем случае осуществляется посредством подходящего взвешивания ортогональных компонентов несущих.

Пример 3.1. Ортогональное представление сигналов

На рис. 3.5 иллюстрируется утверждение, что любой произвольный интегрируемый набор сигналов может представляться как линейная комбинация ортогональных сигналов. На рис. 3.5, *а* показан набор из трех сигналов, $s_1(t)$, $s_2(t)$ и $s_3(t)$.

а) Покажите, что данные сигналы *не* взаимно ортогональны.

б) На рис. 3.5, *б* показаны два сигнала $\psi_1(t)$ и $\psi_2(t)$. Докажите, что эти сигналы ортогональны.

в) Покажите, как неортогональные сигналы из п. а можно выразить как линейную комбинацию ортогональных сигналов из п. б.

г) На рис. 3.5, *в* показаны другие два сигнала $\psi_1'(t)$ и $\psi_2'(t)$. Покажите, как неортогональные сигналы, показанные на рис. 3.5, *а*, выражаются через линейную комбинацию сигналов, изображенных на рис. 3.5, *в*.

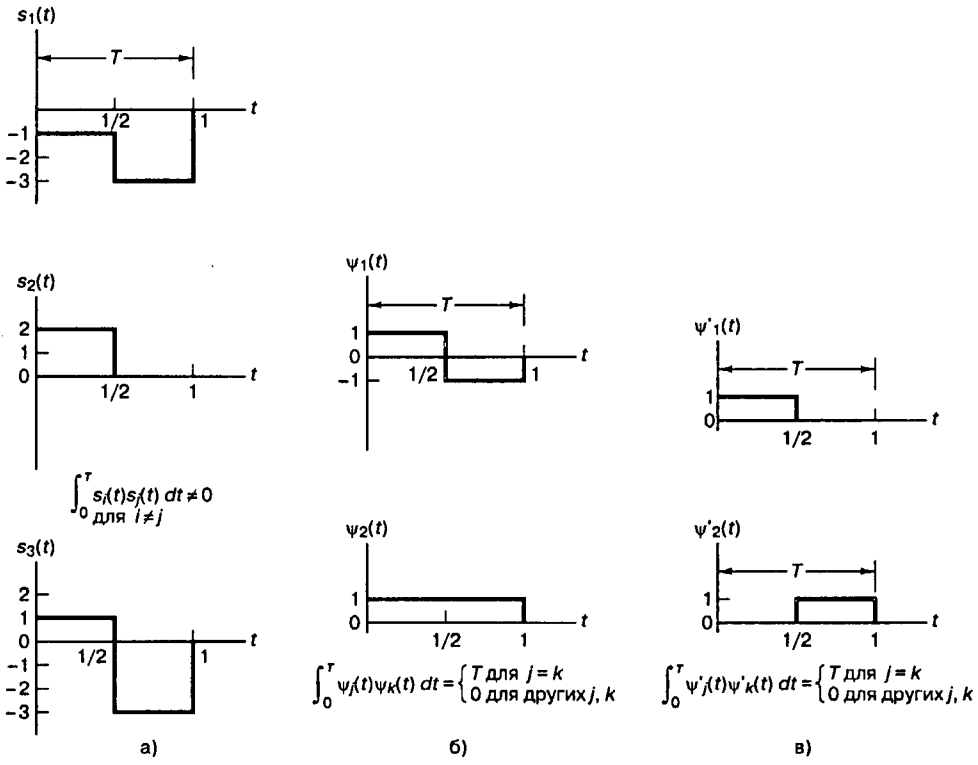


Рис. 3.5. Пример выражения произвольного набора сигналов через ортогональный набор: а) произвольный набор сигналов; б) набор ортогональных базисных функций; в) другой набор ортогональных базисных функций

Решение

а) Сигналы $s_1(t)$, $s_2(t)$ и $s_1(t)$, очевидно, не являются взаимно ортогональными, поскольку не удовлетворяют требованиям, указанным в формуле (3.8), т.е. интегрирование по времени (по периоду передачи символа) скалярного произведения любых двух из трех сигналов не равно нулю. Покажем это для сигналов $s_1(t)$ и $s_2(t)$.

$$\int_0^T s_1(t)s_2(t)dt = \int_0^{T/2} s_1(t)s_2(t)dt + \int_{T/2}^T s_1(t)s_2(t)dt =$$

$$= \int_0^{T/2}(-1)(2)dt = \int_{T/2}^T(-3)(0)\,dt = -T$$

Подобным образом интегрирование по интервалу времени T каждого из скалярных произведений $s_1(t)s_3(t)$ и $s_2(t)s_3(t)$ дает ненулевой результат. Следовательно, множество сигналов $\{s_i(t)\}$ $(i = 1, 2, 3)$ на рис. 3.5, *а* не является ортогональным

б) Используя формулу (3.8), докажем, что $\psi_1(t)$ и $\psi_2(t)$ ортогональны.

$$\int_0^T \psi_1(t)\psi_2(t)dt = \int_0^{T/2}(1)(1)dt + \int_{T/2}^{T}(-1)(1)dt = 0$$

в) С использованием формулы (3.11) при $K_i = T$, неортогональное множество сигналов $\{s_i(t)\}$ $(i = 1, 2, 3)$ можно выразить через линейную комбинацию ортогональных базисных сигналов $\{\psi_j(t)\}$ $(j = 1, 2)$.

$$s_1(t) = \psi_1(t) - 2\psi_2(t)$$
$$s_2(t) = \psi_1(t) + \psi_2(t)$$
$$s_3(t) = 2\psi_1(t) - \psi_2(t)$$

г) Подобно тому, как было сделано в п. в, неортогональное множество $\{s_i(t)\}$ $(i = 1, 2, 3)$ можно выразить через ортогональный набор базисных функций $\{\psi_j'(t)\}$ $(j = 1, 2)$, изображенный на рис. 3.5, *в*.

$$s_1(t) = \psi_1'(t) - 3\psi_2'(t)$$
$$s_2(t) = 2\psi_1'(t)$$
$$s_3(t) = \psi_1'(t) - 3\psi_2'(t)$$

Эти соотношения показывают, как произвольный набор сигналов $\{s_i(t)\}$ выражается через линейную комбинацию сигналов ортогонального набора $\{\psi_j(t)\}$, что описывается формулами (3.10) и (3.11). Какое практическое значение имеет возможность представления сигналов $s_1(t)$, $s_2(t)$ и $s_3(t)$ через сигналы $\psi_1(t)$, $\psi_2(t)$ и соответствующие коэффициенты? Если мы хотим, чтобы система передавала сигналы $s_1(t)$, $s_2(t)$ и $s_3(t)$, достаточно, чтобы передатчик и приемник реализовывались только с использованием двух базисных функций $\psi_1(t)$ и $\psi_2(t)$, а не трех исходных сигналов. Получить ортогональный набор базисных функций $\{\psi_1(t)\}$ из любого данного набора сигналов $\{s_i(t)\}$ позволяет *процесс ортогонализации Грамма-Шмидта.* (Подробно этот процесс описан в приложении 4А работы [4].)

3.1.3.3. Представление белого шума через ортогональные сигналы

Аддитивный белый гауссов шум (additive white Gaussian noise — AWGN), как и любой другой сигнал, можно выразить как линейную комбинацию ортогональных сигналов. Для последующего рассмотрения процесса обнаружения сигналов шум удобно разделить на два компонента:

$$n(t) = \hat{n}(t) + \tilde{n}(t), \tag{3.20}$$

где

$$\hat{n}(t) = \sum_{j=1}^{N} n_j \psi_j(t) \tag{3.21}$$

является шумом в пространстве сигналов или проекцией компонентов шума на координаты сигнала $\psi_1(t)$, ..., $\psi_N(t)$, а

$$\tilde{n}(t) = n(t) - \hat{n}(t) \tag{3.22}$$

есть шумом вне пространства сигналов. Другими словами, $\tilde{n}(t)$ можно рассматривать как шум, эффективно отсеиваемый детектором, а $\hat{n}(t)$ — как шум, который будет "вмешиваться" в процесс обнаружения. Итак, сигнал шума $n(t)$ можно выразить следующим образом.

$$n(t) = \sum_{j=1}^{N} n_j \psi_j(t) + \tilde{n}(t), \tag{3.23}$$

где

$$n_j = \frac{1}{K_j} \int_0^T n(t)\psi_j(t)\, dt \quad \text{для всех } j \tag{3.24}$$

и

$$\int_0^T \tilde{n}(t)\psi_j(t)\, dt = 0 \quad \text{для всех } j \tag{3.25}$$

Компонент $\hat{n}(t)$ шума, выраженный в формуле (3.21), следовательно, можно считать просто равным $n(t)$. Выразить шум $n(t)$ можно через вектор его коэффициентов, подобно тому, как это делалось для сигналов в формуле (3.12). Имеем

$$\mathbf{n} = (n_1, n_2, \ldots, n_N), \tag{3.26}$$

где $\mathbf{n}$ — случайный вектор с нулевым средним и гауссовым распределением, а компоненты шума n_i ($i = 1, \ldots, N$) являются независимыми.

3.1.3.4. Дисперсия белого шума

Белый шум — это *идеализированный процесс* с двусторонней спектральной плотностью мощности, равной постоянной величине $N_0/2$ для всех частот от $-\infty$ до $+\infty$. Следовательно, дисперсия шума (средняя мощность шума, поскольку шум имеет нулевое среднее) равна следующему.

$$\sigma^2 = \text{var}[n(t)] = \int_{-\infty}^{\infty} \left(\frac{N_0}{2}\right) df = \infty \tag{3.27}$$

Хотя дисперсия AWGN равна бесконечности, дисперсия фильтрованного шума AWGN конечна. Например, если AWGN коррелирует с одной из набора ортонормированных функций $\psi_j(t)$, дисперсия на выходе коррелятора описывается следующим выражением.

$$\sigma^2 = \text{var}\, n_j = \mathbf{E}\left\{\left[\int_0^T n(t)\psi_j(t)\, dt\right]^2\right\} = \frac{N_0}{2} \tag{3.28}$$

Доказательство формулы (3.28) приводится в приложении В. С этого момента будем считать, что интересующий нас шум процесса обнаружения является шумом на выходе коррелятора или согласованного фильтра с дисперсией $\sigma_2 = N_0/2$, как указано в формуле (3.28).

3.1.4. Важнейший параметр систем цифровой связи — отношение сигнал/шум

Любой, кто изучал аналоговую связь, знаком с критерием качества, именуемым *отношением средней мощности сигнала к средней мощности шума* (S/N или SNR). В цифровой связи в качестве критерия качества чаще используется нормированная версия SNR, E_b/N_0. E_b — это энергия бита, и ее можно описать как мощность сигнала S, умноженную на время передачи бита T_b. N_0 — это спектральная плотность мощности шума, и ее можно выразить как мощность шума N, деленную на ширину полосы W. Поскольку время передачи бита и скорость передачи битов R_b взаимно обратны, T_b можно заменить на $1/R_b$.

$$\frac{E_b}{N_0} = \frac{S\,T_b}{N/W} = \frac{S/R_b}{N/W} \tag{3.29}$$

Еще одним параметром, часто используемым в цифровой связи, является скорость передачи данных в битах в секунду. В целях упрощения выражений, встречающихся в книге, для представления скорости передачи битов вместо записи R_b будем писать просто R. С учетом сказанного перепишем, выражение (3.29) так, чтобы было явно видно, что отношение E_b/N_0 представляет собой отношение S/N, нормированное на ширину полосы и скорость передачи битов.

$$\frac{E_b}{N_0} = \frac{S}{N}\left(\frac{W}{R}\right) \tag{3.30}$$

Одной из важнейших метрик производительности в системах цифровой связи является график зависимости вероятности появления ошибочного бита P_B от E_b/N_0. На рис. 3.6 показан "водопадоподобный" вид большинства подобных кривых. При $E_b/N_0 \geq x_0$, $P_B \leq P_0$. Безразмерное отношение E_b/N_0 — это стандартная качественная мера производительности систем цифровой связи. Следовательно, необходимое отношение E_b/N_0 можно рассматривать как метрику, позволяющую сравнивать производительность различных систем; чем меньше требуемое отношение E_b/N_0, тем эффективнее процесс обнаружения при данной вероятности ошибки.

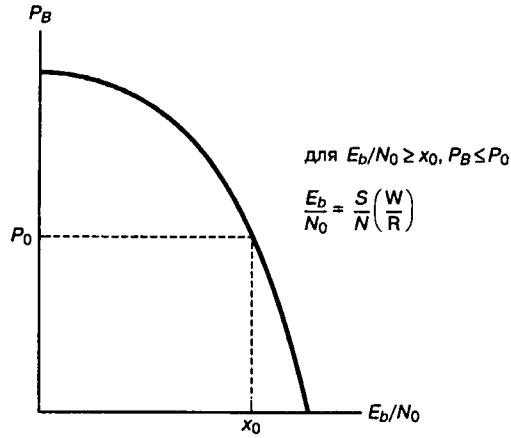


Рис. 3.6. Общий вид зависимости P_B от E_b/N_0

3.1.5. Почему отношение E_b/N_0 — это естественный критерий качества

У неспециалистов в области цифровой связи может возникнуть вопрос о полезности параметра E_b/N_0. Отношение S/N — это удобный критерий качества для аналоговых систем связи: числитель представляет меру мощности сигнала, которую желательно сохранить, а знаменатель — ухудшение вследствие электрических помех. Более того, отношение S/N интуитивно воспринимается как мера качества. Итак, почему в цифровых системах связи мы не можем продолжать использовать отношение S/N как критерий качества? Зачем для цифровых систем нужна другая метрика — отношение энергии бита к спектральной плотности мощности шума? Объяснению этого вопроса и посвящен данный раздел.

В разделе 1.2.4 сигнал в представлении через мощность определялся как сигнал с конечной средней мощностью и бесконечной энергией. Энергетический сигнал определялся как сигнал с нулевой средней мощностью и конечной энергией. Подобная классификация полезна при сравнении аналоговых и цифровых сигналов. Аналоговый сигнал мы относим к сигналам, представляемым через мощность. Почему это имеет смысл? Об аналоговом сигнале можно думать как о сигнале, имеющем бесконечную длительность, который не требуется разграничивать во времени. Неограниченно длительный аналоговый сигнал содержит бесконечную энергию; следовательно, использование энергии — это не самый удобный способ описания характеристик такого сигнала. Значительно более удобным параметром для аналоговых волн является мощность (или скорость доставки энергии).

В то же время в системах цифровой связи мы передаем (и принимаем) символы путем передачи некоторого сигнала в течение конечного промежутка времени, времени передачи символа T_s. Сконцентрировав внимание на одном символе, видим, что мощность (усредненная по времени) стремится к нулю. Значит, для описания характеристик цифрового сигнала мощность не подходит. Для подобного сигнала нам нужна метрика, "достаточно хорошая" в пределах конечного промежутка времени. Другими словами, энергия символа (мощность, проинтегрированная по T_s) — это гораздо более удобный параметр описания цифровых сигналов.

То, что цифровой сигнал лучше всего характеризует полученная им энергия, еще не дает ответа на вопрос, почему E_b/N_0 — это естественная метрика для цифровых систем, так что продолжим. Цифровой сигнал — это транспортное средство, представляющее цифровое сообщение. Сообщение может содержать один бит (двоичное сообщение), два (четверичное), ..., 10 бит (1024-ричное). В аналоговых системах нет ничего подобного такой дискретной структуре сообщения. Аналоговый информационный источник — это бесконечно квантованная непрерывная волна. Для цифровых систем критерий качества должен позволять сравнивать одну систему с другой на битовом уровне. Следовательно, описывать цифровые сигналы в терминах S/N практически бесполезно, поскольку сигнал может иметь однобитовое, 2-битовое или 10-битовое значение. Предположим, что для данной вероятности возникновения ошибки в цифровом двоичном сигнале требуемое отношение S/N равно 20. Будем считать, что понятия сигнала и его значения взаимозаменяемы. Поскольку двоичный сигнал имеет однобитовое значение, требуемое отношение S/N на бит равно 20 единицам. Предположим, что наш сигнал является 1024-ричным, с теми же 20 единицами требуемого отношения S/N. Теперь, поскольку сигнал имеет 10-битовое значение, требуемое отношение S/N на один бит равно всего 2. Возникает вопрос: почему мы должны выполнять такую цепочку вычислений, чтобы найти метрику, представляющую критерий качества? Почему бы сразу не выразить метрику через то, что нам

действительно надо, — параметр, связанный с энергией на битовом уровне, E_b/N_0? В заключение отметим, что поскольку отношение S/N является безразмерным, таким же есть и отношение E_b/N_0. Для проверки можно вычислить единицы измерения.

$$\frac{E_b}{N_0}=\frac{\text{Джоуль}}{\text{Ватт на герц}}=\frac{\text{Ватт–секунда}}{\text{Ватт–секунда}}$$

3.2. Обнаружение двоичных сигналов в гауссовом шуме

3.2.1. Критерий максимального правдоподобия приема сигналов

Критерий принятия решения, используемый в этапе 2 (рис. 3.1), описывался формулой (3.7) следующим образом.

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} \gamma$$

Популярный критерий выбора порога γ для принятия двоичного решения в выражении (3.7) основан на минимизации вероятности ошибки. Вычисление этого *минимального значения ошибки* $\gamma = \gamma_0$ начинается с записи связи отношения плотностей условных вероятностей и отношения априорных вероятностей появления сигнала. Поскольку плотность условной вероятности $p(z|s_i)$ также называется *правдоподобием* s_i, формулировка

$$\frac{p(z \mid s_1)}{p(z \mid s_2)} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{P(s_2)}{P(s_1)} \tag{3.31}$$

есть *критерием отношения правдоподобий* (см. приложение Б). В этом неравенстве $P(s_1)$ и $P(s_2)$ являются априорными вероятностями передачи сигналов $s_1(t)$ и $s_2(t)$, а H_1 и H_2 — две возможные гипотезы. Правило минимизации вероятности ошибки (формула (3.31)) гласит, что если отношение правдоподобий больше отношения априорных вероятностей, то следует выбирать гипотезу H_1.

В разделе Б.3.1 показано, что при $P(s_1) = P(s_2)$ и симметричных правдоподобиях $p(z \mid s_i)$ $(i = 1, 2)$ подстановка формул (3.5) и (3.6) в формулу (3.31) дает

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} \frac{a_1 + a_2}{2} = \gamma_0, \tag{3.32}$$

где a_1 — сигнальный компонент $z(T)$ при передаче $s_1(t)$, а a_2 — сигнальный компонент при передаче $s_2(t)$. Порог γ_0, представленный выражением $(a_1 + a_2)/2$, — это *оптимальный порог* для минимизации вероятности принятия неверного решения в этом важном частном случае. Описанный подход называется *критерием минимальной ошибки*.

Для равновероятных сигналов оптимальный порог γ_0, как показано на рис. 3.2, проходит через пересечение функций правдоподобия. Следовательно, используя формулу (3.32), видим, что этап принятия решения заключается в эффективном выборе

гипотезы, соответствующей сигналу с *максимальным правдоподобием*. Пусть, например, значение выборки принятого сигнала равно $z_a(T)$, а правдоподобия того, что $z_a(T)$ принадлежит к одному из двух классов $s_1(t)$ или $s_2(t)$, отличны от нуля. В этом случае критерий принятия решения можно рассматривать как сравнение правдоподобий $p(z_a|s_1)$ и $p(z_a|s_1)$. Более вероятное значение переданного сигнала соответствует наибольшей плотности вероятности. Другими словами, детектор выбирает $s_1(t)$, если

$$p(z_a|s_1) > p(z_a|s_2) \tag{3.33}$$

В противном случае детектор выбирает $s_2(t)$. Детектор, минимизирующий вероятность ошибки (для классов равновероятных сигналов), называется *детектором максимального правдоподобия*.

Из рис. 3.2 можно видеть, что выражение (3.32) — это "метод здравого смысла" принятия решения при наличии статистических знаний о классах. Имея на выходе детектора значение $z_a(T)$, видим (рис. 3.2), что $z_a(T)$ пересекается с графиком правдоподобия $s_1(t)$ в точке l_1 и с графиком правдоподобия $s_2(t)$ в точке l_2. Какое наиболее разумное решение должен принять детектор? В описанном случае наиболее здравым является выбор класса $s_1(t)$, имеющего большее правдоподобие. Если бы пример был M-мерным, а не бинарным, всего существовало бы M функций правдоподобия, представляющих M классов сигналов, к которым может принадлежать принятый сигнал. Решение по принципу максимального правдоподобия в этом случае представляло бы выбор класса, имеющего самое большое правдоподобие из M возможных. (Основы теории принятия решений даются в приложении Б.)

3.2.1.1. Вероятность ошибки

В процессе принятия бинарного решения, показанном на рис. 3.2, существует две возможности возникновения ошибки. Ошибка e появится при передаче $s_1(t)$, если вследствие шума канала уровень переданного сигнала $z(t)$ упадет ниже γ_0. Вероятность этого равна следующему.

$$P(e|s_1) = P(H_2|s_1) = \int_{-\infty}^{\gamma_0} p(z|s_1)\, dz \tag{3.34}$$

Эта возможность показана заштрихованной областью слева от γ_0 (рис. 3.2). Подобным образом ошибка появляется при передаче $s_2(t)$, если вследствие шума канала уровень переданного сигнала $z(t)$ поднимется выше γ_0. Вероятность этого равна следующему.

$$P(e|s_2) = P(H_1|s_2) = \int_{\gamma_0}^{\infty} p(z|s_2)\, dz \tag{3.35}$$

Суммарная вероятность ошибки равна сумме вероятностей всех возможностей ее появления. Для бинарного случая вероятность возникновения ошибочного бита можно выразить следующим образом.

$$P_B = \sum_{i=1}^{2} P(e, s_i) = \sum_{i=1}^{2} P(e|s_i) P(s_i) \tag{3.36}$$

Объединяя формулы (3.34)–(3.36), получаем

$$P_B = P(e|s_1)P(s_1) + P(e|s_2)P(s_2) \tag{3.37,а}$$

или, что равносильно,

$$P_B = P(H_2|s_1)P(s_1) + P(H_2|s_2)P(s_2) \tag{3.37,б}$$

Иными словами, при передаче сигнала $s_1(t)$ ошибка происходит при выборе гипотезы H_2; или при передаче сигнала $s_2(t)$ ошибка происходит при выборе гипотезы H_1. Для равных априорных вероятностей (т.е. $P(s_1) = P(s_2) = 1/2$) имеем следующее.

$$P_B = \tfrac{1}{2}P(H_2|s_1) + \tfrac{1}{2}P(H_1|s_2) \tag{3.38}$$

Используя симметричность плотностей вероятности, получаем следующее.

$$P_B = P(H_2|s_1) = P(H_1|s_2) \tag{3.39}$$

Вероятность появления ошибочного бита, P_B, численно равна площади под "хвостом" любой функции правдоподобия, $p(z|s_1)$ или $p(z|s_2)$, "заползающим" на "неправильную" сторону порога. Таким образом, для вычисления P_B мы можем проинтегрировать $p(z|s_1)$ от $-\infty$ до γ_0 или $p(z|s_2)$ — от γ_0 до ∞.

$$P_B = \int_{\gamma_0 = (a_1 + a_2)/2}^{\infty} p(z|s_2)\, dz \tag{3.40}$$

Здесь $\gamma_0 = (a_1 + a_2)/2$ — оптимальный порог из уравнения (3.32). Заменяя правдоподобие $p(z|s_2)$ его гауссовым эквивалентом из формулы (3.6), имеем

$$P_B = \int_{\gamma_0 = (a_1 + a_2)/2}^{\infty} \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - a_2}{\sigma_0}\right)^2\right] dz\,, \tag{3.41}$$

где σ_0^2 — дисперсия шума вне коррелятора.

Сделаем замену $u = (z - a_2)/\sigma_0$. Тогда $\sigma_0 du = dz$ и

$$P_B = \int_{u = (a_1 - a_2)/2\sigma_0}^{u = \infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du = Q\left(\frac{a_1 - a_2}{2\sigma_0}\right) \tag{3.42}$$

Здесь $Q(x)$ называется *гауссовым интегралом ошибок* и часто используется при описании вероятности с гауссовой плотностью распределения. Определяется эта функция следующим образом.

$$Q(x) \approx \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp\left(-\frac{u^2}{2}\right) du \tag{3.43}$$

Отметим, что гауссов интеграл ошибок может определяться несколькими способами (см. приложение Б); впрочем, все определения одинаково пригодны для описания вероятности ошибки при гауссовом шуме. $Q(x)$ нельзя вычислить в аналитическом виде. В табл. Б.1 она представлена в форме таблицы. Хорошие аппроксимации функции

$Q(x)$ через более простые функции можно найти в работе [5]. Вот одна из таких аппроксимаций, справедливая для $x > 3$.

$$Q(x) \approx \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{3.44}$$

Итак, мы оптимизировали (в смысле минимизации P_B) порог γ, но не оптимизировали принимающий фильтр в блоке 1 (рис. 3.1). Далее нашей целью является оптимизация этого фильтра путем максимизации аргумента $Q(x)$ в формуле (3.42).

3.2.2. Согласованный фильтр

Согласованный фильтр (matched filter) — это линейное устройство, спроектированное, чтобы давать на выходе максимально возможное для данного передаваемого сигнала отношение сигнал/шум. Предположим, что на вход линейного, инвариантного относительно времени (принимающего) фильтра, за которым следует устройство дискретизации (рис. 3.1), подается известный сигнал $s(t)$ плюс шум AWGN $n(t)$. В момент времени $t = T$ сигнал на выходе устройства дискретизации $z(T)$ состоит из компонента сигнала a_i и компонента шума n_0. Дисперсия шума на выходе (средняя мощность шума) записывается как σ_0^2, так что отношение мгновенной мощности шума к средней мощности шума, $(S/N)_T$, в момент $t = T$ вне устройства дискретизации на этапе 1 равно следующему.

$$\left(\frac{S}{N}\right)_T = \frac{a_i^2}{\sigma_0^2} \tag{3.45}$$

Нам нужно найти передаточную функцию фильтра $H_0(f)$ с *максимальным* отношением $(S/N)_T$. Сигнал $a_i(t)$ на выходе фильтра можно выразить через передаточную функцию фильтра $H_0(f)$ (до оптимизации) и Фурье-образ сигнала на входе

$$a_i(t) = \int_{-\infty}^{\infty} H(f)S(f)e^{2\pi ift}df\,, \tag{3.46}$$

где $S(f)$ — Фурье-образ сигнала на входе, $s(t)$. Если двусторонняя спектральная плотность мощности шума на входе равна $N_0/2$ Вт/Гц, то с помощью формул (1.19) и (1.53) мощность шума на выходе можно записать следующим образом.

$$\sigma_0^2 = \frac{N_0}{2}\int_{-\infty}^{\infty} |H(f)|^2 df \tag{3.47}$$

Объединяя формулы (3.45) и (3.47), получаем выражение для $(S/N)_T$.

$$\left(\frac{S}{N}\right)_T = \frac{\left|\int_{-\infty}^{\infty} H(f)S(f)e^{2\pi ift}df\right|^2}{N_0/2\int_{-\infty}^{\infty} |H(f)|^2\, df} \tag{3.48}$$

Найдем теперь значение $H(f) = H_0(f)$, при котором $(S/N)_T$ достигает максимума. Для этого нам понадобится *неравенство Шварца*, одна из форм записи которого представлена ниже.

$$\left|\int_{-\infty}^{\infty} f_1(x) f_2(x)\, dx\right|^2 \le \int_{-\infty}^{\infty} |f_1(x)|^2\, dx \int_{-\infty}^{\infty} |f_2(x)|^2\, dx \tag{3.49}$$

Равенство достигается при $f_1(x) = kf_2^*(x)$, где k — произвольная константа, а знак "*" обозначает комплексно сопряженное значение. Если отождествить $H(f)$ с $f_1(x)$ и $S(f)e^{2\pi ifT}$ с $f_2(x)$, можем записать следующее.

$$\left|\int_{-\infty}^{\infty} H(f) S(f) e^{2\pi ifT} df\right|^2 \le \int_{-\infty}^{\infty} |H(f)|^2\, df \int_{-\infty}^{\infty} |S(f)|^2\, df \tag{3.50}$$

Подстановка в выражение (3.48) дает

$$\left(\frac{S}{N}\right)_T \le \frac{2}{N_0} \int_{-\infty}^{\infty} |S(f)|^2\, df \tag{3.51}$$

или

$$\max\left(\frac{S}{N}\right)_T = \frac{2E}{N_0}, \tag{3.52}$$

где энергия E входящего сигнала $s(t)$ равна следующему.

$$E = \int_{-\infty}^{\infty} |S(f)|^2\, df \tag{3.53}$$

Следовательно, максимальный выход $(S/N)_T$ зависит от *энергии входящего сигнала* и спектральной плотности мощности шума, *но не от конкретной формы* сигнала.

Равенство в выражении (3.52) получается только при использовании оптимальной передаточной функции фильтра $H_0(f)$.

$$H(f) = H_0(f) = kS^*(f) e^{2\pi ifT} \tag{3.54}$$

или

$$h(t) = \mathfrak{F}^{-1}\left\{kS^*(f) e^{2\pi ifT}\right\} \tag{3.55}$$

Поскольку $s(t)$ — вещественный сигнал, с помощью формул (A.29) и (A.31) можно записать следующее.

$$h(t) = \begin{cases} ks(T-t) & 0 \le t \le T \\ 0 & \text{для остальных } t \end{cases} \tag{3.56}$$

Итак, импульсная характеристика фильтра, дающего максимальное отношение сигнал/шум на выходе, является зеркальным отображением сигнала сообщения $s(t)$, за-

паздывающим на время передачи символа T. Отметим, что задержка в T секунд делает уравнение (3.56) *причинным*, т.е. запаздывание на T секунд делает $h(t)$ функцией положительного времени в промежутке $0 \le t \le T$. Без задержки в T секунд отклик $s(-t)$ нереализуем, поскольку в этом случае он является функцией отрицательного времени.

3.2.3. Реализация корреляции в согласованном фильтре

В формуле (3.56) и на рис. 3.7, *а* отражено основное свойство согласованного фильтра: *импульсная характеристика такого фильтра — это зеркальное отображение сигнала с некоторой задержкой (относительно оси $t = 0$)*. Следовательно, если сигнал равен $s(t)$, его зеркальное отображение равно $s(-t)$, а зеркальное отображение, запаздывающее на T секунд, — это $s(T - t)$. Выход $z(t)$ причинного фильтра во временной области можно описать как свертку принятого входного сигнала $r(t)$ с импульсной характеристикой фильтра (см. раздел А.5).

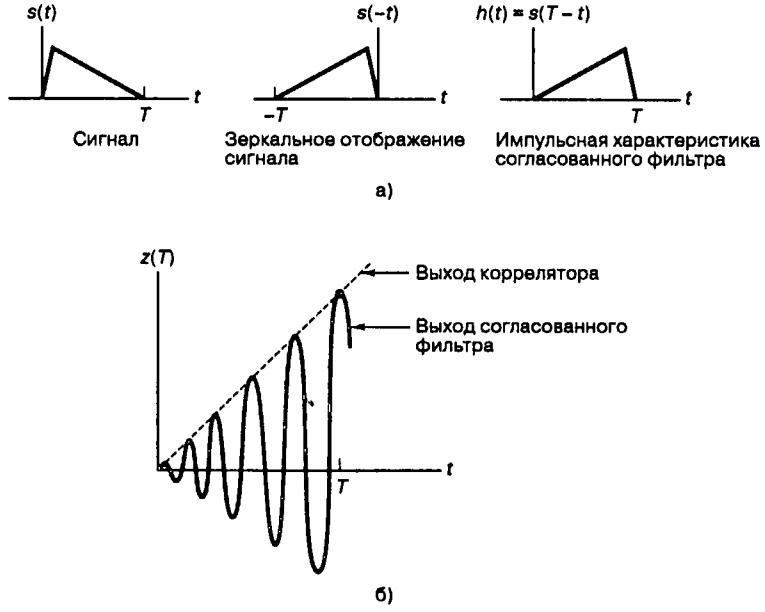


Рис. 3.7. Коррелятор и согласованный фильтр: а) характеристика согласованного фильтра; б) сравнение выходов коррелятора и согласованного фильтра

$$z(t) = r(t) * h(t) = \int_0^t r(\tau)h(t - \tau)\,d\tau \tag{3.57}$$

Подставляя $h(t)$ из формулы (3.56) в $h(t - \tau)$ в формуле (3.57) и выбирая произвольную константу k равной единице, получаем следующее.

$$\begin{aligned} z(t) &= \int_0^t r(\tau)s[T - (t - \tau)]\,d\tau = \\ &= \int_0^t r(\tau)s(T - t + \tau)\,d\tau \end{aligned} \tag{3.58}$$

Для момента времени $t = T$ формулу (3.58) можно переписать следующим образом.

$$z(T) = \int_0^T r(\tau)s(\tau)d\tau \tag{3.59}$$

Из последнего выражения видно, что интеграл от произведения принятого сигнала $r(t)$ на копию переданного сигнала $s(t)$ на интервале передачи символа представляет собой *корреляцию* $r(t)$ с $s(t)$. Предположим, что принятый сигнал $r(t)$ коррелирует со всеми сигналами-прототипами $s_i(t)$ ($i = 1, \ldots, M$) и для этого используется набор из M корреляторов. Сигнал $s_i(t)$, корреляция которого (или интеграл от произведения) с $r(t)$ дает максимальное значение $z_i(T)$, — и есть сигнал, который согласуется с $r(t)$ лучше остальных. Далее это свойство корреляции мы будем использовать для оптимального обнаружения сигналов.

3.2.3.1. Сравнение свертки и корреляции

Работа согласованного фильтра описывается математической операцией *свертки*; сигнал сворачивается с импульсной характеристикой фильтра. Работа коррелятора описывается математической операцией *корреляции*; сигнал коррелирует с копией самого себя. Довольно часто термин "согласованный фильтр" используется как синоним термина "коррелятор". Как такое возможно, если математические операции различны? Напомним, что процесс свертки двух сигналов использует один из сигналов, обращенный во времени. Кроме того, импульсная характеристика согласованного фильтра определяется именно через сигнал, обращенный во времени. Следовательно, свертка в согласованном фильтре с обращенной во времени функцией дает еще одно обращение во времени, подавая на выход (в конце интервала передачи символа) то, что является корреляцией сигнала с собственной копией. Значит, принимающий фильтр, изображенный на рис. 3.1, можно реализовать либо как согласованный фильтр, либо как коррелятор. Важно отметить, что выходы коррелятора и согласованного фильтра одинаковы *только в момент времени* $t = T$. Для синусоидального входа выход коррелятора, $z(t)$, на промежутке $0 \le t \le T$ приблизительно описывается линейной функцией. В то же время выход согласованного фильтра приблизительно описывается синусоидой, амплитуда которой в том же промежутке времени модулирована линейной функцией (см. рис. 3.7, *б*). Поскольку при соизмеримых входах выходы согласованного фильтра и коррелятора идентичны в момент взятия выборки $t = T$, функции согласованного фильтра и коррелятора, изображенные на рис. 3.8, часто используются как взаимозаменяемые.

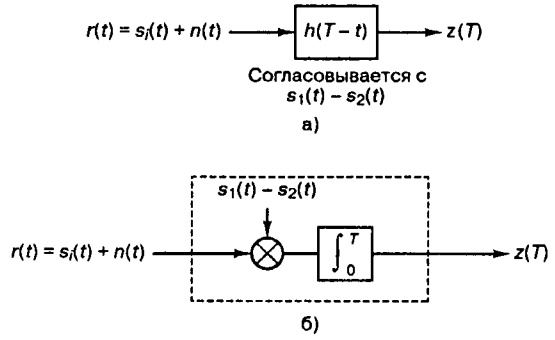


Рис. 3.8. Эквивалентность согласованного фильтра и коррелятора: а) согласованный фильтр; б) коррелятор

3.2.3.2. Дилемма в представлении упорядоченных во времени событий

При представлении упорядоченных во времени событий существует серьезная проблема. Возникает частая ошибка в области электротехники — путаница между самым старшим битом и самым младшим. На рис. 3.9, *а* показано, как обычно изображается функция времени; самое раннее событие представлено слева, а наиболее позднее — справа. Людям, привыкшим читать слева направо, такое изображение кажется единственно правильным. Рассмотрим рис. 3.9, *б*, где показано, как импульсы поступают в сеть (или канал) и покидают ее. Здесь самое раннее событие изображено справа, а наиболее позднее — слева. Изучение этого рисунка позволяет понять, что при записи упорядоченных событий возможна путаница между двумя возможными форматами записи. Чтобы избежать затруднений, зачастую необходимо дать некоторые пояснения (например, указать, что крайний справа бит — это первый бит).

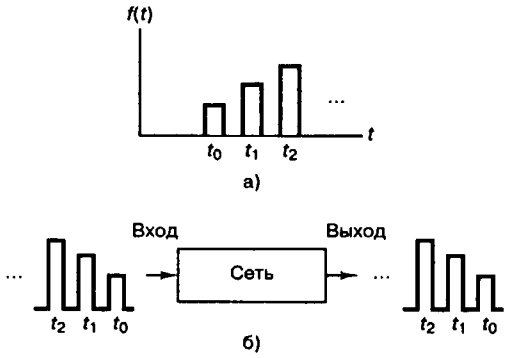


Рис. 3.9. Дилемма в представлении упорядоченных во времени событий

Математические соотношения часто имеют "врожденные" особенности, гарантирующие соответствующее упорядочение событий. Например, в разделе 3.2.3 согласованный фильтр определялся как имеющий импульсную характеристику $h(t)$ — запаздывающую версию обращенной во времени копии сигнала. Иными словами, $h(t) = s(T - t)$. Запаздывание на один интервал передачи символа T необходимо для того, чтобы фильтр был причинным (выход должен быть функцией положительного времени). Обращение во времени можно рассматривать как "предварительную коррекцию", где крайняя правая часть временного графика теперь соответствует наиболее раннему событию. Поскольку свертка навязывает другое обращение во времени, поступающий сигнал и импульсный отклик фильтра будут "идти в ногу" (ранний с ранним, поздний с поздним).

3.2.4. Оптимизация вероятности ошибки

Для оптимизации (минимизации) P_B в среде канала и приемника с шумом AWGN, показанных на рис. 3.1, нужно выбрать оптимальный принимающий фильтр на этапе 1 и оптимальный порог принятия решения на этапе 2. Для двоичного случая оптимальный порог принятия решения уже выбран и дается формулой (3.32), а в формуле (3.42) показано, что вероятность ошибки при таком пороге равна $P_B = Q[(a_1 - a_2)/2\sigma_0]$. Для минимального P_B в общем случае необходимо выбрать фильтр (согласованный) с максимальным аргументом функции $Q(x)$. Следовательно, нужно определить максимальное $(a_1 - a_2)/2\sigma_0$, что равносильно максимальному

$$\frac{(a_1 - a_2)^2}{\sigma_0^2}, \tag{3.60}$$

где $(a_1 - a_2)$ — разность желательных компонентов сигнала на выходе фильтра в момент $t = T$, а квадрат этого разностного сигнала представляет его мгновенную мощность. В разделе 3.2.2 описывался согласованный фильтр с максимальным отношением сигнал/шум (signal-to-noise ratio — SNR) для данного известного сигнала. Здесь мы решаем вопрос двоичной передачи сигналов и ищем оптимальный фильтр с максимальной разностью двух возможных выходных сигналов. В выводе, приведенном в уравнениях (3.45)–(3.52), было показано, что согласованный фильтр дает на выходе максимально возможное отношение SNR, равное $2E/N_0$. Допустим, что фильтр согласовывает входящий разностный сигнал $[s_1(t) - s_2(t)]$. Следовательно, в момент $t = T$ можем записать отношение SNR на выходе.

$$\left(\frac{S}{N}\right)_T = \frac{(a_1 - a_2)^2}{\sigma_0^2} = \frac{2E_d}{N_0}, \tag{3.61}$$

где $N_0/2$ — двусторонняя спектральная плотность мощности шума на входе фильтра и

$$E_d = \int_0^T [s_1(t) - s_2(t)]^2 dt \tag{3.62}$$

является энергией разностного сигнала на входе фильтра. Отметим, что уравнение (3.61) не представляет отношения SNR для какой-то отдельной передачи, $s_1(t)$ или $s_2(t)$. Это отношение дает метрику разности сигналов на выходе фильтра. Максимизируя выходное отношение SNR, как показано в уравнении (3.61), согласованный фильтр обеспечивает максимальное расстояние (нормированное на шум) между двумя возможными выходами — сигналами a_1 и a_2.

Далее, объединяя уравнения (3.42) и (3.61), получаем следующее.

$$P_B = Q\left(\sqrt{\frac{E_d}{2N_0}}\right) \tag{3.63}$$

Для согласованного фильтра уравнение (3.63) является важным промежуточным результатом, включающим энергию разностного сигнала на входе фильтра. Из этого уравнения можно вывести более общее соотношение для энергии принятого бита. Для начала определим временной коэффициент взаимной корреляции ρ, который будем использовать в качестве меры подобия двух сигналов $s_1(t)$ и $s_2(t)$. Имеем

$$\rho = \frac{1}{E_b} \int_0^T s_1(t) s_2(t)\, dt \tag{3.64,а}$$

и

$$\rho = \cos\theta, \tag{3.64,б}$$

где $-1 \le \rho \le 1$. Формула (3.64,а) — это классический математический способ выражения корреляции. Впрочем, если рассматривать $s_1(t)$ и $s_2(t)$ как векторы сигналов $\mathbf{s}_1$ и $\mathbf{s}_2$, то более удобным представлением ρ является формула (3.64,б). Векторное представление позволяет получать удобные графические изображения. Векторы $\mathbf{s}_1$ и $\mathbf{s}_2$ разделены углом θ; при малом угле векторы достаточно подобны (сильно коррелируют), а при больших углах они отличаются. Косинус угла θ дает ту же нормированную метрику корреляции, что и формула (3.64,а).

Расписывая выражение (3.62), получаем следующее.

$$E_d = \int_0^T s_1^2(t)\,dt + \int_0^T s_2^2(t)\,dt - 2\int_0^T s_1(t)s_2(t)\,dt \tag{3.65}$$

Напомним, что два первых члена формулы (3.65) представляют энергию, связанную с битом, E_b.

$$E_b = \int_0^T s_1^2(t)\,dt = \int_0^T s_2^2(t)\,dt \tag{3.66}$$

Подставляя уравнения (3.64,а) и (3.66) в формулу (3.65), получаем следующее.

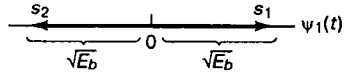
$$E_d = E_b + E_b - 2\rho E_b = 2E_b(1-\rho) \tag{3.67}$$

Подставляя уравнение (3.67) в (3.63), получаем следующее.

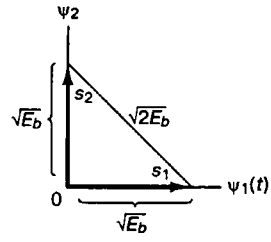
$$P_B = Q\left(\sqrt{\frac{E_b(1-\rho)}{N_0}}\right) \tag{3.68}$$

Рассмотрим случай $\rho = 1$, соответствующий наилучшей корреляции сигналов $s_1(t)$ и $s_2(t)$ в течение времени передачи символа (если сигналы изобразить как векторы, угол между ними будет равен нулю). Возможно ли, чтобы подобные сигналы использовались кем-то в реальной системе? Разумеется, нет, поскольку сигналы связи (элементы алфавита) должны быть максимально несопоставимы, чтобы их можно было легко различать (обнаруживать). В данный момент мы просто рассматриваем возможные значения ρ. Следующий частный случай $\rho = -1$ соответствует "антикорреляции" $s_1(t)$ и $s_2(t)$ в течение времени передачи символа. Другими словами, угол между векторами сигналов составляет 180°. В этом случае, когда векторы являются зеркальными отображениями друг друга, как показано на рис. 3.10, *а*, сигналы называются *антиподными*. Рассмотрим также случай $\rho = 0$, соответствующий нулевой корреляции между $s_1(t)$ и $s_2(t)$ (угол между векторами равен 90°). Такие сигналы, показанные на рис. 3.10, *б*, именуются *ортогональными*. Чтобы два сигнала были ортогональными, они не должны коррелировать в течение времени передачи символа, т.е. должно выполняться следующее условие.

$$\int_0^T s_1(t)s_2(t)\,dt = 0 \tag{3.69}$$



а)



б)

Рис. 3.10. Векторы двоичных сигналов: а) антиподные; б) ортогональные

Вопрос ортогональности рассматривался ранее, в разделе 3.1.3. При обнаружении антиподных сигналов (т.е. при $\rho = -1$) с помощью согласованного фильтра, уравнение (3.68) можно записать следующим образом.

$$P_B = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \tag{3.70}$$

Точно так же при обнаружении ортогональных сигналов (т.е. при $\rho = 0$) с помощью согласованного фильтра, формулу (3.68) можно записать следующим образом.

$$P_B = Q\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{3.71}$$

На рис. 3.10, где амплитуды сигналов выбраны равными $\sqrt{E_b}$, показано, что вероятность ошибки, описываемая уравнениями (3.70) и (3.71), является функцией расстояния между $\mathbf{s}_1$ и $\mathbf{s}_2$ (чем больше расстояние, тем меньше P_B). Если взять антиподные сигналы (рис. 3.10, *а*), расстояние между ними будет равно $2\sqrt{E_b}$, а энергия E_d, связанная с расстоянием, будет выражаться как квадрат расстояния, или $4E_b$. При подстановке $E_d = 4E_b$ в уравнение (3.63) получаем уравнение (3.70). Если взять ортогональные сигналы (рис. 3.10, *б*), расстояние между ними будет равно $\sqrt{2E_b}$; следовательно, $E_d = 2E_b$. При подстановке $E_d = 2E_b$ в уравнение (3.63) получим уравнение (3.71).

Пример 3.2. Обнаружение антиподных сигналов с помощью согласованного фильтра

Рассмотрим бинарную систему связи, принимающую равновероятные сигналы $s_1(t)$ и $s_2(t)$ плюс шум AWGN (рис. 3.11). Предположим, что в качестве принимающего фильтра используется согласованный фильтр и спектральная плотность мощности шума N_0 равна 10^{-12} Вт/Гц. С помощью значения напряжения и времени принятого сигнала, показанных на рис. 3.11, вычислите вероятность появления ошибочного бита.

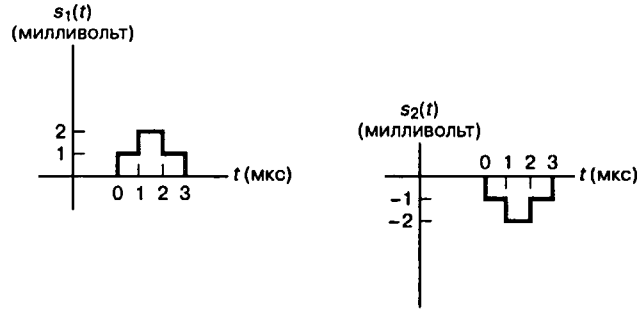


Рис. 3.11. Узкополосные аналоговые сигналы

Решение

Мы можем графически определить отношение принятой энергии на бит сигнала, используя для этого один из двух графиков, либо $s_1(t)$, либо $s_2(t)$, представленных на рис. 3.11. Энергия — это площадь под графиком импульса, которая находится путем интегрирования.

$$E_b = \int_0^3 v^2(t)\,dt = (10^{-3}\text{В})^2 \times (10^{-6}\text{ с}) + (2 \times 10^{-3}\text{В})^2 \times (10^{-6}\text{ с}) + (10^{-3}\text{В})^2 \times (10^{-6}\text{ с}) = 6 \times 10^{-12}\text{ Дж}$$

Поскольку сигналы, изображенные на рис. 3.11, являются антиподными и обнаруживаются с помощью согласованного фильтра, используем формулу (3.70) для вычисления вероятности появления ошибочного бита.

$$Q\left(\sqrt{\frac{12\times10^{-12}}{10^{-12}}}\right) = Q\left(\sqrt{12}\right) = Q(3{,}46)$$

Из табл. Б.1 находим, что $P_B = 3 \times 10^{-4}$. Кроме того, поскольку аргумент $Q(x)$ больше 3, можно также использовать приближенное соотношение, приведенное в формуле (3.44), которое дает вероятность $P_B \approx 2{,}9 \times 10^{-4}$.

Поскольку принятые сигналы являются антиподными и принимаются согласованным фильтром, весьма вероятно, что формула (3.70) дает верное выражение для нахождения вероятности возникновения ошибочного бита. Сигналы $s_1(t)$ и $s_2(t)$ могут выглядеть гораздо более странно, но до тех пор, пока они являются антиподными и обнаруживаются с помощью согласованного фильтра, их внешний вид не влияет на вычисление P_B. Формы сигналов, разумеется, *имеют значение*, но только когда дело доходит до определения импульсного отклика согласованного фильтра, необходимого для обнаружения этих сигналов.

3.2.5. Вероятность возникновения ошибки при двоичной передаче сигналов

3.2.5.1. Униполярная передача сигналов

На рис. 3.12, *а* приведен пример узкополосной ортогональной передачи сигналов, называемой униполярной.

$$\begin{aligned} s_1(t) &= A \qquad 0 \le t \le T \quad \text{для двоичной 1} \\ s_2(t) &= 0 \qquad 0 \le t \le T \quad \text{для двоичного 0} \end{aligned} \tag{3.72}$$

Здесь $A > 0$ — амплитуда сигнала $s_1(t)$. Определение ортогональной передачи сигналов дается выражением (3.69), требующим, чтобы $s_1(t)$ и $s_2(t)$ имели нулевую корреляцию в течение периода передачи символа.

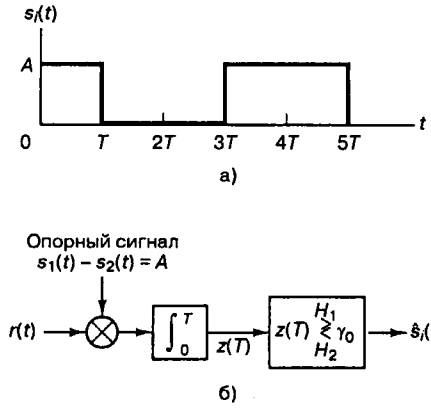


Рис. 3.12. Обнаружение при униполярной узкополосной передаче сигналов: а) пример униполярной передачи сигналов; б) обнаружение с помощью коррелятора

Поскольку в формуле (3.72) $s_2(t)$ равно нулю в течение периода передачи символа, множество униполярных импульсов полностью удовлетворяет условию, приведенному в уравнении (3.69), а следовательно, они формируют ортогональное множество сигналов. Рассмотрим униполярную передачу сигналов (рис. 3.12, *а*) и коррелятор (рис. 3.12, *б*), который может использоваться для обнаружения подобных импульсов. Коррелятор перемножает входящий сигнал $r(t)$ и разность сигналов-прототипов, $[s_1(t) - s_2(t)] = A$, после чего результат интегрируется. По окончании периода передачи символа T устройство дискретизации (включающееся в момент, определенный как верхний предел интегрирования) дает тестовую статистику $z(T)$, которая затем сравнивается с порогом γ_0. В случае приема $s_1(t)$ и шума AWGN (т.е. когда $r(t) = s_1(t) + n(t)$) сигнальный компонент $z(T)$ находится с помощью уравнения (3.69).

$$a_1(T) = \mathrm{E}\{z(T) \mid s_1(t)\} = \mathrm{E}\left\{\int_0^T A^2 + An(t)\,dt\right\} = A^2T$$

Здесь $\mathrm{E}\{z(T)|s_1(t)\}$ — *математическое ожидание* того, что при принятой выборке $z(T)$ был передан сигнал $s_1(t)$. Далее использовано равенство $\mathrm{E}\{n(t)\} = 0$. Подобным образом при $r(t) = s_2(t) + n(t)$, $a_2(T) = 0$. Таким образом, в рассматриваемом случае оптимальный порог принятия решения (см. уравнение (3.32)) равен $\gamma_0 = (a_1 + a_2)/2 = 1/2\, A^2T$. Если тестовая статистика $z(T)$ больше γ_0, сигнал считается равным $s_1(t)$; в противном случае принимается решение, что был передан сигнал $s_2(t)$.

Из уравнения (3.62) получаем, что энергетический разностный сигнал равен $E_d = A^2T$. Тогда из формулы (3.63) получаем вероятность появления на выходе ошибочного бита.

$$P_B = Q\left(\sqrt{\frac{E_d}{2N_0}}\right) = Q\left(\sqrt{\frac{A^2T}{2N_0}}\right) = \left(\sqrt{\frac{E_b}{N_0}}\right), \tag{3.73}$$

где при равновероятной передаче сигналов *средняя* энергия на бит равна $E_b = A^2T/2$. Уравнение (3.73) совпадает с уравнением (3.71), полученным с помощью общих рассуждений для ортогональной передачи сигналов.

Отметим, что вне блока перемножения, подобного показанному на рис. 3.12, *б*, единицей измерения сигнала является вольт. Следовательно, для сигналов напряжения на каждом из двух входов передаточная функция блока перемножения должна иметь размерность 1/вольт, а функция $r(t)\,s_i(t)$ вне блока перемножения — вольт/вольт в квадрате. Подобным образом вне блока интегрирования также используется единица измерения вольт. Следовательно, для сигнала напряжения в блоке интегрирования передаточная функция интегратора должна иметь размерность 1/секунду, а значит, общая передаточная функция блока перемножения-интегрирования должна иметь размерность 1/вольт-секунда. Итак, для сигнала, поступающего на интегратор и имеющего размерность энергии (вольт в квадрате-секунда), получаем с выхода сигнал, пропорциональный энергии принятого сигнала (вольт/джоуль).

3.2.5.2. Биполярная передача сигналов

На рис. 3.13, *а* приведен пример узкополосной антиподной передачи сигналов, называемой биполярной, где

$$s_1(t) = +A \qquad 0 \le t \le T \qquad \text{для двоичной 1}$$

и (3.74)

$$s_2(t) = -A \qquad 0 \le t \le T \qquad \text{для двоичного 0}$$

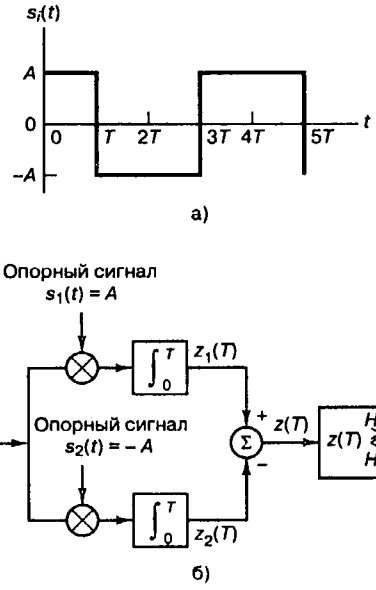


Рис. 3.13. Обнаружение при биполярной узкополосной передаче сигналов: а) пример биполярной передачи сигналов; б) обнаружение с помощью коррелятора

Как определялось ранее, термин "антиподный" относится к двоичным сигналам, которые являются зеркальными отображениями друг друга, т.е. $s_1(t) = -s_2(t)$. Приемник-коррелятор таких антиподных сигналов может иметь схему, подобную представленной на рис. 3.13, *б*. Один коррелятор перемножает входящий сигнал $r(t)$ и сигнал-прототип $s_1(t)$, после чего интегрирует результат; второй выполняет те же действия с сигналом $s_2(t)$.

На рис. 3.13, *б* изображена сама суть основной функции цифрового приемника. Иными словами, в течение периода передачи символа входящий зашумленный сигнал пускается по множественным различным "проходам" для проверки его корреляции со всеми возможными прототипами. После этого приемник определяет наибольшее выходное напряжение (наилучшее соответствие) и принимает соответствующее решение относительно значения переданного символа. В бинарном случае имеем два возможных прототипа. В квадратичном случае могут существовать 4 возможности и т.д. На рис. 3.13, *б* выходы коррелятора обозначены как $z_i(T)$ ($i = 1, 2$). Тестовая статистика, сформированная из разности выходов коррелятора, выглядит следующим образом.

$$z(T) = z_1(T) - z_2(T) \tag{3.75}$$

Решение принимается с использованием порога, указанного в формуле (3.32). Для антиподных сигналов $a_1 = -a_2$; следовательно, $\gamma_0 = 0$. Значит, если тестовая статистика $z(T)$ положительна, считается, что передан сигнал $s_1(T)$; если же тестовая статистика отрицательна, считается, что передан сигнал $s_2(T)$.

Из уравнения (3.62) энергетический разностный сигнал равен $E_d = (2A)^2T$. Следовательно, можем использовать уравнение (3.63) для вычисления вероятности появления ошибочного бита.

$$P_B = Q\left(\sqrt{\frac{E_d}{2N_0}}\right) = Q\left(\sqrt{\frac{2A^2T}{N_0}}\right) = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \tag{3.76}$$

Здесь средняя энергия на бит равна $E_b = A^2T$. Уравнение (3.76) совпадает с уравнением (3.70), полученным с помощью общих рассуждений для ортогональной передачи сигналов.

3.2.5.3. Использование базисных функций для описания передачи сигналов

В корреляторе, приведенном на рис. 3.13, *б*, в качестве опорных могут использоваться не только сигналы $s_i(t)$; с этой же целью могут применяться *базисные функции*, описанные в разделе 3.1.3. Проиллюстрируем этот подход на бинарной передаче сигналов с помощью униполярных или биполярных импульсов, поскольку в этом случае все сигнальное пространство можно охарактеризовать одной базисной функцией. Если нормировать пространство, т.е. в уравнении (3.76) положить $K_j = 1$, то базисная функция $\psi_1(t)$ будет равна $\sqrt{1/T}$.

Для униполярной передачи импульсов можем записать следующее.

$$s_1(t) = a_{11}\psi_1(t) = A\sqrt{T} \times \left(\sqrt{\frac{1}{T}}\right) = A$$

и

$$s_2(t) = a_{21}\psi_1(t) = 0 \times \left(\sqrt{\frac{1}{T}}\right) = 0$$

Здесь коэффициенты a_{11} и a_{21} равны, соответственно, $A\sqrt{T}$ и 0.

Для биполярной передачи импульсов можем записать

$$s_1(t) = a_{11}\psi_1(t) = A\sqrt{T} \times \left(\sqrt{\frac{1}{T}}\right) = A$$

и

$$s_2(t) = a_{21}\psi_1(t) = -A\sqrt{T} \times \left(\sqrt{\frac{1}{T}}\right) = -A,$$

где коэффициенты a_{11} и a_{21} равны, соответственно, $A\sqrt{T}$ и $-A\sqrt{T}$. При использовании антиподных импульсов можно считать, что приемник-коррелятор имеет вид, показанный на рис. 3.12, *б*, с опорным сигналом, равным $\sqrt{1/T}$. Итак, при передаче $s_1(t) = A$, можем записать следующее.

$$a_1(T) = \mathbf{E}\{z(T)|s_1(t)\} = \mathbf{E}\left\{\int_0^T \frac{A}{\sqrt{T}} + \frac{n(t)}{\sqrt{T}}dt\right\} = A\sqrt{T}$$

Поскольку $\mathbf{E}\{n(t)\} = 0$, а значит для антиподной передачи сигналов $E_b = A^2T$, то $a_1(T) = \sqrt{E_b}$. Аналогично при приеме сигнала $r(t) = s_2(t) + n(t)$ получаем $a_2(T) = -\sqrt{E_b}$. Если опорные сигналы рассматривать именно таким образом, то математическое ожидание $z(T)$ равно $\sqrt{E_b}$ (измеряется в нормированных вольтах, пропорциональных принятой энергии). Приведенный подход к описанию корреляторa дает удобное выражение $z(T)$, имеющее те же единицы измерения (вольт), что используются вне блоков перемножения и интегрирования. Еще раз повторим важный момент: на выходе устройства дискретизации (в додетекторной точке) тестовая статистика $z(T)$ — это сигнал напряжения, пропорциональный энергии принятого сигнала.

На рис. 3.14 показана зависимость P_B от E_b/N_0 для биполярной и униполярной передачи сигналов. Существует только два точных способа сравнения этих кривых. Проведем вертикальную линию при некотором данном отношении E_b/N_0, скажем 10 дБ. Видим, что униполярная передача сигналов дает вероятность P_B порядка 10^{-3}, а биполярная — порядка 10^{-6}. Нижняя кривая соответствует лучшей достоверности передачи. Можно также провести горизонтальную линию при некотором требуемом уровне P_B, скажем 10^{-5}. Видим, что при униполярной передаче сигналов каждый принятый бит потребует отношения E_b/N_0 порядка 12,5 дБ, а при биполярной передаче — не более 9,5 дБ. Разумеется, более низкие требования лучше (требуется меньшая мощность, меньшая полоса). Вообще, более достоверным схемам соответствуют кривые, расположенные ближе к левой и нижней осям. Изучая кривые на рис. 3.14, видим, что биполярная схема имеет выигрыш в 3 дБ по сравнению с униполярной. Это отличие могло быть предсказано ранее, поскольку отношение E_b/N_0 в формулах (3.70) и (3.71) отличалось в 2 раза. В главе 4 будет показано, что при обнаружении с использованием согласованного фильтра *полосовая* антиподная передача сигналов (например, двоичная фазовая манипуляция) дает такое же значение P_B, как и *узкополосная* антиподная передача сигналов (например, с помощью биполярных импульсов). Также будет показано, что при обнаружении с помощью согласованного фильтра *полосовая* ортогональная передача сигналов (например, ортогональная частотная манипуляция) дает такое же значение P_B, как и *узкополосная* ортогональная передача сигналов (например, с использованием униполярных импульсов).

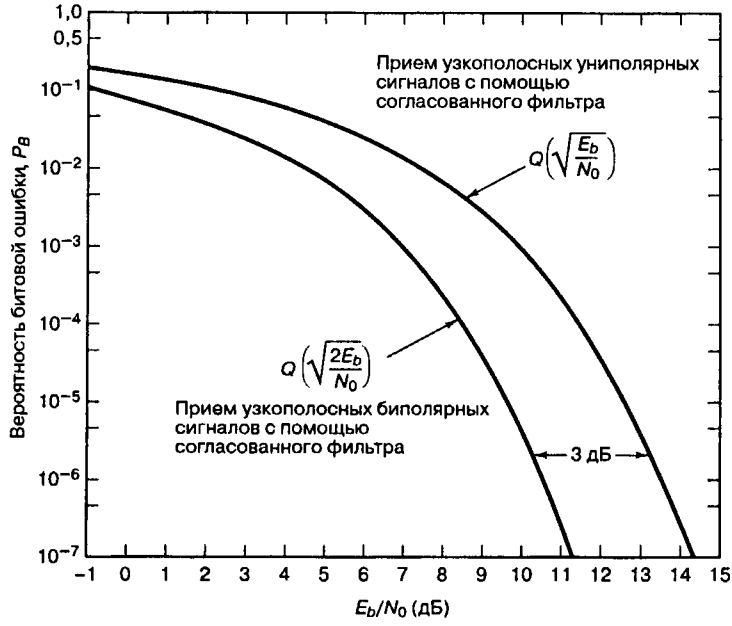


Рис. 3.14. Вероятность появления ошибочного бита при униполярной и биполярной передаче сигналов

3.3. Межсимвольная интерференция

На рис. 3.15, *а* представлены фильтрующие элементы типичной системы цифровой связи. В системе — передатчике, приемнике и канале — используется множество разнообразных фильтров (и реактивных элементов, таких как емкость и индуктивность). В передатчике информационные символы, описываемые как импульсы или уровни напряжения, модулируют импульсы, которые затем фильтруются для согласования с определенными ограничениями полосы. В узкополосных системах канал (кабель) имеет распределенное реактивное сопротивление, искажающее импульсы. Некоторые полосовые системы, такие как беспроводные, являются, по сути, каналами с замираниями (см. главу 15), которые проявляют себя как нежелательные фильтры, также искажающие сигнал. Если принимающий фильтр настраивается на компенсацию искажения, вызванного *как* передатчиком, *так и* каналом, он зачастую называется *выравнивающим* (equalizing filter) или *принимающим/выравнивающим* (recciving/equalizing). На рис. 3.15, *б* приведена удобная модель системы, объединяющая все следствия фильтрации в одну общесистемную передаточную функцию.

$$H(f) = H_t(f)\, H_c(f)\, H_r(f) \tag{3.77}$$

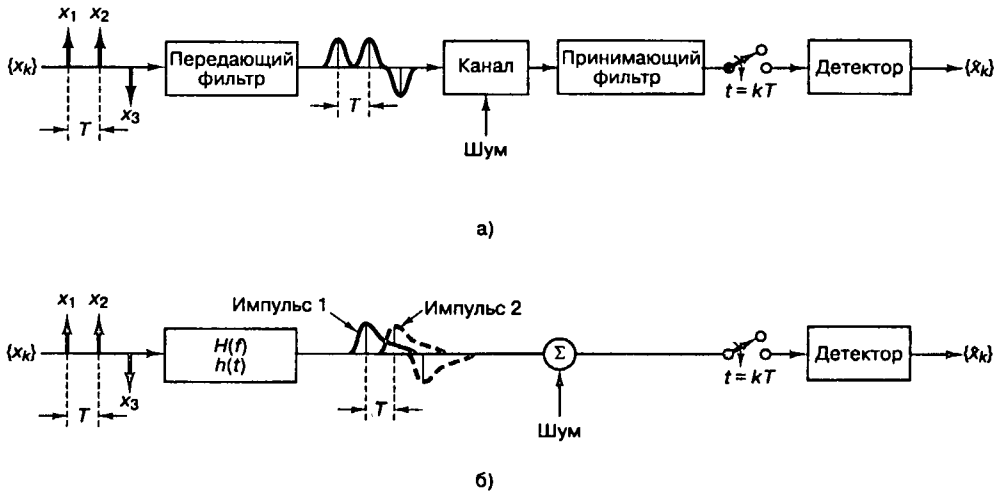


Рис. 3.15. Межсимвольная интерференция в процессе обнаружения: а) типичная узкополосная цифровая система; б) эквивалентная модель

Здесь $H_t(f)$ характеризует передающий фильтр, $H_c(f)$ — фильтрацию в канале, а $H_r(f)$ — принимающий/выравнивающий фильтр. Таким образом, характеристика $H_c(f)$ представляет передаточную функцию всей системы, отвечающую за все этапы фильтрации в различных местах цепочки передатчик-канал-приемник. В бинарной системе, использующей какую-нибудь распространенную кодировку PCM, например NRZ-L, детектор принимает решение относительно значения символа путем сравнения выборки принятого импульса с порогом. Например, детектор, изображенный на рис. 3.15, решает, что была послана двоичная единица, если принятый импульс положителен, или двоичный нуль — в противном случае. Вследствие системной фильтрации принятые импульсы могут перекрываться, как показано на рис. 3.15, *б*. Хвост импульса может "размываться" на соседний интервал передачи символа, таким образом мешая процессу обнаружения и повышая вероятность появления ошибки; подобный процесс получил название *межсимвольной интерференции* (intersymbol interference — ISI). Даже при отсутствии шумов воздействие фильтрации и искажение, вызванное каналом, приводят к возникновению ISI. Иногда функция $H_c(f)$ задается, и задача состоит в определении $H_t(f)$ и $H_r(f)$, минимизирующих ISI на выходе $H_r(f)$.

Исследованием проблемы задания формы принятого импульса с тем, чтобы предотвратить появление ISI на детекторе, долгое время занимался Найквист [6]. Он показал, что минимальная теоретическая ширина полосы системы, требуемая для определения R_s символов/секунду без ISI, равна $R_s/2$ Гц. Это возможно, если передаточная функция системы $H(f)$ имеет прямоугольную форму, как показано на рис. 3.16, *а*. Для узкополосных систем с такой $H(f)$, что односторонняя ширина полосы фильтра равна $1/2T$ (*идеальный фильтр Найквиста*), импульсная характеристика функции $H(f)$, вычисляемая с помощью обратного преобразования Фурье (см. табл. А.1), имеет вид $h(t) = \text{sinc}\,(t/T)$; она показана на рис. 3.16, *б*. Импульс, описываемый функцией sinc (t/T), называется *идеальным импульсом Найквиста*; он имеет бесконечную длительность и состоит из множественных лепестков: главного и боковых, именуемых *хвостами*. Найк-

вист установил, что если каждый импульс принятой последовательности имеет вид sinc (t/T), импульсы могут обнаруживаться без межсимвольной интерференции. На рис. 3.16, *б* показано, как удается обойти ISI. Итак, имеем два последовательных импульса, $h(t)$ и $h(t-T)$. Несмотря на то что хвосты функции $h(t)$ имеют бесконечную длительность, из рисунка видно, что в момент $t=T$ взятия выборки функции $h(t-T)$ хвост функции $h(t)$ проходит через точку нулевой амплитуды, и подобным образом он будет иметь нулевую амплитуду в моменты взятия выборок всех остальных импульсов последовательности $h(t-kT)$, $k=\pm1, \pm2, \ldots$. Следовательно, предполагая идеальную синхронизацию процесса взятия выборок, получаем, что межсимвольная интерференция не будет влиять на процесс обнаружения. Чтобы узкополосная система могла обнаруживать $1/T$ таких импульсов (символов) в секунду, ширина ее полосы должна быть равна $1/2T$; другими словами, система с шириной полосы $W=1/2T=R_s/2$ Гц может поддерживать максимальную скорость передачи $2W=1/T=R_s$ символов/с (*ограничение полосы по Найквисту*) без ISI. Следовательно, при идеальной фильтрации Найквиста (и нулевой межсимвольной интерференции) максимальная возможная скорость передачи символов на герц полосы, называемая *уплотнением скорости передачи символов* (symbol-rate packing), равна 2 символа/с/Гц. Вследствие прямоугольной формы передаточной функции идеального фильтра Найквиста и бесконечной длины соответствующего импульса, подобные идеальные фильтры нереализуемы; реализовать их можно только приближенно.

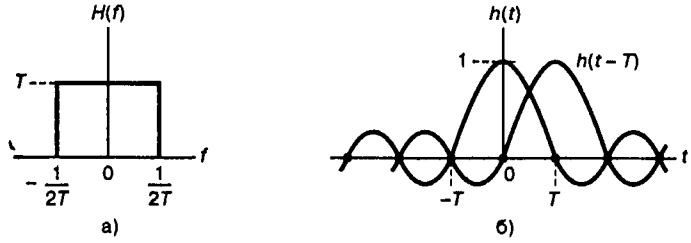


Рис. 3.16. Каналы Найквиста для нулевой межсимвольной интерференции: а) прямоугольная передаточная функция системы $H(f)$; б) принятый импульс $h(t)$ = sinc (t/T)

Стоит отметить, что названия "фильтр Найквиста" и "импульс Найквиста" часто используются для описания обширного класса фильтраций и формообразований, удовлетворяющих условию нулевой межсимвольной интерференции в точках взятия выборок. Фильтр Найквиста — это фильтр, передаточная функция которого может быть представлена прямоугольной функцией, свернутой с любой четно-симметричной частотной функцией. Импульс Найквиста — это импульс, форма которого может быть описана функцией sinc (t/T), умноженной на другую временную функцию. Следовательно, существует бесконечное множество фильтров Найквиста и соответствующих импульсов. В классе фильтров Найквиста наиболее популярными являются фильтры с характеристикой типа приподнятого косинуса или корня из приподнятого косинуса. Несколько позже эти фильтры будут рассмотрены подробно.

Основным параметром систем связи является *эффективность использования полосы*, R/W, измеряемая в бит/с/Гц. Как можно понять из единиц измерения, R/W представляет меру скорости переноса данных на единицу ширины полосы, а значит, показывает, насколько эффективно любой метод передачи сигналов использует ресурс поло-

сы. Поскольку ограничение ширины полосы по Найквисту устанавливает теоретическое максимальное уплотнение скорости передачи символов без межсимвольной интерференции, равное 2 символа/с/Гц, может возникнуть вопрос, можно ли что-то сказать об ограничении величин, измеряемых в бит/с/Гц. О последних ничего нельзя сказать прямо; ограничение связано только с импульсами или символами и возможностью обнаружения их амплитудных значений без искажения со стороны других импульсов. При нахождении *R/W* для любой схемы передачи сигналов необходимо знать, сколько битов представляет каждый символ, что само по себе является темой отдельного рассмотрения. Допустим, сигналы кодируются с использованием *M*-уровневой кодировки PAM. Каждый символ (включающий *k* бит) представляется одной из *M* импульсных амплитуд. Для $k = 6$ бит на символ размер набора символов составляет $M = 2^k = 64$ амплитуды. Таким образом, при 64-уровневой кодировке PAM теоретическая максимальная эффективность использования полосы, не допускающая межсимвольной интерференции, равна 12 бит/с/Гц. (Подробнее об эффективности использования полосы в главе 9.)

3.3.1. Формирование импульсов с целью снижения ISI

3.3.1.1. Цели и компромиссы

Чем компактнее спектр передачи сигналов, тем выше разрешенная скорость передачи данных или больше число пользователей, которые могут обслуживаться одновременно. Это имеет большое значение для поставщиков услуг связи, поскольку более эффективное использование доступной ширины полосы приносит больший доход. Для большинства систем связи (за исключением систем расширенного спектра, рассмотренных в главе 12) нашей задачей является максимальное сужение требуемой полосы системы. Найквист определил основное ограничение для такого сужения полосы. Но что произойдет, если заставить систему работать с меньшей полосой, чем определяется ограничением? Импульсы станут протяженнее по времени, что, вследствие увеличения межсимвольной интерференции, отрицательно скажется на достоверности передачи. Более разумным было бы сжатие полосы информационных импульсов до некоторого разумного значения, которое больше минимума, определенного Найквистом. Это выполняется путем формирования импульсов с помощью фильтра Найквиста. Если край полосы пропускания фильтра крутой, приблизительно соответствующий прямоугольной форме (рис. 3.16, *а*), то спектр сигнала можно сделать более компактным. В то же время использование подобного фильтра приводит к тому, что длительность импульсного отклика становится приблизительно равна бесконечности, как показано на рис. 3.16, *б*. Каждый импульс накладывается на все импульсы последовательности. Длительные отклики дают хвосты больших амплитуд около главного лепестка каждого импульса. Подобные хвосты нежелательны, поскольку, как видно из рис. 3.16, *б*, они вносят нулевую межсимвольную интерференцию *только в том случае*, если выборка производится *точно* в соответствующий момент времени; при больших хвостах даже небольшие ошибки синхронизации приведут к межсимвольной интерференции. Следовательно, хотя компактный спектр и позволяет оптимальным образом использовать полосу, он оказывается очень чувствительным к ошибкам синхронизации, приводящим к увеличению межсимвольной интерференции.

3.3.1.2. Фильтр с характеристикой типа приподнятого косинуса

Ранее говорилось, что принимающий фильтр часто называется *выравнивающим*, если он настраивается на компенсацию искажений, вносимых передатчиком и каналом. Другими словами, конфигурация этого фильтра выбрана так, чтобы оптимизировать общесистемную частотную передаточную функцию $H(f)$, описанную формулой (3.77). Одна из часто используемых передаточных функций $H(f)$ принадлежит к классу функций Найквиста (нулевая ISI в моменты взятия выборок) и называется *приподнятым косинусом* (raised-cosine). Описывается эта функция следующим выражением.

$$H(f) = \begin{cases} 1 & \text{для } |f| < 2W_0 - W \\ \cos^2\left(\frac{\pi}{4}\frac{|f| + W - 2W_0}{W - W_0}\right) & \text{для } 2W_0 - W < |f| < W \\ 0 & \text{для } |f| > W \end{cases} \tag{3.78}$$

Здесь W — максимальная ширина полосы, а $W_0 = 1/2T$ — минимальная ширина полосы по Найквисту для прямоугольного спектра и ширина полосы по уровню –6 дБ (или точка половинной амплитуды) для косинусоидального спектра. Разность $W - W_0$ называется "избытком полосы" (excess bandwidth); она означает дополнительную ширину полосы по сравнению с минимумом Найквиста (например, для прямоугольного спектра $W = W_0$). *Коэффициент сглаживания* (roll-off factor) определяется как $r = (W - W_0)/W_0$, где $0 \le r \le 1$. Коэффициент сглаживания — это избыток полосы, деленный на ширину полосы по уровню –6 дБ (т.е. относительный избыток полосы). Для данного W_0 выравнивание r задает требуемый избыток относительно W_0 и характеризует крутизну фронта характеристики фильтра. На рис. 3.17, *а* для нескольких значений коэффициента сглаживания r ($r = 0$, $r = 0{,}5$ и $r = 1$) показана характеристика типа приподнятого косинуса. Случай $r = 0$ соответствует минимальной ширине полосы по Найквисту. Отметим, что при $r = 1$ требуемый избыток полосы равен 100% и хвосты характеристики достаточно малы. Система с подобной спектральной характеристикой может поддерживать скорость передачи символов R_s символов/с при использовании полосы в R_s Гц (удвоенная минимальная полоса по Найквисту), что дает уплотнение скорости передачи, равное 1 символ/с/Гц. Импульсный отклик, соответствующий функции $H(f)$ и определяемый выражением (3.78), равен следующему.

$$h(t) = 2W_0(\operatorname{sinc} 2W_0 t)\frac{\cos\left[2\pi(W - W_0)t\right]}{1 - \left[4(W - W_0)t\right]^2} \tag{3.79}$$

Этот импульсный отклик изображен на рис. 3.17, *б* для $r = 0$, $r = 0{,}5$ и $r = 1$. Хвост имеет нулевые значения в каждый момент взятия выборки, вне зависимости от значения коэффициента сглаживания.

Фильтр, описанный уравнением (3.78), и импульс, представленный уравнением (3.79), можно реализовать только приблизительно, поскольку, строго говоря, спектр типа приподнятого косинуса физически не может быть реализован (причина та же, что и при реализации идеального фильтра Найквиста). Реализуемый фильтр должен иметь импульсный отклик конечной длительности и давать нулевой выход до момента включения импульса (см. раздел 1.7.2), что невозможно для семейства характеристик типа приподнятого косинуса. Эти нереализуемые фильтры являются *непри-*

чинными (импульсный отклик фильтра имеет бесконечную продолжительность и фильтрованный импульс начинается в момент $t = -\infty$). На практике фильтр формирования импульсов должен удовлетворять двум требованиям. Он должен обеспечивать желаемое сглаживание и должен быть реализуем (импульсный отклик должен усекаться до конечного размера).

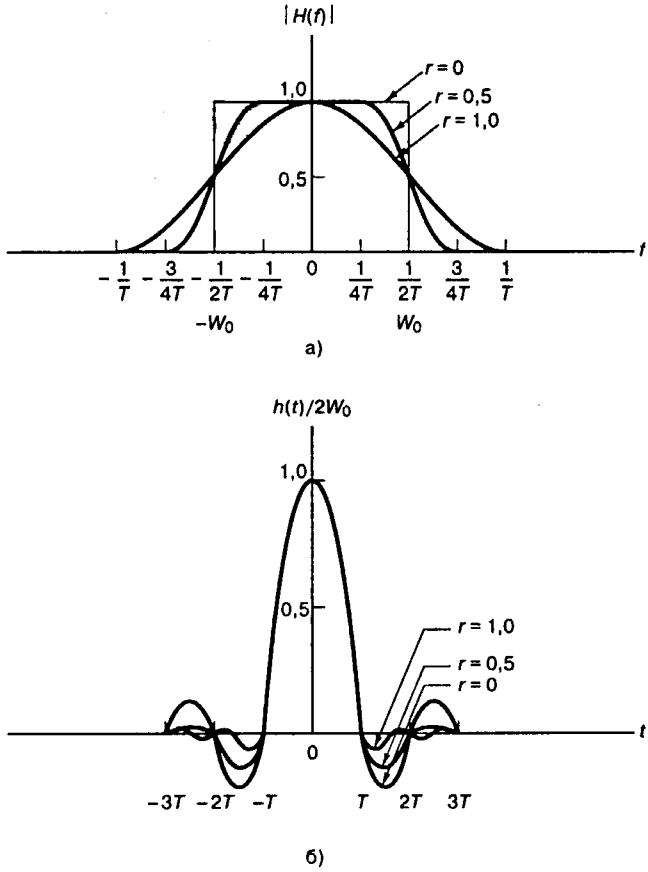


Рис. 3.17. Характеристики фильтров типа приподнятого косинуса: а) передаточная функция системы; б) импульсный отклик системы

Используя ограничение ширины полосы по Найквисту (минимальная ширина полосы W, требуемая для поддержания скорости R_s символов/с без межсимвольной интерференции, равна $R_s/2$ Гц), можно вывести более общее соотношение между требуемой полосой и скоростью передачи символов, включающее коэффициент сглаживания r.

$$W = \frac{1}{2}(1 + r)R_s \tag{3.80}$$

Таким образом, при $r = 0$ формула (3.80) описывает минимальную требуемую полосу для обеспечения идеальной фильтрации по Найквисту. При $r > 0$ ширина полосы превышает минимум Найквиста; следовательно, для этого случая R_s меньше удвоенной

ширины полосы. Если демодулятор подает на выход одну выборку на символ, теорема о дискретном представлении Найквиста нарушается, поскольку у нас остается слишком мало выборок для однозначного восстановления аналогового сигнала (присутствует наложение). Впрочем, в системах цифровой связи нас и не интересует восстановление аналоговых сигналов. Кроме того, поскольку семейство фильтров с характеристикой типа приподнятого косинуса характеризуется нулевой межсимвольной интерференцией в каждый момент произведения выборки из символа, мы по-прежнему можем добиться однозначного обнаружения.

Сигналы с полосовой модуляцией (см. главу 4), такие как сигналы с амплитудной (amplitude-shift keying — ASK) и фазовой манипуляцией (phase-shift keying — PSK), требуют вдвое большей полосы передачи, чем эквивалентные узкополосные сигналы (см. раздел 1.7.1). Такие смещенные по частоте сигналы занимают полосу, вдвое большую по ширине соответствующей узкополосной; зачастую их называют двухполосными (double-sideband — DSB). Следовательно, для сигналов в кодировках ASK и PSK соотношение между требуемой шириной полосы W_{DSB} и скоростью передачи символов R_s принимает следующий вид.

$$W_{DSB} = (1 + r)R_s \tag{3.81}$$

Напомним, что передаточная функция, имеющая вид приподнятого косинуса, — это общесистемная функция $H(f)$, описывающая "полный проход" сообщения, отправленного передатчиком (в виде импульса), через канал и принимающий фильтр. Фильтрация в приемнике описывается частью общей передаточной функции, тогда как подавление межсимвольной интерференции обеспечивает передаточная функция, имеющая вид приподнятого косинуса. Как следствие сказанного, принимающий и передающий фильтры часто выбираются (согласовываются) так, чтобы передаточная функция каждого имела вид квадратного корня из приподнятого косинуса. Подавление любой межсимвольной интерференции, внесенной каналом, обеспечивает произведение этих двух функций, которое дает общую передаточную функцию системы, имеющую вид приподнятого косинуса. Если же для уменьшения последствий привнесенной каналом межсимвольной интерференции вводится отдельный выравнивающий фильтр, принимающий и выравнивающий фильтры могут совместно настраиваться так, чтобы компенсировать искажение, вызванное как передатчиком, так и каналом; при этом общая передаточная функция системы характеризуется нулевой межсимвольной интерференцией.

Рассмотрим компромиссы, с которыми приходится сталкиваться при выборе фильтров формирования импульсов. Чем больше сглаживание фильтра, тем короче будут хвосты импульсов (из этого следует, что амплитуды хвостов также будут меньше). Меньшие хвосты менее чувствительны к ошибкам синхронизации, а значит, подвержены меньшему искажению вследствие межсимвольной интерференции. Отметим, что на рис. 3.17, *б* даже для $r = 1$ ошибка синхронизации по-прежнему приводит к некоторому увеличению межсимвольной интерференции. Но в то же время в этом случае проблема менее серьезна, чем при $r = 0$, поскольку при $r = 0$ хвосты сигнала $h(t)$ больше, чем при $r = 1$. Увеличение хвостов — это плата за повышение избытка полосы. С другой стороны, чем меньше сглаживание фильтра, тем меньше избыток полосы, а это позволяет повысить скорость передачи сигналов или число пользователей, которые могут одновременно использовать систему. В этом случае мы платим более длительными хвостами импульсов, большими их амплитудами, а следовательно, большей восприимчивостью к ошибкам синхронизации.

3.3.2. Факторы роста вероятности ошибки

Факторы повышения вероятности возникновения ошибки в цифровой связи могут быть следующими. Во-первых, это связано с простым падением мощности принятого сигнала или с повышением мощности шума или интерференции, что в любом случае приводит к уменьшению отношения сигнал/шум, или E_b/N_0. Во-вторых, это искажение сигнала, вызванное, например, межсимвольной интерференцией (intersymbol interference — ISI). Ниже показывается, чем отличаются эти факторы.

Предположим, нам нужна система связи с такой зависимостью вероятности появления ошибочного бита P_B от отношения E_b/N_0, какая изображена сплошной линией на рис. 3.18, *а*. Предположим, что после настройки системы и проведения измерений оказывается, к нашему разочарованию, что вероятность P_B соответствует не теоретической кривой, а кривой, показанной на рис. 3.18, *а* пунктиром. Причина снижения отношения E_b/N_0 — потеря сигналом мощности или повышение шума или интерференции. Желаемой вероятности ошибочного бита в 10^{-5} соответствует теоретическая величина $E_b/N_0 = 10$ дБ. Поскольку производительность реальной системы не соответствует теоретическим расчетам, нам следует использовать пунктирный график и добиться отношения E_b/N_0, равного 12 дБ (для получения той же вероятности $P_b = 10^{-5}$). Если причины проблемы устранить нельзя, то насколько большее отношение E_b/N_0 требуется теперь для получения необходимой вероятности ошибочного бита? Ответ, разумеется, — 2 дБ. Вообще, это может оказаться серьезной проблемой, особенно если система располагает ограниченной мощностью и получить дополнительные 2 дБ весьма сложно. Но все же уменьшение отношения E_b/N_0 *не смертельно*, по сравнению с ухудшением качества, вызванным искажением.

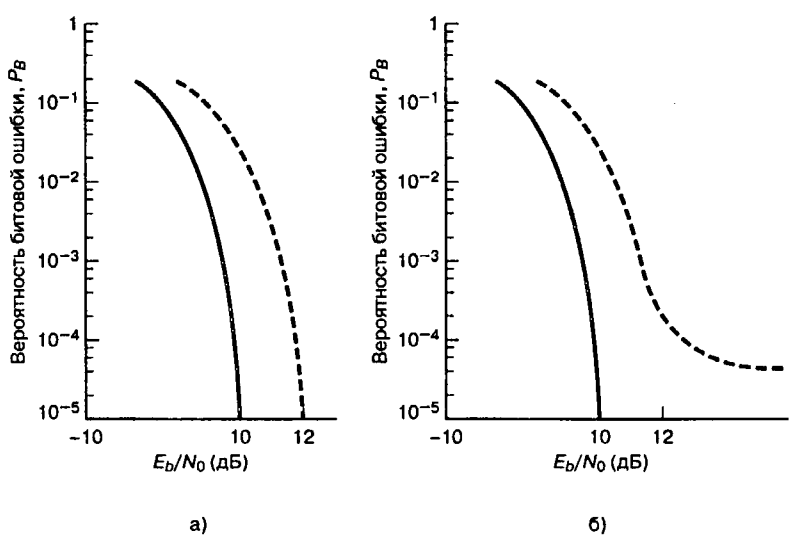


Рис. 3.18. Факторы роста вероятности ошибки: а) снижение E_b/N_0; б) непреодолимое ухудшение, вызванное искажением

Обратимся к рис 3.18, *б* и представим, что мы снова не получили желаемой вероятности, описываемой сплошной кривой. Но в этот раз причиной стало не уменьшение отношения сигнал/шум, а искажение, вызванное межсимвольной интерференцией (реальная кривая показана пунктиром). Если причину проблемы устранить нельзя,

то насколько большее отношение E_b/N_0 требуется теперь для получения необходимой вероятности ошибочного бита? В этом случае потребуется бесконечное увеличение. Другими словами, не существует такого E_b/N_0, которое позволило бы устранить проблему. Если непреодолимое ухудшение описывается такой кривой, как показана на рис. 3.18, *б*, то никакое увеличение E_b/N_0 не может дать желаемого результата (предполагается, что нижняя точка пунктирной кривой находится выше требуемой вероятности P_B). Безусловно, каждая кривая зависимости P_B от E_b/N_0 имеет где-то нижнюю точку, но если эта точка находится далеко за областью, представляющей практический интерес, то она уже не имеет значения.

Итак, увеличение отношения E_b/N_0 не всегда помогает решить проблему межсимвольной интерференции (особенно если кривая зависимости P_B от E_b/N_0 выходит за область практического интереса). Это можно понять, взглянув на перекрывающиеся импульсы на рис. 3.15, *б* — увеличение отношения E_b/N_0 никак не влияет на длительность области перекрытия, и степень искажения импульсов не изменится. Так что же обычно противопоставляют искажающему эффекту межсимвольной интерференции? В данной ситуации наиболее приемлемым является метод, именуемый выравниванием (см. раздел 3.4). Поскольку причиной межсимвольной интерференции является искажение вследствие фильтрации в передатчике и канале, выравнивание можно рассматривать как процесс, компенсирующий подобные неоптимальные эффекты фильтрации.

Пример 3.3. Требования к ширине полосы

а) Найдите минимальную ширину полосы, требуемую для узкополосной передачи последовательности четырехуровневых импульсов в кодировке PAM со скоростью $R = 2400$ бит/с, если передаточная характеристика системы имеет вид приподнятого косинуса со 100%-ным избытком полосы ($r = 1$).

б) Та же последовательность модулируется несущей, так что теперь узкополосный спектр смещен и центрирован на частоте f_0. Определите минимальную двустороннюю полосу, требуемую для передачи модулированной последовательности PAM. Передаточная характеристика считается такой же, как и в п. а.

Решение

а) $M = 2^k$, поскольку $M = 4$ уровня, $k = 2$.

Скорость передачи символов или импульсов $R_s = \dfrac{R}{k} = \dfrac{2400}{2} = 1200$ символов / с;

минимальная ширина полосы $W = \frac{1}{2}(1 + r)R_s = \frac{1}{2}(2)(1200) = 1200$ Гц.

На рис. 3.19, *а* во временной области показан принятый узкополосный импульс в кодировке PAM; из выражения (3.79) получим функцию $h(t)$. На рис. 3.19, *б* показан Фурье-образ функции $h(t)$ — функция типа приподнятого косинуса. Отметим, что требуемая ширина полосы, W, находится в диапазоне от $f = 0$ до $f = 1/T$; она вдвое превышает теоретическую минимальную полосу по Найквисту.

б) Здесь, как и в п. а,

$R_s = 1200$ символов/с;

$W_{DSB} = (1 + r)R_s = 2(1200) = 2400$ Гц.

На рис. 3.20, *а* показан модулированный принятый импульс. Этот сигнал в кодировке PAM можно рассматривать как произведение высокочастотной синусоидальной несущей и сигнала с формой импульса, показанной на рис. 3.19, *а*. Односторонний спектральный график на рис. 3.20, *б* показывает модулированный сигнал, полоса которого выражается следующей формулой.

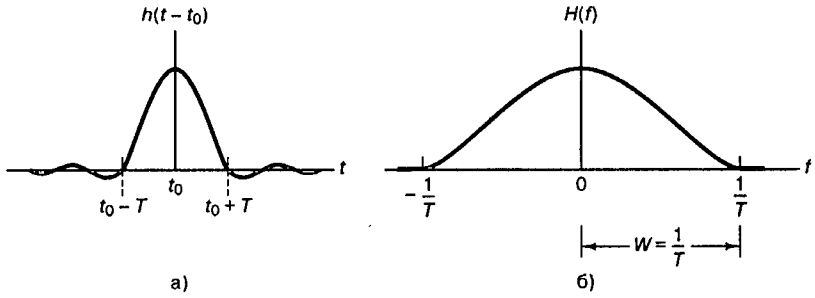


Рис. 3.19. Сформированный импульс и узкополосный спектр типа приподнятого косинуса

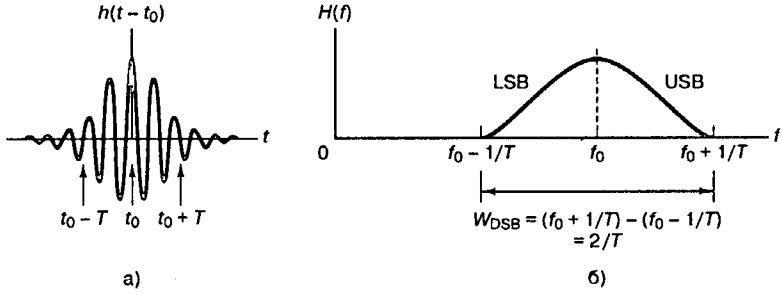


Рис. 3.20. Модулированный сформированный импульс и двухполосный модулированный спектр типа приподнятого косинуса

$$W_{\text{DSB}} = \left(f_0 + \frac{1}{T}\right) - \left(f_0 - \frac{1}{T}\right) = \frac{2}{T}$$

При смещении вверх по частоте спектра, показанного на рис. 3.19, *а*, смещаются отрицательная и положительная половины узкополосного спектра, таким образом требуемая полоса передачи дублируется. Как указывает название, двусторонний сигнал имеет две боковые полосы: верхнюю боковую полосу (upper sideband — USB), получаемую из положительной половины узкополосного сигнала, и нижнюю боковую полосу (lower sideband — LSB), получаемую из отрицательной половины.

Пример 3.4. Цифровые телефонные каналы

Сравните требования к ширине полосы системы для наземного аналогового телефонного канала передачи в речевом диапазоне (3 кГц) и цифрового канала. Для цифрового канала речь форматируется как поток битов в кодировке PCM с частотой дискретизации аналогового сигнала 8000 выборок/с. Каждая речевая выборка квантуется на один из 256 уровней. Затем поток битов передается с использованием сигналов PCM и принимается с нулевой межсимвольной интерференцией.

Решение

Процесс дискретизации и квантования дает слова PCM, каждое из которых представляет одну выборку и относится к одному из $L = 256$ различных уровней. Если каждая выборка передается как 256-уровневый импульс (символ) в кодировке PAM, то из формулы (3.82) получим ширину полосы (без межсимвольной интерференции), требуемую для передачи R_s символов/с.

$$W \geq \frac{R_s}{2} \text{ Гц}$$

Здесь равенство достигается только при использовании идеальной фильтрации Найквиста. Поскольку цифровая телефонная система использует (двоичные) сигналы PCM, каждое слово PCM преобразовывается в $l = \log_2 L = \log_2 256 = 8$ бит. Следовательно, полоса, необходимая для передачи речи с использованием PCM, равна следующему выражению.

$$W_{\text{PCM}} \geq (\log_2 L)\frac{R_s}{2}\ \text{Гц} \geq$$

$$\geq \frac{1}{2}(8\ \text{бит/символ})\ (8000\ \text{симолов/с}) = 32\ \text{кГц}$$

Описанный аналоговый канал передачи речи (3 кГц) обычно требует полосы порядка 4 кГц, включая некоторые разделительные полосы между каналами, называемые *защитными* (guard band). Следовательно, при использовании формата PCM, 8-битового квантования и двоичной передачи с сигналами PCM требуется примерно в 8 раз большая полоса, чем при использовании аналогового канала.

3.3.3. Демодуляция/обнаружение сформированных импульсов

3.3.3.1. Согласованные и обычные фильтры

Обычные фильтры отсекают нежелательные спектральные компоненты принятого сигнала при поддержании некоторой точности воспроизведения сигналов в выбранной области спектра, называемой *полосой пропускания* (pass-band). В общем случае эти фильтры разрабатываются для обеспечения приблизительно одинакового усиления, их характеристика дает линейное увеличение фазы в зависимости от частоты в пределах полосы пропускания и минимальное поглощение в остальной части спектра, именуемой *полосой заграждения* (stop-band). Согласованный фильтр имеет несколько иные "проектные приоритеты", направленные на максимизацию отношения SNR известного сигнала при шуме AWGN. В обычных фильтрах используются случайные сигналы, и результат фильтрации определяется только полосами сигналов, тогда как согласованные фильтры применяются с *известными сигналами*, имеющими произвольные параметры (такие, как амплитуда и время). Согласованный фильтр можно рассматривать как *шаблон*, который согласовывает обрабатываемый сигнал с известной формой. Обычный фильтр сохраняет временную или спектральную структуру сигнала. Согласованный фильтр, наоборот, в значительной степени модифицирует временную структуру путем сбора энергии сигнала, которая согласовывается с его шаблоном, и в завершение каждого интервала передачи символа представляет результат фильтрации в виде значения максимальной амплитуды. Вообще, в цифровой связи приемник обрабатывает поступающие сигналы с помощью фильтров обоих типов. Задачей обычного фильтра является изоляция и извлечение высокоточной аппроксимации сигнала с последующей передачей результата согласованному фильтру. Согласованный фильтр накапливает энергию принятого сигнала, и в момент взятия выборки ($t = T$) на выход фильтра подается напряжение, пропорциональное этой энергии, после чего следует обнаружение и дальнейшая обработка сигнала.

3.3.3.2. Импульсы Найквиста

Рассмотрим последовательность информационных импульсов на входе передатчика и последовательность импульсов, получаемую на выходе согласованного фильтра с характеристикой типа приподнятого косинуса (перед дискретизацией). На рис. 3.21 переданные данные представлены импульсными сигналами, которые появляются в мо-

менты времени τ_0, τ_1, ... Фильтрование приводит к расширению входящих сигналов, а следовательно, к запаздыванию их во времени. Время поступления импульсов обозначим t_0, t_1, ... Импульс, переданный в момент времени τ_0, поступает в приемник в момент времени t_0. Хвост, предшествующий основному лепестку демодулированного импульса, называется его *предтечей* (precursor). Для реальной системы с заданным системным эталонным временем принцип причинности дает условие $t_0 \geq \tau_0$, а разность времен $\tau_0 - t_0$ выражает задержку распространения в системе. В данном примере интервал времени от начала предтечи демодулированного импульса и до появления его главного лепестка или максимальной амплитуды равен $3T$ (утроенное время передачи импульса). Каждый выходящий импульс последовательности накладывается на другие импульсы; каждый импульс воздействует на основные лепестки трех предшествующих и трех последующих импульсов. В подобном случае, когда импульс фильтруется (формируется) так, что занимает более одного интервала передачи символа, определяется параметр, называемый *временем поддержки* (support time) импульса. Время поддержки — это количество интервалов передачи символа в течение длительности импульса. На рис. 3.21 время поддержки импульса равно 6 интервалам передачи символа (7 информационных точек с 6 интервалами между ними).

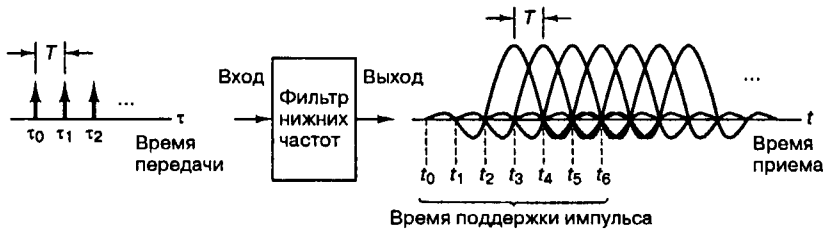


Рис. 3.21. Фильтрованная последовательность импульсов: выход и вход

На рис. 3.22, *а* показан импульсный отклик фильтра с характеристикой типа корня из приподнятого косинуса (максимальное значение нормированного фильтра равно единице, сглаживание фильтра $r = 0{,}5$), а на рис. 3.22, *б* изображен импульсный отклик фильтра с характеристикой типа приподнятого косинуса, называемый *импульсом Найквиста* (нормирование и значение коэффициента сглаживания такие же, как и на рис. 3.22, *а*). Изучая эти два импульса, можно заметить, что они очень похожи. Однако первый имеет несколько более быстрые переходы, а значит, его спектр (корень квадратный из приподнятого косинуса) не так быстро затухает, как спектр (приподнятый косинус) импульса Найквиста. Еще одним малозаметным, но важным отличием является то, что импульс Найквиста с характеристикой типа корня из приподнятого косинуса *не* дает нулевой межсимвольной интерференции (можно проверить, что хвосты импульса на рис. 3.22, *а* не проходят через точку нулевой амплитуды в моменты взятия выборок). В то же время, если фильтр с характеристикой типа корня из приподнятого косинуса используется и в передатчике, и в приемнике, произведение передаточных функций двух фильтров дает характеристику типа приподнятого косинуса, что означает нулевую межсимвольную интерференцию на выходе.

Было бы неплохо рассмотреть, как импульсы Найквиста с характеристикой типа корня из приподнятого косинуса выглядят на выходе передатчика и какую форму они имеют после демодуляции на согласованном фильтре, характеристика которого также представляет собой корень из приподнятого косинуса.

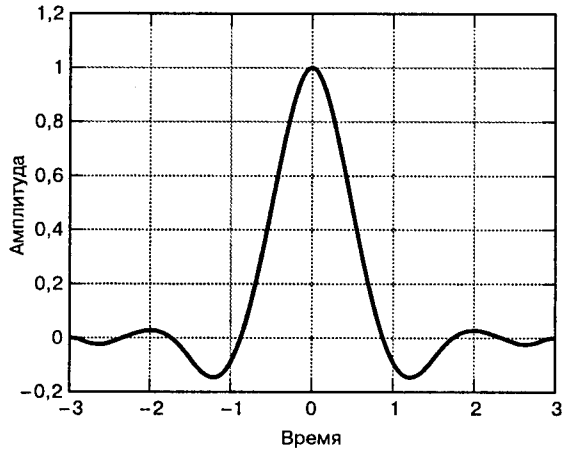


Рис. 3.22, а. Импульс Найквиста с характеристикой типа корня из приподнятого косинуса

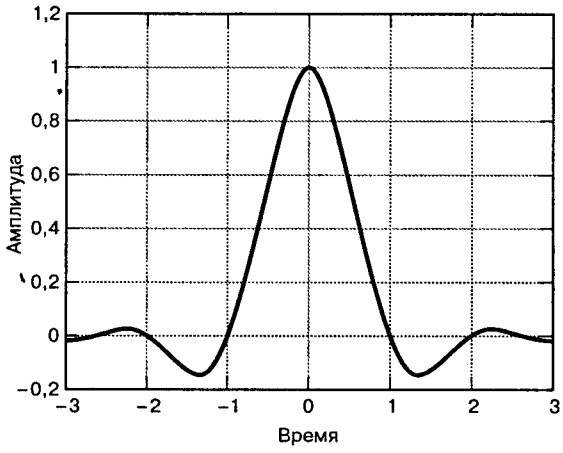


Рис. 3.22, б. Импульс Найквиста с характеристикой типа приподнятого косинуса

На рис. 3.23, *а* в качестве примера передачи приведена последовательность символов сообщения {+1 +1 −1 +3 +1 +3} из четверичного набора символов, где алфавит состоит из символов {±1, ±3}. Будем считать, что импульсы модулируются с помощью четверичной кодировки PAM, а их форма определяется фильтром с характеристикой типа корня из приподнятого косинуса с коэффициентом сглаживания $r = 0{,}5$. Аналоговый сигнал на рис. 3.23, *а* описывает выход передатчика. Сигнал на выходе (последовательность импульсов Найквиста, форма которых получена с выхода фильтра с характеристикой типа корня из приподнятого косинуса) запаздывает относительно сигнала на входе (показанного в виде импульсов), но для удобства визуального представления, чтобы читатель мог сравнить выход фильтра с его входом, оба сигнала изображены как одновременные. В действительности передается (или модулируется) только аналоговый сигнал.

На рис. 3.23, *б* показаны те же задержанные символы сообщения, а также сигнал с выхода согласованного фильтра с характеристикой типа корня из приподнятого

косинуса, что для всей системы в сумме дает передаточную функцию типа приподнятого косинуса.

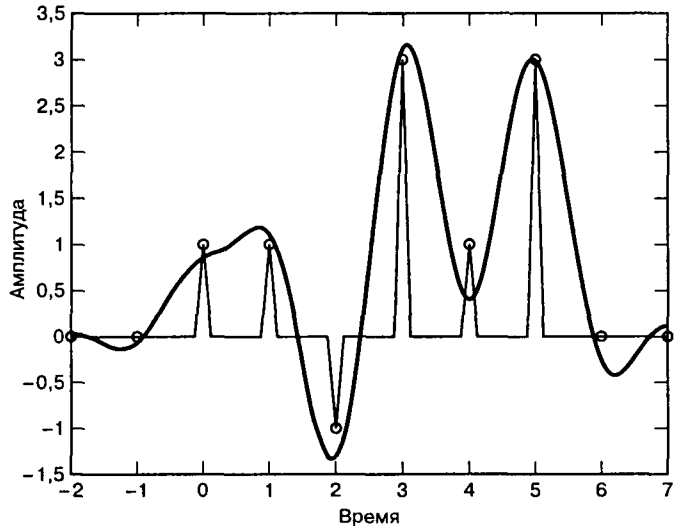


Рис. 3.23, а. М-уровневый сигнал Найквиста, пропущенный через фильтр с характеристикой типа корня из приподнятого косинуса, и входные дискретные значения, задержанные на некоторое время

Существует простой тест, позволяющий проверить, содержит ли фильтрованный сигнал с выхода межсимвольную интерференцию (предполагается отсутствие шума). Для этого требуется всего лишь произвести выборку фильтрованного сигнала в моменты времени, соответствующие исходным входящим выборкам; если полученные сигналы в результате выборки не отличаются от выборок исходного сообщения, то сигналы с выхода фильтра имеют нулевую межсимвольную интерференцию (в моменты взятия выборок). При сравнении рис. 3.23, *а* и 3.23, *б* на предмет межсимвольной интерференции видно, что дискретизация сигнала Найквиста на рис. 3.23, *а* (выход передатчика) *не дает* точных исходных выборок; в то же время дискретизация сигнала Найквиста на рис. 3.23, *б* (выход согласованного фильтра) *дает* точные исходные выборки. Это еще раз подтверждает, что фильтр Найквиста дает нулевую межсимвольную интерференцию в моменты взятия выборок, тогда как другие фильтры не имеют такой особенности.

3.4. Выравнивание

3.4.1. Характеристики канала

Многие каналы связи (например, телефонные или беспроводные) можно охарактеризовать как узкополосные линейные фильтры с импульсной характеристикой $h_c(t)$ и частотной характеристикой

$$H_c(f) = |H_c(f)|e^{i\theta_c(f)}, \tag{3.82}$$

где $h_c(t)$ и $H_c(f)$ — Фурье-образы друг друга, $|H_c(f)|$ — амплитудная характеристика канала, а $\theta_c(f)$ — фазовая характеристика канала.

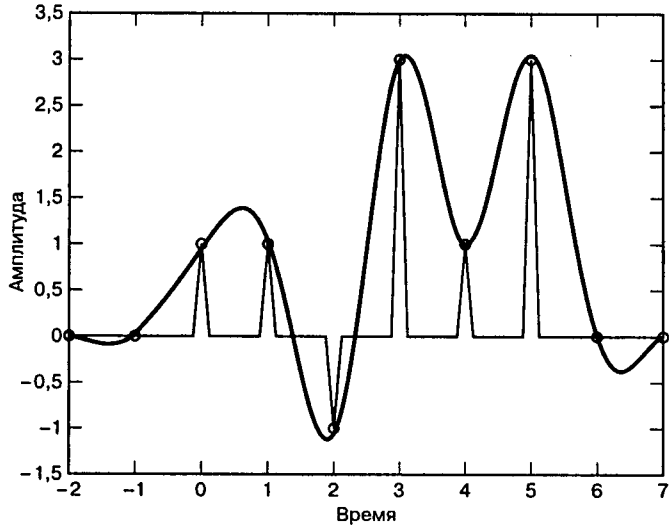


Рис. 3.23, б. Выход фильтра с характеристикой типа приподнятого косинуса и входные дискретные значения, задержанные на некоторое время

В разделе 1.6.3 было показано, что для получения идеальных (неискажающих) передающих характеристик канала в пределах полосы сигнала W, функция $|H_c(f)|$ *должна быть константой.* Кроме того, $\theta_c(f)$ *должна быть линейной функцией частоты*, что эквивалентно утверждению "запаздывание должно быть постоянным для всех спектральных компонентов сигнала". Если $|H_c(f)|$ не является константой в пределах полосы W, то канал будет искажать амплитуду сигнала. Если $\theta_c(f)$ не является линейной функцией частоты в пределах полосы W, канал будет искажать фазу. Во многих каналах, искажающих подобным образом информацию, например каналах с замираниями, искажение фазы и амплитуды обычно проявляется одновременно. При передаче последовательности импульсов подобное искажение проявляется в виде рассеивания или "размывания" импульсов, так что ни один импульс принятой демодулированной последовательности не определяется однозначно. В разделе 3.3 описывалось перекрытие импульсов, известное как *межсимвольная интерференция* (intersymbol interference — ISI). Это эффект, который проявляется в большинстве систем модуляции и является одной из основных помех надежной высокоскоростной передачи по узкополосным каналам. Совокупность методов обработки или фильтрации сигнала, направленных на устранение или снижение межсимвольной интерференции, именуется как "выравнивание" (equalization) и рассматривается в данном разделе.

На рис. 2.1 выравнивание разбито на две большие категории. Первая категория, *оценка последовательности с максимальным правдоподобием* (maximum-likelihood sequence estimation — MLSE), подразумевает измерение $h_c(t)$ с последующей подстройкой приемника под среду передачи. Цель такой подстройки — позволить детектору произвести точную оценку демодулированной искаженной последовательности импульсов. При использовании приемника MLSE искаженные выборки не изменяются и не проходят этап непосредственной компенсации последствий помех; вместо этого приемник перенастраивается так, чтобы максимально эффективно работать с искаженными выборками. (Пример этого метода, известный как выравнивание Витерби, рассмотрен в разделе 15.7.1.) Вторая категория, *вы-*

равнивание с помощью фильтров, включает использование фильтров для компенсации искажения импульсов. В этом случае детектору предоставляется последовательность демодулированных выборок, модифицированных или "очищенных" эквалайзером от последствий ISI. Выравнивание с помощью фильтров (более популярный подход из двух описанных выше) также имеет несколько подтипов. Фильтры могут быть линейными устройствами, содержащими только элементы с прямой связью (*трансверсальные эквалайзеры*), или нелинейными, включающими элементы с обратной связью (*эквалайзеры с решающей обратной связью*). Кроме того, фильтры могут различаться алгоритмом работы, который может быть *заданным* или *адаптивным*. Также они могут различаться разрешением или частотой обновления. Если выборки производятся только в пределах символа, т.е. одна выборка на символ, то это *символьное разделение*. Если каждому символу соответствует несколько выборок, то это *фракционное разделение*.

Модифицируем уравнение (3.77), заменив принимающий/выравнивающий фильтр отдельными (принимающим и выравнивающим) фильтрами, определяемыми частотными передаточными функциями $H_r(f)$ и $H_e(f)$. Будем также считать, что общая передаточная функция системы $H(f)$ имеет вид приподнятого косинуса (raised-cosine), и обозначим ее $H_{RC}(f)$. Таким образом, можем записать следующее.

$$H_{RC}(f) = H_t(f)\, H_c(f)\, H_r(f)\, H_e(f) \tag{3.83}$$

В системах, представляющих практический интерес, частотная передаточная функция системы $H_c(f)$ и ее импульсная характеристика $h_c(t)$ не известны с точностью, достаточной для разработки приемника, который в любой момент времени дает нулевую межсимвольную интерференцию. Передающий и принимающий фильтры, как правило, выбираются так, чтобы

$$H_{RC}(f) = H_t(f)\, H_r(f) \tag{3.84}$$

Таким образом, характеристики $H_t(f)$ и $H_r(t)$ имеют вид корней из приподнятого косинуса. Следовательно, передаточная функция эквалайзера, необходимая для компенсации искажения, внесенного каналом, является обратной передаточной функции канала.

$$H_e(f) = \frac{1}{H_c(f)} = \frac{1}{|H_c(f)|} e^{-i\theta_c(f)}. \tag{3.85}$$

Иногда частотная передаточная функция системы допускает межсимвольную интерференцию в специально выбранных точках дискретизации (например, передаточная функция гауссового фильтра). Такие передаточные функции позволяют повысить эффективность использования полосы, по сравнению с фильтром с характеристикой типа приподнятого косинуса. При выборе такого конструкторского решения выравнивающий фильтр должен компенсировать не только внесенную каналом межсимвольную интерференцию, но и ISI, внесенную передающим и принимающим фильтрами [7].

3.4.2. Глазковая диаграмма

Глазковая диаграмма — это изображение, полученное в результате измерения отклика системы на заданные узкополосные сигналы. На вертикальные пластины осциллографа подается отклик приемника на случайную последовательность импульса, а на горизонтальные — пилообразный сигнал сигнальной частоты. Другими словами, горизонтальная временная развертка осциллографа устанавливается равной длительности символа (импульса). В течение каждого сигнального промежутка очередной сигнал накладывается

на семейство кривых в интервале (0, T). На рис. 3.24 приведена глазковая диаграмма, получаемая при двоичной антиподной (биполярные импульсы) передаче сигналов. Поскольку символы поступают из случайного источника, они могут быть как положительными, так и отрицательными, и отображение послесвечения электронного луча позволяет видеть изображение, имеющее форму глаза. Ширина открытия глаза указывает время, в течение которого должна быть произведена выборка сигнала. Разумеется, оптимальное время взятия выборки соответствует максимально распахнутому глазу, что дает максимальную защиту от воздействия помех. Если в системе не используется фильтрация, т.е. если передаваемым информационным импульсам соответствует бесконечная полоса, то отклик системы дает импульсы идеальной прямоугольной формы. В этом случае диаграмма будет выглядеть уже не как глаз, а как прямоугольник. Диапазон разностей амплитуд, обозначенный через D_A, является мерой искажения, вызванного межсимвольной интерференцией, а диапазон разностей времен перехода через нуль, обозначенный через J_T, есть мерой неустойчивой синхронизации. На рисунке также показана мера запаса помехоустойчивости M_N и чувствительность к ошибкам синхронизации S_T. Чаще всего глазковая диаграмма используется для качественной оценки степени межсимвольной интерференции. По мере закрытия глаза ISI увеличивается, а по мере открытия — уменьшается.

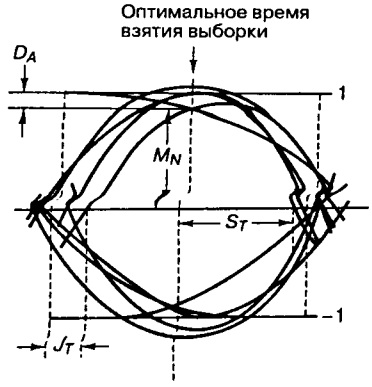


Рис. 3.24. Глазковая диаграмма

3.4.3. Типы эквалайзеров

3.4.3.1. Трансверсальный эквалайзер

В качестве тестовой последовательности, используемой для выравнивания, часто выбирается шумоподобная последовательность с широкополосным спектром, с помощью которой оценивается отклик канала. В простейшем смысле настройка может заключаться в передаче простого короткого импульса (приблизительно, идеального импульса) с последующим изучением импульсного отклика канала. На практике в качестве тестовой последовательности предпочтителен не единичный импульс, а псевдошумовой сигнал, поскольку последний имеет большую среднюю мощность, а значит, большее отношение сигнал/шум при одинаковых максимальных переданных мощностях. Для изучения трансверсального фильтра предположим, что через систему был передан единственный импульс, причем система спроектирована таким образом, что общая передаточная функция имеет вид приподнятого косинуса $H_{RC} = H_t(f)\, H_r(f)$. Также будем считать, что канал вводит межсимвольную интерференцию, так что принятый демодулированный импульс искажается,

как показано на рис. 3.25, поэтому боковые лепестки, ближайшие к главному лепестку импульса, не проходят через нуль в моменты взятия выборок. Искажение можно рассматривать как положительное или отрицательное отражение, появляющееся до и после главного лепестка. Для получения желаемой передаточной функции с характеристикой типа приподнятого косинуса выравнивающий фильтр, как следует из уравнения (3.85), должен иметь частотный отклик $H_e(f)$, тогда отклик канала при умножении на $H_e(f)$ будет $H_{RC}(f)$. Другими словами, мы хотим, чтобы выравнивающий фильтр вырабатывал набор подавляющих отражений. Поскольку нас интересуют выборки выровненного сигнала только в определенные моменты времени, проектирование подобного выравнивающего фильтра может быть довольно простой задачей.

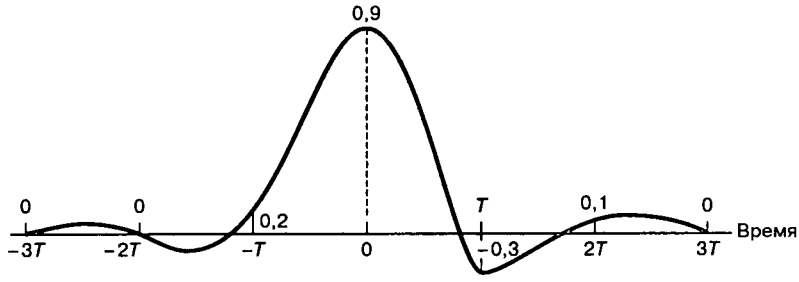


Рис. 3.25. Принятый искаженный импульс

Трансверсальный фильтр, изображенный на рис. 3.26, — это наиболее популярная форма легко настраиваемого выравнивающего фильтра, состоящего из канала задержки с отводами задержки на T секунд (где T — длительность символа). В подобном эквалайзере текущее и предыдущее значения принятого сигнала линейно взвешиваются коэффициентами эквалайзера или весовыми коэффициентами отводов $\{c_n\}$, а затем суммируются для формирования выхода. Основной вклад вносит центральный отвод; вклады остальных отводов связаны с отражениями основного сигнала в течение последующих (и предыдущих) интервалов T. Если бы можно было создать фильтр с бесконечным числом отводов, можно было бы так подобрать весовые коэффициенты, чтобы импульсный отклик системы равнялся всегда нулю, за исключением моментов взятия выборок; таким образом $H_e(f)$ была бы точно равна обратной передаточной функции канала в формуле (3.85). Несмотря на то что фильтр с бесконечным числом отводов не относится к числу реализуемых, все же можно создать фильтр, достаточно хорошо аппроксимирующий идеальный случай.

На рис. 3.26 выходы взвешенных отводов усиливаются, суммируются и подаются на устройство принятия решения. Весовые коэффициенты отводов $\{c_n\}$ должны выбираться так, чтобы вычитать эффекты интерференции из символов, соседствующих во времени с искомым символом. Предположим, что существует $(2N + 1)$ отводов с весовыми коэффициентами $c_{-N}, c_{-N+1}, \ldots, c_N$. Выборки на выходе эквалайзера $\{z(k)\}$ находятся путем следующей свертки выборок на входе $\{x(k)\}$ и весовых коэффициентов $\{c_n\}$.

$$z(k) = \sum_{n=-N}^{N} x(k-n)c_n \quad k = -2N, \ldots, 2N \quad n = -N, \ldots, N, \tag{3.86}$$

где $k = 0, \pm 1, \pm 2, \ldots$ — временные коэффициенты, показанные в круглых скобках. (Время может быть как положительным, так и отрицательным.)

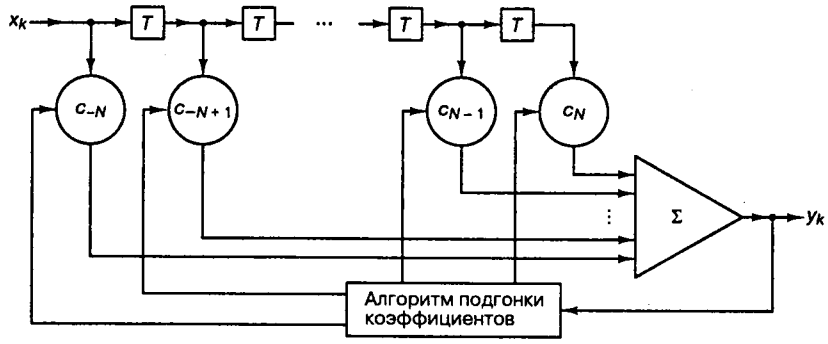


Рис. 3.26. Трансверсальный фильтр

Коэффициент n используется для обозначения смещения во времени и как идентификатор коэффициентов фильтра (адрес фильтра). В последнем случае n показан как индекс. Если ввести векторы $\mathbf{z}$ и $\mathbf{c}$ и матрицу $\mathbf{x}$

$$\mathbf{z} = \begin{bmatrix} z(-2N) \\ \vdots \\ z(0) \\ \vdots \\ z(2N) \end{bmatrix} \qquad \mathbf{c} = \begin{bmatrix} c_{-N} \\ \vdots \\ c_0 \\ \vdots \\ c_N \end{bmatrix} \tag{3.87}$$

и

$$\mathbf{x} = \begin{bmatrix} x(-N) & 0 & 0 & \dots & 0 & 0 \\ x(-N+1) & x(-N) & 0 & \dots & \dots & \dots \\ \vdots & & & \vdots & & \vdots \\ x(N) & x(N-1) & x(N-2) & \dots & x(-N+1) & x(-N) \\ \vdots & & & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & x(N) & x(N-1) \\ 0 & 0 & 0 & \dots & 0 & x(N) \end{bmatrix} \tag{3.88}$$

то соотношение между $\{z(k)\}$, $\{x(k)\}$ и $\{c_n\}$ можно записать в более компактной форме.

$$\mathbf{z} = \mathbf{x}\,\mathbf{c} \tag{3.89,а}$$

Если матрица $\mathbf{x}$ является квадратной, а число строк и столбцов соответствует числу элементов вектора $\mathbf{c}$, то $\mathbf{c}$ можно выразить в следующем виде.

$$\mathbf{C} = \mathbf{x}^{-1}\,\mathbf{z} \tag{3.89,б}$$

Отметим, что в общем случае размер вектора $\mathbf{z}$ и число строк матрицы $\mathbf{x}$ могут быть любыми, поскольку нас может интересовать межсимвольная интерференция в точках взятия выборок, достаточно удаленных от основного лепестка рассматриваемого импульса. В формулах (3.86)–(3.88) индекс k выбирался так, чтобы число точек взятия выборок равнялось $4N+1$. Векторы $\mathbf{z}$ и $\mathbf{c}$ имеют размерность $4N+1$ и $2N+1$, соответственно, а матрица $\mathbf{x}$ не является квадратной и имеет размер $4N+1$ на $2N+1$. В этом случае система уравне-

ний (3.89,а) называется переопределенной (т.е. число уравнений превышает число неизвестных). Решать подобные уравнения можно с помощью детерминистского способа — *метода обращения в нуль незначащих коэффициентов* или статистического — *метода решения с минимальной среднеквадратической ошибкой* (mean-square error — MSE).

Обращение в нуль незначащих коэффициентов

Это решение начинается с отделения N верхних и N нижних строк матрицы $\mathbf{x}$ в уравнении (3.88). Таким образом, матрица $\mathbf{x}$ становится квадратной размером $2N+1$ на $2N+1$, вектор $\mathbf{z}$ также имеет теперь размер $2N+1$, а формула (3.89,а) определяет детерминированную систему $2N+1$ уравнений. Предлагаемое решение минимизирует максимальное искажение, вызванное межсимвольной интерференцией, путем выбора весовых коэффициентов $\{c_n\}$ таким образом, чтобы сигнал на выходе эквалайзера был равен нулю в N точках взятия выборок по обе стороны от искомого импульса. Другими словами, весовые коэффициенты выбираются так, чтобы

$$z(k)=\begin{cases}1 & \text{для } k=0\\ 0 & \text{для } k=\pm 1,\pm 2,\ldots,\pm N\end{cases} \tag{3.90}$$

Для нахождения $2N+1$ весовых коэффициентов $\{c_n\}$ из системы $2N+1$ уравнений используется выражение (3.90). Требуемая длина фильтра (число отводов) зависит от того, насколько сильно канал может "размазать" импульс. Для эквалайзера конечного размера максимальное искажение гарантированно будет минимизировано только в том случае, если глазковая диаграмма изначально имеет вид открытого глаза. В то же время при высокоскоростной передаче и в каналах, вводящих значительную межсимвольную интерференцию, до выравнивания глаз всегда закрыт [8]. Кроме того, эквалайзер, использующий метод обращения в нуль незначащих коэффициентов, не учитывает воздействие шума, поэтому такое решение не всегда является оптимальным.

Пример 3.5. Трехотводный эквалайзер, использующий метод обращения в нуль незначащих коэффициентов

Путем передачи единственного импульса или настроечного сигнала требуется определить весовые коэффициенты отводов выравнивающего трансверсального фильтра. Выравнивающий канал, изображенный на рис. 3.26, состоит всего из трех отводов. Пусть принят искаженный набор выборок импульса $\{x(k)\}$ со значениями напряжения 0,0; 0,2; 0,9; –0,3; 0,1, как показано на рис. 3.25. Используйте метод обращения в нуль незначащих коэффициентов для нахождения коэффициентов $\{c_{-1}, c_0, c_1\}$, уменьшающих межсимвольную интерференцию так, чтобы выборки импульса после выравнивания имели значения $\{z(-1)=0, z(0)=1, z(1)=0\}$. Используя эти весовые коэффициенты, вычислите значения выборок выровненного импульса в моменты $k=\pm 2, \pm 3$. Чему равен вклад наибольшей амплитуды в межсимвольную интерференцию и чему равна сумма амплитуд всех вкладов?

Решение

При заданном импульсном отклике канала из формулы (3.89) получим следующее.

$$\mathbf{z}=\mathbf{x}\,\mathbf{c}$$

или

$$\begin{bmatrix}0\\1\\0\end{bmatrix}=\begin{bmatrix}x(0) & x(-1) & x(-2)\\ x(1) & x(0) & x(-1)\\ x(2) & x(1) & x(0)\end{bmatrix}\begin{bmatrix}c_{-1}\\ c_0\\ c_1\end{bmatrix}$$

$$= \begin{bmatrix} 0,9 & 0,2 & 0 \\ -0,3 & 0,9 & 0,2 \\ 0,1 & -0,3 & 0,9 \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \end{bmatrix}$$

Решая систему трех уравнений, получаем следующие значения весовых коэффициентов.

$$\begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} -0,2140 \\ 0,9631 \\ 0.3448 \end{bmatrix}$$

Значения выравненных выборок импульса $\{z(k)\}$, соответствующих временам взятия выборок $k = -3, -2, -1, 0, 1, 2, 3$, вычисляются с помощью формулы (3.89,а).

0,0000; –0,0428; 0,0000; 1,0000; 0,0000; –0,0071; 0,0345

Вклад наибольшей амплитуды в межсимвольную интерференцию равен 0,0428, а сумма амплитуд всех вкладов равна 0,0844. Очевидно, что эквалайзер с тремя отводами дает нулевое значение выровненного импульса в точках взятия выборки, соседствующих с основным лепестком. Если создать эквалайзер большего размера, он будет давать нулевое значение в большем числе точек взятия выборок.

Решение с минимальной среднеквадратической ошибкой

Более устойчивый эквалайзер можно получить, выбрав весовые коэффициенты $\{c_n\}$, минимизирующие среднеквадратическую ошибку (mean-square error — MSE) всех членов, вносящих вклад в межсимвольную интерференцию, плюс мощности шума на выходе эквалайзера [9]. Среднеквадратическая ошибка определяется как математическое ожидание квадрата разности желаемого и обнаруженного информационных символов. Для получения решения с минимальной MSE можно использовать переопределенную систему уравнений (3.89,а), умножив обе ее части на $\mathbf{x}^T$, что дает [10]

$$\mathbf{x}^T\mathbf{z} = \mathbf{x}^T\mathbf{x}\mathbf{c} \tag{3.91,а}$$

и

$$\mathbf{R}_{xz} = \mathbf{R}_{xx}\mathbf{c}, \tag{3.91,б}$$

где $\mathbf{R}_{xz} = \mathbf{x}^T\mathbf{z}$ является *вектором взаимной корреляции*, а $\mathbf{R}_{xx} = \mathbf{x}^T\mathbf{c}$ — *автокорреляционной* матрицей входного шумового сигнала. На практике $\mathbf{R}_{xz}$ и $\mathbf{R}_{xx}$ *априори* неизвестны, но могут быть вычислены приблизительно путем передачи через канал тестового сигнала и использования усреднения по времени для нахождения весовых коэффициентов из уравнения (3.91).

$$\mathbf{c} = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xz} \tag{3.92}$$

При детерминистском решении метода обращения в нуль незначащих коэффициентов матрица $\mathbf{x}$ должна быть квадратной. Но для получения (статистического) решения с минимальной MSE начинать следует с переопределенной системы уравнений, а значит, *неквадратной* матрицы $\mathbf{x}$, которая впоследствии преобразовывается в *квадратную* автокорреляционную матрицу $\mathbf{R}_{xx} = \mathbf{x}^T\mathbf{x}$, порождающую систему $2N + 1$ уравнений, решение которой дает значения весовых коэффициентов, минимизирующих MSE. Размер вектора $\mathbf{c}$ и число столбцов матрицы $\mathbf{x}$ соответствуют числу отводов выравнивающего фильтра. Большинство высокоскоростных модемов для выбора весовых коэффи-

циентов используют критерий MSE, поскольку он лучше равновесного; он является более устойчивым при наличии шумов и большой межсимвольной интерференции [8].

Пример 3.6. Семиотводный эквалайзер с минимальной среднеквадратической ошибкой

Путем передачи единственного импульса или настроечного сигнала требуется определить весовые коэффициенты отводов выравнивающего трансверсального фильтра. Выравнивающий канал, изображенный на рис. 3.26, состоит из семи отводов. Пусть принят искаженный набор выборок импульса $\{x(k)\}$ со значениями напряжения 0,0108; –0,0558; 0,1617; 1,0000; –0,1749; 0,0227; 0,0110. Используйте решение с минимальной среднеквадратической ошибкой для нахождения весовых коэффициентов $\{c_n\}$, минимизирующих межсимвольную интерференцию. Используя эти весовые коэффициенты, вычислите значения выборок выровненного импульса в моменты $\{k = 0, \pm1, \pm2, \pm3, \ldots, \pm6\}$. Чему равен вклад наибольшей амплитуды в межсимвольную интерференцию и чему равна сумма амплитуд всех вкладов?

Решение

С помощью формулы (3.93) для семиотводного фильтра ($N = 3$), можно записать матрицу $\mathbf{x}$ размером $4N + 1$ на $2N + 1 = 13 \times 7$.

$$\mathbf{x} = \begin{bmatrix} 0{,}0110 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0{,}0227 & 0{,}0110 & 0 & 0 & 0 & 0 & 0 \\ -0{,}1749 & 0{,}0227 & 0{,}0110 & 0 & 0 & 0 & 0 \\ 1{,}0000 & -0{,}1749 & 0{,}0227 & 0{,}0110 & 0 & 0 & 0 \\ 0{,}1617 & 1{,}0000 & -0{,}1749 & 0{,}0227 & 0{,}0110 & 0 & 0 \\ -0{,}0558 & 0{,}1617 & 1{,}0000 & -0{,}1749 & 0{,}0227 & 0{,}0110 & 0 \\ 0{,}0108 & -0{,}0558 & 0{,}1617 & 1{,}0000 & -0{,}1749 & 0{,}0227 & 0{,}0110 \\ 0 & 0{,}0108 & -0{,}0558 & 0{,}1617 & 1{,}0000 & -0{,}1749 & 0{,}0227 \\ 0 & 0 & 0{,}0108 & -0{,}0558 & 0{,}1617 & 1{,}0000 & -0{,}1749 \\ 0 & 0 & 0 & 0{,}0108 & -0{,}0558 & 0{,}1617 & 1{,}0000 \\ 0 & 0 & 0 & 0 & 0{,}0108 & -0{,}0558 & 0{,}1617 \\ 0 & 0 & 0 & 0 & 0 & 0{,}0108 & -0{,}0558 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0{,}0108 \end{bmatrix}$$

Используя матрицу $\mathbf{x}$, можно получить автокорреляционную матрицу $\mathbf{R}_{xx}$ и вектор взаимной корреляции $\mathbf{R}_{xz}$, определенные формулами (3.91). С помощью компьютера матрица $\mathbf{R}_{xx}$ обращается, выполняется умножение матриц (см. формулу (3.92)), в результате чего получаются следующие весовые коэффициенты $\{c_{-3}, c_{-2}, c_{-1}, c_0, c_1, c_2, c_3\}$.

$$-0{,}0116;\ 0{,}0108;\ 0{,}1659;\ 0{,}9495;\ -0{,}1318;\ 0{,}0670;\ -0{,}0269$$

Подставляя эти весовые коэффициенты в систему уравнений (3.89,а), находим 13 выравненных выборок $\{z(k)\}$ в моменты времени $k = -6, -5, \ldots, 5, 6$.

$$-0{,}0001;\ -0{,}0001;\ 0{,}0041;\ 0{,}0007;\ 0{,}0000;\ -0{,}0000;\ 1{,}0000;$$

$$0{,}0003;\ -0{,}0007;\ 0{,}0015;\ -0{,}0095;\ 0{,}0022;\ -0{,}0003$$

Вклад наибольшей амплитуды в межсимвольную интерференцию равен 0,0095, а сумма амплитуд всех вкладов равна 0,0195.

3.4.3.2. Эквалайзер с решающей обратной связью

Основное ограничение линейного эквалайзера, такого как трансверсальный фильтр, заключается в плохой производительности в каналах, имеющих спектральные нули [11]. Подобные каналы часто встречаются в приложениях мобильной радиосвязи. Эквалайзер с решающей обратной связью (decision feedback equalizer — DFE) — это нелинейное устройство, использующее предыдущее решение детектора для устранения межсимвольной интерференции из импульсов, демодулируемых в настоящий момент. Поскольку причиной интерференции являются хвосты предыдущих импульсов, по сути, из текущего импульса вычитается искажение, вызванное предыдущими импульсами.

На рис. 3.27 в виде блочной диаграммы изображен эквалайзер DFE, причем направляющий фильтр и фильтр обратной связи могут быть линейными; например, это может быть трансверсальный фильтр. На рисунке также показано адаптивное обновление весовых коэффициентов фильтра (см. следующий раздел). Нелинейность DFE вытекает из нелинейной характеристики детектора, обеспечивающего подачу сигнала на вход фильтра обратной связи. В основе работы DFE лежит следующее: если значения ранее обнаруженных символов известны (предыдущее решение предполагается точным), то межсимвольную интерференцию, внесенную символами, можно точно уравновесить на выходе направляющего фильтра путем вычитания значений предыдущих символов с соответствующими весовыми коэффициентами. Для удовлетворения выбранного критерия (например, минимальности среднеквадратической ошибки) весовые коэффициенты направляющего отвода и отвода обратной связи могут подгоняться одновременно.

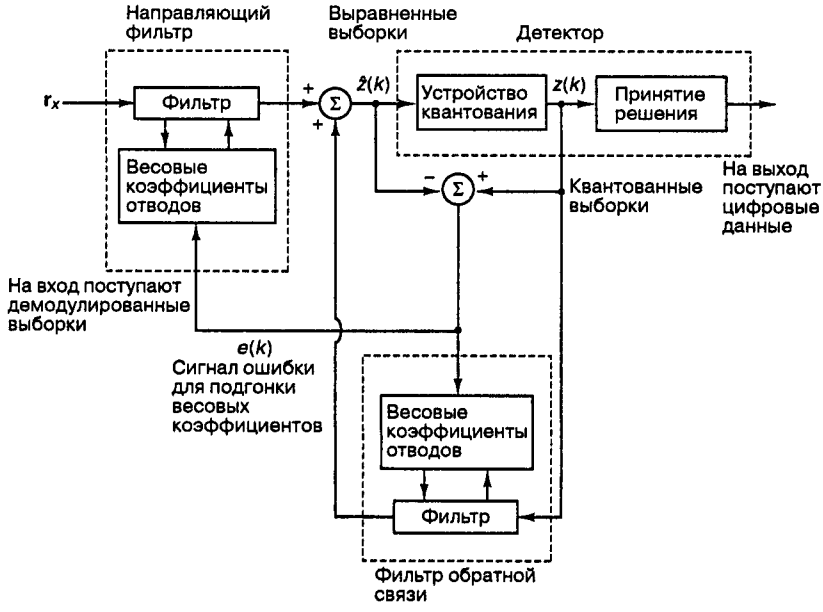


Рис. 3.27. Эквалайзер с решающей обратной связью

При использовании единственного направляющего фильтра выход содержит шум канала, внесенный каждой выборкой, произведенной в фильтре. Преимуществом реализации DFE является то, что фильтр обратной связи не только используется для удаления межсимвольной интерференции, но и работает на бесшумных уровнях квантования, а значит на его выходе отсутствует шум канала.

3.4.4. Заданное и адаптивное выравнивание

В инвариантных относительно времени каналах с известными частотными характеристиками, характеристики канала могут измеряться, и, соответственно, могут подгоняться значения весовых коэффициентов отводов. Если весовые коэффициенты остаются фиксированными в течение всего процесса передачи данных, выравнивание называется *заданным* (preset); простой метод заданного выравнивания заключается в установке весовых коэффициентов $\{c_n\}$, согласно некоторым усредненным знаниям о канале. Такой метод использовался для передачи информации по телефонным каналам со скоростью, не превышающей 2400 бит/с. Еще один метод заданного выравнивания состоит в передаче настроечной последовательности, которая в приемнике сравнивалась с последовательностью, сгенерированной локально. Отличия последовательностей позволяют установить весовые коэффициенты $\{c_n\}$. Важным моментом использования любой разновидности заданного выравнивания является то, что установка параметров производится либо единожды, либо в исключительно редких случаях (например, при прерывании передачи и необходимости ее повторной настройки).

Тип выравнивания, способный отслеживать постепенные изменения, называется *адаптивным* (adaptive). Его реализация может включать периодическую или непрерывную "подборку" весовых коэффициентов отводов. Периодическая корректировка выполняется путем периодической передачи начальной комбинации битов или краткой настроечной последовательности, заранее известной приемнику. Кроме того, начальная комбинация битов используется приемником для определения начала передачи, установки уровня автоматической регулировки усиления (automatic gain control — AGC) и для согласования с принятым сигналом внутренних часов и местных гетеродинов. Непрерывная подстройка осуществляется посредством замещения известной тестовой последовательности набором информационных символов, которые получены на выходе эквалайзера и считаются известными данными. При непрерывной и автоматической (наиболее распространенный подход) настройке используется метод, *управляемый решением* (decision directed) [11]. Название метода не стоит путать с DFE — эквалайзером с решающей обратной связью. Управление решением связано только со способом юстировки (с помощью сигнала от детектора) весовых коэффициентов отводов фильтра. Эквалайзер DFE — это наличие дополнительного фильтра на выходе детектора, рекурсивным образом возвращающего сигнал на вход детектора. Следовательно, при использовании DFE существует два фильтра (направляющий и фильтр обратной связи), обрабатывающие данные для снижения межсимвольной интерференции.

Недостатком заданного выравнивания является то, что оно требует предварительной настройки в начале каждой новой передачи. Кроме того, нестационарные каналы, вследствие межсимвольной интерференции и фиксированных весовых коэффициентах отводов, могут приводить к ухудшению производительности системы. Адаптивное выравнивание, в частности адаптивное выравнивание, *управляемое решением*, успешно устраняет межсимвольную интерференцию, если первоначальная вероятность ошибки не превышает один процент (эмпирическое правило). Если вероятность ошибки превышает один процент, эквалайзер, управляемый решением, может и не дать требуемого результата. Общее решение этой проблемы — инициализировать эквалайзер с альтернативным процессом, (таким, как передача начальной комбинации битов), что позволит обеспечить низкую вероятность ошибки в канале, а затем переключиться в режим управления решением. Чтобы избежать служебных издержек, вносимых начальной комбинацией битов, проекты многих систем предусматривают работу в режиме непрерывного широковещания с использованием для первоначальной оценки канала алгоритмов *слепого выравнивания* (blind equalization). Эти

алгоритмы согласовывают коэффициенты фильтра со статистикой выборок, а не с решениями относительно значений выборок [11].

Для оценки оптимальных коэффициентов автоматические эквалайзеры используют итеративные методы. Система уравнений, приведенная в выражении (3.93), не учитывает воздействие шума канала. При получении устойчивого решения для значений весовых коэффициентов фильтра, требуется усреднять либо данные для устойчивой статистики сигнала, либо зашумленное решение, полученное из зашумленных данных. Сложность алгоритма и проблемы численной устойчивости часто приводит к разработке алгоритмов, усредняющих зашумленные решения. Наиболее надежным из этого класса алгоритмов является алгоритм минимальной среднеквадратической (least-mean-square — LMS) ошибки. Каждая итерация этого алгоритма использует зашумленную оценку *градиента* ошибок для регулировки весовых коэффициентов относительно снижения среднеквадратической ошибки. Градиент шума — это просто произведение $e(k)\,\mathbf{r}_x$ скалярного значения ошибки $e(k)$ и вектора данных $\mathbf{r}_x$. Вектор $\mathbf{r}_x$ — это вектор выборок канала, которые подверглись воздействию шума и в момент k находились на выравнивающем фильтре. Выше использовалось следующее математическое представление: передавался импульс, и выравнивающий фильтр работал с последовательностью выборок (вектором), представляющей импульсный отклик канала. Эти принятые выборки (в виде сдвига во времени) изображались как матрица **x**. Теперь, вместо использования отклика на импульс, предполагается передача данных на вход фильтра (рис. 3.27), соответственно определяется вектор принятых выборок $\mathbf{r}_x$, представляющий информационный отклик канала. Ошибка записывается как разность желаемого сигнала и сигнала, полученного на выходе фильтра.

$$e(k) = z(k) - \hat{z}(k) \tag{3.93}$$

Здесь $z(k)$ — желаемый выходной сигнал (выборка без межсимвольной интерференции), а $\hat{z}(k)$ — оценка $z(k)$ в момент времени k (производится в устройстве квантования, показанном на рис. 3.27), имеющая следующий вид.

$$\hat{z}(k) = \mathbf{c}^T\mathbf{r}_x = \sum_{n=-N}^{N} x(k-n)c_n \tag{3.94}$$

В формуле (3.94) суммирование представляет свертку входящих информационных выборок с весовыми коэффициентами отводов $\{c_n\}$, где c_n — коэффициент n-го отвода в момент времени k, а $\mathbf{c}^T$ — транспонированный вектор весовых коэффициентов в момент времени k. Далее будет показано, что итеративный процесс, обновляющий значения весовых коэффициентов в каждый момент времени k, имеет следующий вид.

$$\mathbf{c}(k+1) = \mathbf{c}(k) + \Delta e(k)\mathbf{r}_x \tag{3.95}$$

Здесь $\mathbf{c}(k)$ — вектор весовых коэффициентов фильтра в момент времени k, а Δ — малый член, ограничивающий шаг коэффициентов, а значит, контролирующий скорость сходимости алгоритма и дисперсию устойчивого решения. Это простое соотношение является следствием принципа ортогональности, утверждающего, что ошибка, сопровождающая оптимальное решение, ортогональна обрабатываемым данным. Поскольку алгоритм рекурсивен (по отношению к весовым коэффициентам), необходимо следить за его устойчивостью. Устойчивость гарантируется, если параметр Δ меньше значения обратной энергии данных в фильтре. Если алгоритм является устойчивым, он в среднем сходится к оптимальному решению, при этом его дисперсия пропорциональна параметру Δ. Таким обра-

зом, желательно, чтобы параметр сходимости Δ был больше (для более быстрой сходимости), но не настолько, чтобы привести к неустойчивости, хотя, с другой стороны, малый параметр Δ обеспечивает малую дисперсию. Обычно для получения низкодисперсного устойчивого решения Δ выбирается равным фиксированной небольшой величине [12]. Существуют схемы [13], позволяющие Δ меняться от больших значений к меньшим в процессе получения устойчивого решения.

Отметим, что уравнения (3.93)–(3.95) приведены в контексте вещественных сигналов. Если используется квадратурная реализация, так что сигнал описывается вещественной и мнимой (или синфазной и квадратурной) упорядоченными парами, то каждый канал на рис. 3.27 в действительности состоит из двух каналов, и уравнения (3.93)–(3.95) необходимо записывать в комплексной форме. (Квадратурная реализация подробно рассмотрена в разделах 4.2.1 и 4.6.)

3.4.5. Частота обновления фильтра

Выравнивающие фильтры классифицируются по частоте дискретизации входящего сигнала. Трансверсальный фильтр с отводами, размещенными через T секунд, где T — длительность передачи символа, называется *эквалайзером с символьным разделением* (symbol-spaced equalizer). Процесс дискретизации выхода эквалайзера с частотой $1/T$ приводит к наложению, если полоса сигнала не ограничена строго величиной $1/T$ Гц, т.е. спектральные компоненты сигнала, не разделенные промежутком $1/T$ Гц, накладываются. Наложенная версия сигнала может давать спектральные нули [8]. Частота обновления фильтра, превышающая скорость передачи символов, помогает смягчить эту проблему. Эквалайзеры, использующие подобный метод, называются *эквалайзерами с фракционным разделением* (fractionally-spaced equalizer). В таких устройствах отводы фильтра разделены промежутками

$$T' \le \frac{T}{(1+r)} \text{ секунд}, \tag{3.96}$$

где через r обозначен избыток полосы. Другими словами, ширина принятого сигнала равна следующему.

$$W \le \frac{(1+r)}{T} \tag{3.97}$$

T' необходимо выбрать так, чтобы передаточная функция эквалайзера $H_e(f)$ была значительно шире и охватывала весь спектр сигнала. Отметим, что сигнал на выходе эквалайзера по-прежнему выбирается с частотой $1/T$, но поскольку весовые коэффициенты отводов разделены промежутками T' (входящий сигнал эквалайзера выбирается с частотой $1/T'$), выравнивание принятого сигнала происходит до наложения его частотных компонентов. Моделирование эквалайзеров в телефонных линиях с $T' = T/2$ показывает, что эквалайзеры с фракционным разделением превосходят эквалайзеры с символьным разделением [14].

3.5. Резюме

В данной главе описаны два этапа процесса обнаружения двоичных сигналов в гауссовом шуме. Первый этап — это сжатие принятого сигнала до одного символа $z(T)$, а второй — принятие решения относительно первоначального значения принятого сигнала, для чего $z(T)$ сравнивается с определенным порогом. В главе рассказывается, как

выбрать оптимальный порог. Также показано, что линейный фильтр, известный как согласованный фильтр или коррелятор, — это оптимальный выбор для максимизации выходного отношения сигнал/шум, а значит, для минимизации вероятности ошибки.

Здесь дано определение межсимвольной интерференции (intersymbol interference — ISI) и объясняется значение работ Найквиста по установлению теоретической минимальной ширины полосы для обнаружения символов без ISI. Факторы роста вероятности ошибки были разбиты на две основные категории. Первая — это простое снижение отношения сигнал/шум. Вторая, проистекающая из искажения, — это выход зависимости вероятности ошибки от E_b/N_0 за область, представляющую практический интерес. В заключение описываются методы выравнивания, позволяющие уменьшить последствия межсимвольной интерференции.

Литература

1. Nyquist H. *Thermal Agitation of Electric Charge in Conductors.* Phys. Rev., vol. 32, July 1928, pp. 110–113.
2. Van Trees H. L. *Detection, Estimation and Modulation Theory.* Part 1, John Wiley & Sons, Inc., New York, 1968.
3. Arthurs E. and Dym H. *On the Optimum Detection of Digital Signals in the Presence of White Gaussian Noise — A Geometric Interpretation of Three Basic Data Transmission Systems.* IRE Trans. Commun. Syst., December, 1962.
4. Wozencraft J. M. and Jacobs I. M. *Principles of Communication Engineering.* John Wiley & Sons, Inc., New York, 1965.
5. Borjesson P. O. and Sundberg C. E. *Simple Approximations of the Error Function Q(x) for Communications Applications.* IEEE Trans. Commun., vol. COM27, March, 1979, pp. 639–642.
6. Nyquist H. *Certain Topics of Telegraph Transmission Theory.* Trans. Am. Inst. Electr. Eng., vol. 47, April, 1928, pp. 617–644.
7. Hanzo L. and Stefanov J. *The AN-European Digital Cellular Mobile Radio System — Known as GSM.* Mobile Radio Communications, edited by R. Steele, Chapter 8, Pentech Press, London, 1992.
8. Qureshi S. U. H. *Adaptive Equilization.* Proc. IEEE, vol. 73, n. 9, September, 1985, pp.1340–1387.
9. Lucky R. W., Salz J. and Weldon E. J., Jr. *Principles of Data Communications.* Mc-Graw Hill Book Co., New York, 1968.
10. Harris F. and Adams B. *Digital Signal Processing to Equalize the Pulse response of Non Synchronous Systems Such as Encountered in Sonar and Radar.* Proc. of the Twenty-Fourth Annual ASILOMAR Conference on Signals, Systems, and Computers, Pacific Grove, California, November, 5–7, 1990.
11. Proakis J. G. *Digital Communications.* McGraw-Hill Book Company, New York, 1983.
12. Feuer A. and Weinstein E. *Convergence Analysis of LMS Filters with Uncorrelated Gaussian Data.* IEEE Trans. on ASSP, vol. V-33, pp. 220–230, 1985.
13. Macchi O. *Adaptive Processing: Least Mean Square Approach With Applications in Transmission.* John Wiley & Sons, New York, 1995.
14. Benedetto S., Biglieri E. and Castellani V. *Digital Transmission Theory.* Prentice Hall, 1987.

Задачи

3.1. Определите, являются ли сигналы $s_1(t)$ и $s_2(t)$ ортогональными на интервале $(-1{,}5T_2 < t < 1{,}5T_2)$, где $s_1(t) = \cos(2\pi f_1 t + \varphi_1)$, $s_2(t) = \cos(2\pi f_2 t + \varphi_1)$, $f_2 = 1/T_2$, в следующих случаях.

а) $f_1 = f_2$ и $\varphi_1 = \varphi_2$

б) $f_1 = 1/3f_2$ и $\varphi_1 = \varphi_2$

в) $f_1 = f_2$ и $\varphi_1 = \varphi_2$

г) $f_1 = \pi f_2$ и $\varphi_1 = \varphi_2$

д) $f_1 = f_2$ и $\varphi_1 = \varphi_2 + \pi/2$

е) $f_1 = f_2$ и $\varphi_1 = \varphi_2 + \pi$

3.2. а) Покажите, что три функции, приведенные на рис. 33.1, попарно ортогональны на интервале (–2, 2).

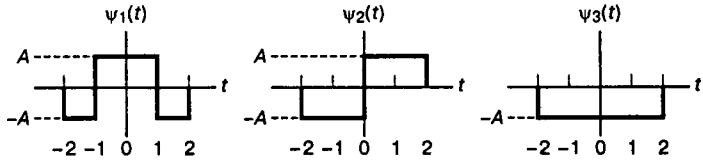


Рис. 33.1

б) Определите значение константы A, преобразующей набор функций из п. а в набор ортонормированных функций.

в) Выразите сигнал $x(t)$ через ортонормированные функции, полученные при выполнении п. б.

$$x(t) = \begin{cases} 1 & \text{для } 0 \le t \le 2 \\ 0 & \text{для остальных } t \end{cases}$$

3.3. Даны следующие функции

$$\psi_1(t) = \exp(-|t|) \text{ и } \psi_2(t) = 1 - A\exp(-2|t|)$$

Определите константу A, при которой функции $\psi_1(t)$ и $\psi_2(t)$ ортогональны на интервале $(-\infty, \infty)$.

3.4. Предположим, что используется некоторая система цифровой связи; сигнальные компоненты вне приемника-коррелятора с равной вероятностью принимают значения $a_i(T) = +1$ или -1 В. Определите вероятность появления ошибочного бита, если гауссов шум на выходе коррелятора имеет единичную дисперсию.

3.5. Биполярный двоичный сигнал $s_i(t)$ — это импульс $+1$ или -1 В на интервале $(0, T)$. К сигналу добавляется аддитивный белый гауссов шум с двусторонней спектральной плотностью мощности 10^{-3} Вт/Гц. Если обнаружение принятого сигнала производится с помощью согласованного фильтра, определите максимальную скорость передачи битов, которую можно поддерживать при вероятности появления ошибочного бита $P_B \le 10^{-3}$.

3.6. Биполярные импульсные сигналы $s_i(t)$ $(i = 1, 2)$ амплитуды ± 1 В принимаются при шуме AWGN с дисперсией 0,1 В^2. Определите оптимальный (дающий минимальную вероятность ошибки) порог γ_0 для обнаружения с использованием согласованного фильтра при следующих априорных вероятностях: (а) $P(s_1) = 0{,}5$; (б) $P(s_1) = 0{,}7$; (в) $P(s_1) = 0{,}2$. Объясните влияние априорных вероятностей на значение γ_0. (*Подсказка*: используйте формулы (Б.10)–(Б.12).)

3.7. Двоичная система связи передает сигналы $s_i(t)$ $(i = 1, 2)$. Тестовая статистика приемника $z(T) = a_i + n_0$, где компонент сигнала a_i равен $a_1 = +1$ или $a_2 = -1$, а компонент шума n_0 имеет равномерное распределение. Плотности условного распределения $p(z|s_i)$ даются выражениями

$$p(z|s_1) = \begin{cases} \frac{1}{2} & \text{для } -0{,}2 \le z \le 1{,}8 \\ 0 & \text{для других } z \end{cases}$$

и

$$p(z|s_2) = \begin{cases} \frac{1}{2} & \text{для } -1{,}8 \le z \le 0{,}2 \\ 0 & \text{для других } z \end{cases}$$

Определите вероятность появления ошибки P_B для равновероятной передачи сигналов и использования оптимального порога принятия решения.

3.8. а) Чему равна минимальная ширина полосы, необходимая для передачи без межсимвольной интерференции сигнала с использованием 16-уровневой кодировки PAM на скорости 10 Мбит/с?

б) Чему равен коэффициент сглаживания, если доступная полоса равна 1,375 МГц?

3.9. Сигнал речевого диапазона (300–3300 Гц) оцифровывается так, что квантовое искажение $\le \pm 0{,}1\%$ удвоенного максимального напряжения сигнала. Предположим, что частота дискретизации равна 8000 выборок/с и используется 32-уровневая кодировка PAM. Определите теоретическую минимальную ширину полосы, при которой еще не возникает межсимвольная интерференция.

3.10. Двоичные данные передаются со скоростью 9600 бит/с с использованием 8-уровневой модуляции PAM и фильтра с характеристикой типа приподнятого косинуса. Частотный отклик системы не превышает 2,4 кГц.

а) Чему равна скорость передачи символов?

б) Чему равен коэффициент сглаживания характеристики фильтра?

3.11. Сигнал речевого диапазона (300–3300 Гц) дискретизируется с частотой 8000 выборок/с. Выборки можно передавать сразу в виде импульсов PAM или каждую выборку вначале можно преобразовать в формат PCM и использовать для передачи двоичные (PCM) сигналы.

а) Чему равна минимальная ширина полосы системы, необходимая для обнаружения импульсов PAM без межсимвольной интерференции и с параметром сглаживания фильтра $r = 1$?

б) Используя ту же характеристику выравнивания, что и в предыдущем пункте, определите минимальную ширину полосы, необходимую для обнаружения двоичных сигналов (кодировка PCM), если выборки квантовались с использованием восьми уровней.

в) Повторите п. б для 128 уровней.

3.12. Аналоговый сигнал форматирован в формате PCM и передается с использованием двоичных сигналов через канал с полосой 100 кГц. Предполагается, что используются 32 уровня квантования и что полная эквивалентная передаточная функция — приподнятый косинус с выравниванием $r = 0{,}6$.

а) Найдите максимальную скорость передачи битов, которую может поддерживать система без межсимвольной интерференции.

б) Найдите максимальную ширину исходного аналогового сигнала, возможную при приведенных параметрах.

в) Повторите пп. а и б для 8-уровневой кодировки PAM.

3.13. Равновероятные двоичные импульсы в кодировке RZ когерентно обнаруживаются в гауссовом канале с $N_0 = 10^{-8}$ Вт/Гц. Предполагается, что синхронизация идеальна, амплитуда принятых импульсов равна 100 мВ и вероятность ошибки $P_B = 10^{-3}$; найдите наибольшую скорость передачи данных, возможную в описанной системе.

3.14. Двоичные импульсы в кодировке NRZ передаются по кабелю, ослабляющему сигнал на 3 дБ (на пути от передатчика к приемнику). Эти импульсы когерентно обнаруживаются приемником, а скорость передачи данных равна 56 Кбит/с. Шум считать гауссовым с

$N_0 = 10^{-6}$ Вт/Гц. Чему равна минимальная мощность, необходимая для передачи с вероятностью ошибки $P_B = 10^{-3}$?

3.15. Покажите, что минимальная ширина полосы по Найквисту для случайной двоичной последовательности с биполярными импульсами идеальной формы равна ширине полосы шумового эквивалента. *Подсказка*: спектральная плотность мощности случайной последовательности биполярных импульсов определяется формулой (1.38), а ширина полосы шумового эквивалента дана в разделе 1.7.2.

3.16. Дана четырехуровневая последовательность символов сообщений в кодировке PAM: {+1, +1, –1, +3, +1, +3}, где элементами алфавита являются числа {±1, ±3}. Импульсы формируются фильтром с характеристикой типа корня из приподнятого косинуса; время поддержки каждого фильтрованного импульса составляет 6 периодов передачи символа, передаваемая последовательность — аналоговый сигнал, показанный на рис. 3.23, *а*. Отметим, что сигналы "размываются" вследствие межсимвольной интерференции, вносимой фильтром. Покажите, как можно реализовать набор N корреляторов для выполнения демодуляции принятой последовательности импульсов $r(t)$ на согласованном фильтре, если число символов, переданных в течение длительности импульса, также равно N. (*Подсказка*: для набора корреляторов используйте опорные сигналы вида $s_1(t - kT)$, где $k = 0, \ldots, 5$, а T — время передачи символа.)

3.17. Желательным импульсным откликом системы является идеальный отклик $h(t) = \delta(t)$, где $\delta(t)$ — импульсная функция. Предполагается, что канал так вводит межсимвольную интерференцию, что общий импульсный отклик становится равным $h(t) = \delta(t) + \alpha\delta(t - T)$, где $\alpha < 1$, а T — длительность передачи символа. Выведите выражения для импульсного отклика фильтра, который реализует метод обращения в нуль незначащих коэффициентов и уменьшает последствия межсимвольной интерференции. Покажите, что этот фильтр подавляет межсимвольную интерференцию. Если полученное подавление окажется недостаточным, как можно будет модифицировать фильтр для более сильного подавления межсимвольной интерференции?

3.18. Результатом передачи одного импульса является принятая последовательность выборок (импульсный отклик) со значениями 0,1; 0,3; –0,2; 1,0; 0,4; –0,1; 0,1, где наиболее ранней является крайняя слева выборка. Значение 1,0 соответствует основному лепестку импульса, а другие — соседним выборкам. Спроектируйте трехотводный трансверсальный эквалайзер, подавляющий межсимвольную интерференцию в точках дискретизации по обе стороны основного лепестка. Вычислите значения выровненных импульсов в моменты времени $k = 0, \pm1, \ldots, \pm3$. Чему после выравнивания равен вклад наибольшей амплитуды в межсимвольную интерференцию и чему равна сумма амплитуд всех вкладов?

3.19. Повторите задачу 3.18, если импульсный отклик канала описывается следующими принятыми выборками: 0,01; 0,02; –0,03; 0,1; 1,0; 0,2: –0,1; 0,05; 0,02. С помощью компьютера найдите весовые коэффициенты девятиотводного трансверсального эквалайзера, удовлетворяющие критерию минимальности среднеквадратической ошибки. Вычислите значения импульсов на выходе эквалайзера в моменты времени $k = 0, \pm1, \ldots, \pm8$. Чему после выравнивания равен вклад наибольшей амплитуды в межсимвольную интерференцию и чему равна сумма амплитуд всех вкладов?

3.20. В данной главе отмечалось, что устройства обработки сигналов, такие как блоки перемножения и интегрирования, обычно работают с сигналами, имеющими размерность *вольт*. Таким образом, передаточная функция таких устройств должна выражаться в этих же единицах. Нарисуйте блочную диаграмму интегратора произведений, показывающую единицы сигналов в каждом проводнике и передаточную функцию устройства в каждом блоке. (*Подсказка*: см. раздел 3.2.5.1.)

Вопросы для самопроверки

3.1. При *узкополосной* передаче принятые сигналы уже имеют вид импульсов. Почему для восстановления импульсного сигнала требуется демодулятор (см. начало главы 3)?

3.2. Почему отношение E_b/N_0 является естественным критерием качества систем цифровой связи (см. раздел 3.1.5)?

3.3. При представлении упорядоченных во времени событий какая дилемма может легко привести к путанице между самым старшим битом и самым младшим (см. раздел 3.2.3.1)?

3.4. Термин *согласованный фильтр* часто используется как синоним термина *коррелятор*. Как такое возможно при совершенно разных математических операциях, описывающих их работу (см. раздел 3.2.3.1)?

3.5. Опишите два точных способа сравнения различных кривых, описывающих зависимость вероятности появления ошибочного бита от отношения E_b/N_0 (см. раздел 3.2.5.3).

3.6. Существуют ли функции фильтров формирования импульсов (отличные от приподнятого косинуса), дающие нулевую межсимвольную интерференцию (см. раздел 3.3)?

3.7. До какой степени можно *сжать полосу*, не подвергаясь при этом межсимвольной интерференции (см. раздел 3.3.1.1)?

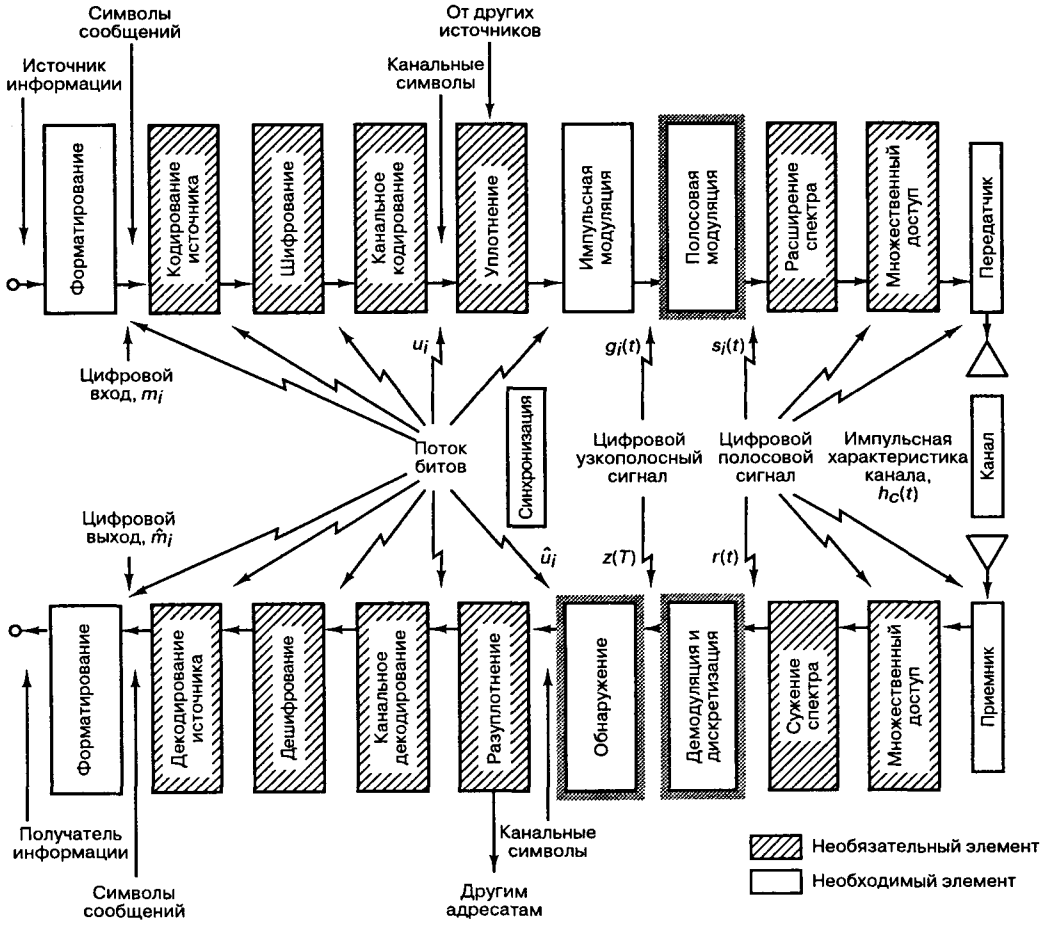
3.8. Ухудшение качества сигнала определяется двумя основными факторами: *снижением* отношения сигнал/шум и *искажением*, приводящим к не поддающейся улучшению вероятности возникновения ошибки. Чем отличаются эти факторы (см. раздел 3.3.2)?

3.9. Иногда увеличение отношения E_b/N_0 не останавливает ухудшение качества, вызванное *межсимвольной интерференцией*. Когда это происходит (см. раздел 3.3.2)?

3.10. Чем отличается эквалайзер, реализовывающий метод *обращения в нуль незначащих коэффициентов*, от эквалайзера, реализовывающего решение с *минимальной среднеквадратической ошибкой* (см. раздел 3.4.3.1)?

ГЛАВА 4

Полосовая модуляция и демодуляция



4.1. Зачем нужна модуляция

Цифровая модуляция — это процесс преобразования цифровых символов в сигналы, совместимые с характеристиками канала. При узкополосной модуляции (baseband modulation) эти сигналы обычно имеют вид импульсов заданной формы. В случае *полосовой модуляции* (bandpass modulation) импульсы заданной формы модулируют синусоиду, называемую *несущей волной* (carrier wave), или просто *несущей* (carrier), затем следует передача на нужное расстояние с использованием радиочастот; для этого несущая преобразовывается в электромагнитное поле. Может возникнуть вопрос: зачем для радиопередачи узкополосных сигналов нужна несущая? Ответ звучит следующим образом. Передача электромагнитного поля через пространство выполняется с помощью антенн. Размер антенны зависит от длины волны λ и текущей задачи. Для переносных телефонов размер антенны обычно равен $\lambda/4$, а длина волны c/f, где c — скорость света, 3×10^8 м/с. Рассмотрим передачу узкополосного сигнала (скажем, имеющего частоту $f = 3000$ Гц) путем сопряжения его непосредственно с антенной без использования несущей. Какая антенна нам понадобится? Возьмем стандарт телефонной промышленности, $\lambda/4$. Получаем, что для узкополосного сигнала 3000 Гц $\lambda/4 = 2{,}5 \times 10^4$м = 25 км. Итак, для передачи через пространство сигнала с частотой 3000 Гц без *модулирования несущей* требуется антенна размером 25 км. При этом, если узкополосная информация модулируется несущей более высокой частоты, например 900 МГц, размер антенны будет составлять порядка 8 см. Приведенные вычисления показывают, что модулирование несущей частоты, или полосовая модуляция, — это этап, необходимый для всех систем, использующих радиопередачу.

Полосовая модуляция имеет и другие важные преимущества при передаче сигналов. При использовании одного канала более чем одним сигналом, модуляция может применяться для выделения различных сигналов. Подобный метод, известный как *уплотнение с частотным разделением* (frequency-division multiplexing — FDM), рассматривается в главе 11. Модуляция может использоваться и для минимизации последствий интерференции. Класс схем модулирования, известный как *модулирование расширенным спектром*, требует полосы, значительно превышающей минимальную полосу, необходимую для передачи сообщения. В главе 12 рассмотрены компромиссы, связанные с выбором полосы, снижающим интерференцию. Кроме того, модуляция может использоваться для перемещения сигнала в диапазон частот, в котором легко удовлетворяются специфические конструктивные требования, например, относящиеся к фильтрации и усилению. Примером такого применения модуляции является преобразование в приемнике радиочастотных сигналов в сигналы промежуточной частоты.

4.2. Методы цифровой полосовой модуляции

Полосовая модуляция (аналоговая или цифровая) — это процесс преобразования информационного сигнала в синусоидальную волну; при цифровой модуляции синусоида на интервале T называется цифровым символом. Синусоиды могут отличаться по амплитуде, частоте и фазе. Таким образом, полосовую модуляцию можно определить как процесс варьирования амплитуды, частоты или фазы (или их комбинаций) радиочастотной несущей согласно передаваемой информации. В общем виде несущая записывается следующим образом.

$$s(t) = A(t) \cos \theta(t) \tag{4.1}$$

Здесь $A(t)$ — переменная во времени амплитуда, а $\theta(t)$ — переменный во времени угол. Угол удобно записывать в виде

$$\theta(t) = \omega_0 t + \phi(t), \tag{4.2}$$

так что

$$s(t) = A(t) \cos [\omega_0 t + \phi(t)], \tag{4.3}$$

где ω — *угловая частота* несущей, а $\phi(t)$ — ее *фаза*. Частота может записываться как переменная f или как переменная ω. В первом случае частота измеряется в герцах (Гц), во втором — в радианах в секунду (рад/с). Эти параметры связаны следующим соотношением $\omega = 2\pi f$.

Основные типы *полосовой модуляции/демодуляции* перечислены на рис. 4.1. Если для обнаружения сигналов приемник использует информацию о фазе несущей, процесс называется *когерентным обнаружением* (coherent detection); если подобная информация не используется, процесс именуется *некогерентным обнаружением* (noncoherent detection). Вообще, в цифровой связи термины "демодуляция" (demodulation) и "обнаружение" (detection) часто используются как синонимы, хотя демодуляция делает акцент на восстановлении сигнала, а обнаружение — на принятии решения относительно символьного значения принятого сигнала. При идеальном когерентном обнаружении приемник содержит прототипы каждого возможного сигнала. Эти сигналы-прототипы дублируют алфавит переданных сигналов по всем параметрам, даже по радиочастотной фазе. В этом случае говорят, что приемник *автоматически подстраивается* под фазу входящего сигнала. В процессе демодуляции приемник перемножает и интегрирует входящий сигнал с каждым прототипом (определяет корреляцию). На рис. 4.1 под общим заголовком когерентной модуляции/демодуляции перечислены: фазовая манипуляция (phase shift keying — PSK), частотная манипуляция (frequency shift keying — FSK), амплитудная манипуляция (amplitude shift keying — ASK), модуляция без разрыва фазы (continuous phase modulation — CPM) и смешанные комбинации этих модуляций. Основные форматы полосовой модуляции рассмотрены в данной главе. Некоторые специализированные форматы, такие как квадратурная фазовая манипуляция со сдвигом (offset quadrature PSK — OQPSK), манипуляция с минимальным сдвигом (minimum shift keying — MSK), принадлежащие к классу модуляций CPM, и квадратурная амплитудная модуляция (quadrature amplitude modulation — QAM), рассмотрены в главе 9.

Некогерентная демодуляция относится к системам, использующим демодуляторы, спроектированные для работы без знания абсолютной величины фазы входящего сигнала; следовательно, определение фазы в этом случае не требуется. Таким образом, преимуществом некогерентных систем перед когерентными является простота, а недостатком — большая вероятность ошибки (P_E). На рис. 4.1 под заголовком некогерентной передачи сигналов перечислены модуляции, подобные используемым при когерентной передаче: DPSK, FSK, ASK, CPM и смешанные их комбинации. Подразумевается, что для некогерентного приема информация о фазе не используется; так почему же под заголовком "некогерентная передача" указана одна из форм фазовой манипуляции? Это вызвано тем, что одну из важных форм PSK можно отнести к некогерентной (или дифференциально когерентной), поскольку она не требует согласования по фазе с принятой несущей. При использовании этой "псевдо-PSK", называемой *дифференциальной* фазовой манипуляцией (differential PSK — DPSK), в процессе обнаружения текущего символа в качестве опорной фазы применяется фаза предыдущего символа. Подробно этот вопрос рассмотрен в разделах 4.5.1 и 4.5.2.

Рис. 4.1. Основные преобразования цифровой связи

4.2.1. Векторное представление синусоиды

Используя известное тригонометрическое равенство, называемое теоремой Эйлера, введем комплексную запись синусоидальной несущей.

$$e^{i\omega_0 t} = \cos \omega_0 t + i \sin \omega_0 t \tag{4.4}$$

Возможно, кто-то чувствует себя уютнее при использовании более простой, привычной записи $\cos \omega_0 t$ или $\sin \omega_0 t$. Возникает естественный вопрос: что нам дает комплексная запись? Далее будет показано (раздел 4.6), что такая форма записи облегчает описание реализации реальных модуляторов и демодуляторов. Здесь же мы рассмотрим общие преимущества представления несущей в комплексной форме, приведенной в формуле (4.4).

Во-первых, при комплексной записи в компактной форме, $e^{i\omega_0 t}$, указаны два важных компонента любой синусоидальной несущей волны, называемые взаимно ортогональными синфазной (действительной) и квадратурной (мнимой) составляющими. Во-вторых, как показано на рис. 4.2, немодулированная несущая удобно представляется в полярной системе координат в виде единичного вектора с постоянной скоростью ω_0 рад/с, вращающегося против часовой стрелки. При увеличении t (от t_0 до t_1) мы можем изобразить переменные во времени проекции вращающегося вектора на синфазной (I) и квадратурной (Q) осях. Эти декартовы оси обычно называются синфазным (I channel) и квадратурным каналом (Q channel), а их проекции представляют взаимно ортогональные составляющие сигнала, связанные с этими каналами. В-третьих, процесс модуляции несущей можно рассматривать как систематическое возмущение вращающегося вектора (и его проекций).

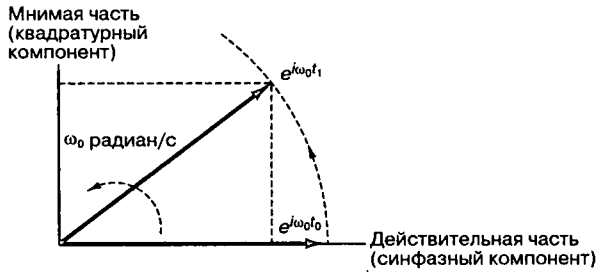


Рис. 4.2. Векторное представление синусоиды

Рассмотрим, например, несущую, *амплитудно-модулированную* синусоидой с единичной амплитудой и частотой ω_m, где $\omega_m << \omega_0$. Переданный сигнал имеет следующий вид.

$$s(t) = \mathrm{Re}\left\{ e^{i\omega_0 t} \left(1 + \frac{e^{i\omega_m t}}{2} + \frac{e^{-i\omega_m t}}{2} \right) \right\}, \tag{4.5}$$

где $\mathrm{Re}\{x\}$ — действительная часть комплексной величины $\{x\}$. На рис. 4.3 показано, что вращающийся вектор $e^{i\omega_0 t}$, представленный на рис. 4.2, возмущается двумя боковыми членами — $e^{i\omega_m t}$, вращающимся против часовой стрелки, и $e^{-i\omega_m t}$, вращающимся по часовой стрелке. Боковые векторы вращаются намного медленнее, чем вектор несущей волны. В результате модулированный вращающийся вектор несущей волны растет и уменьшается согласно указаниям боковых полос, но частота его вращения остается постоянной; отсюда и название "амплитудная модуляция".

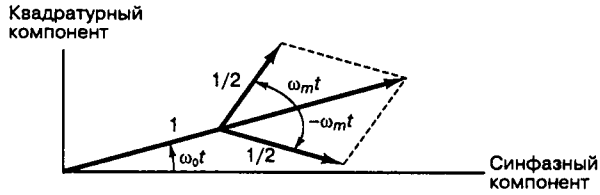


Рис. 4.3. Амплитудная модуляция

Еще один пример, иллюстрирующий полезность векторного представления, — это *частотная модуляция* (frequency modulation — FM) несущей похожей синусоидой с частотой вращения ω_m рад/с. Аналитическое представление *узкополосной* частотной модуляции (narrowband FM — NFM) подобно представлению амплитудной модуляции и описывается выражением:

$$s(t) = \mathrm{Re}\left\{e^{i\omega_0 t}\left(1 - \frac{\beta}{2}e^{-i\omega_m t} + \frac{\beta}{2}e^{i\omega_m t}\right)\right\}, \tag{4.6}$$

где β — коэффициент модуляции [1]. На рис. 4.4 показано, что, как и в предыдущем случае, вектор несущей волны возмущается двумя боковыми векторами. Но поскольку один из них, как указано в формуле (4.6), имеет знак "минус", симметрия боковых векторов, вращающихся по часовой стрелке и против нее, отличается от имеющейся в случае амплитудной модуляции. При модуляции AM симметрия приводит к увеличению и уменьшению вектора несущей волны со временем. В случае модуляции NFM симметрия боковых векторов (на 90° отличающаяся от симметрии AM) приводит к ускорению и замедлению вращения вектора согласно указаниям боковых полос, при этом амплитуда остается неизменной; отсюда название "частотная модуляция".

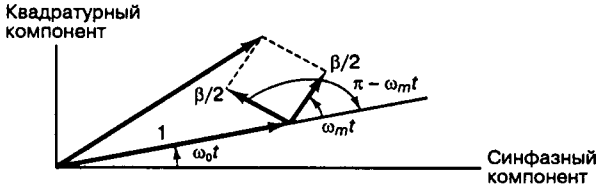


Рис. 4.4. Узкополосная частотная модуляция

На рис. 4.5 изображены наиболее распространенные форматы цифровой модуляции: PSK, FSK, ASK и смешанная комбинация ASK и PSK (обозначаемая как ASK/PSK, или АРК). В первом столбце указаны аналитические выражения, во втором — временная диаграмма, а в третьем — векторная диаграмма. В общем случае M-арной передачи сигналов устройство обработки получает k исходных битов (или канальных битов, если используется кодирование) в каждый момент времени и указывает модулятору произвести один из $M = 2^k$ возможных сигналов. Частным случаем M-уровневой модуляции является бинарная с $k = 1$.

На рис. 4.2 несущая волна представлялась как вектор, вращающийся на плоскости со скоростью, равной частоте несущей, ω_0 рад/с. На рис. 4.5 векторная схема каждой цифровой модуляции представляет совокупность информационных сигналов (векторов или точек пространства сигналов) без указания времени. Другими словами,

на рис. 4.5 не отображено вращение немодулированного сигнала с постоянной скоростью, а представлено только взаимное расположение векторов-носителей информации. Стоит обратить внимание, что в примерах на рис. 4.5 значения размера множества M отличаются.

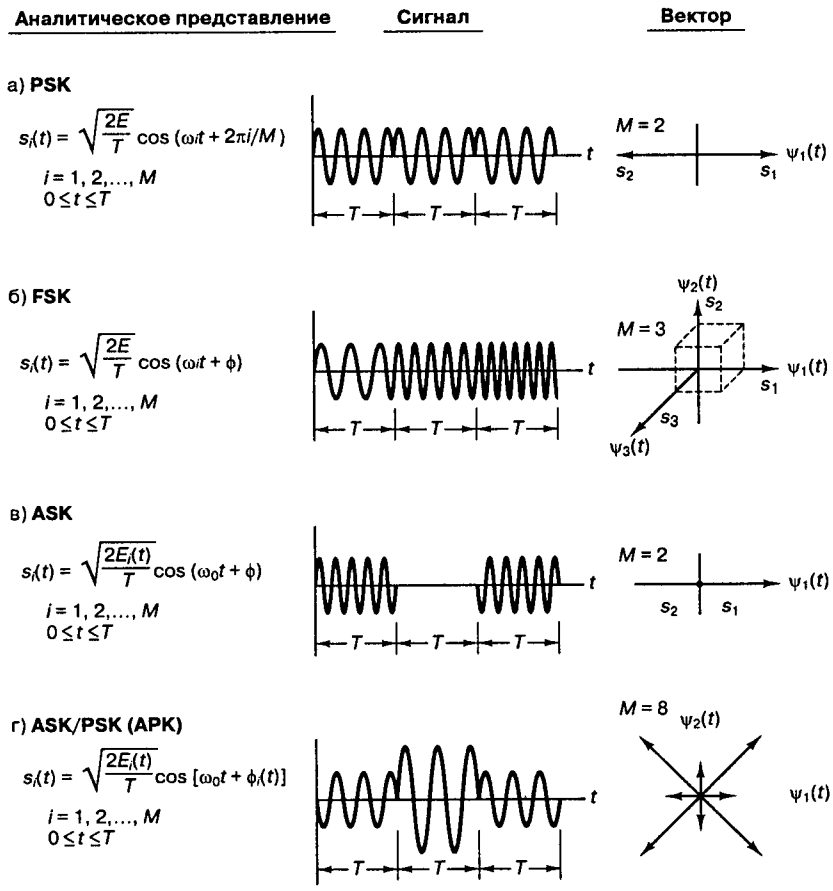


Рис. 4.5. Виды цифровых модуляций: а) PSK; б) FSK; в) ASK; г) ASK/PSK (APK)

4.2.2. Фазовая манипуляция

Фазовая манипуляция (phase shift keying — PSK) была разработана в начале развития программы исследования дальнего космоса; сейчас схема PSK широко используется в коммерческих и военных системах связи. Сигнал в модуляции PSK имеет следующий вид.

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos[\omega_0 t + \phi_i(t)] \quad 0 \le t \le T \quad i = 1, \ldots, M \tag{4.7}$$

Здесь фазовый член $\phi_i(t)$ может принимать M дискретных значений, обычно определяемых следующим образом.

$$\phi_i(t) = \frac{2\pi i}{M} \quad i = 1, \ldots, M$$

На рис. 4.5, *а* приведен пример двоичной ($M = 2$) фазовой манипуляции (binary PSK — BPSK). Параметр E — это энергия символа, T — время передачи символа, $0 \le t \le T$. Работа схемы модуляции заключается в смещении фазы модулируемого сигнала $s_i(t)$ на одно из двух значений, нуль или π (180°). Типичный вид сигнала в модуляции BPSK приведен на рис. 4.5, *а*, где явно видны характерные резкие изменения фазы при переходе между символами; если модулируемый поток данных состоит из чередующихся нулей и единиц, такие резкие изменения будут происходить при каждом переходе. Модулированный сигнал можно представить как вектор на графике в полярной системе координат; длина вектора соответствует амплитуде сигнала, а его ориентация в общем M-арном случае — фазе сигнала относительно других $M - 1$ сигналов набора. При модуляции BPSK векторное представление дает два противофазных (180°) вектора. Наборы сигналов, которые могут быть представлены подобными противофазными векторами, называются *антиподными*.

4.2.3. Частотная манипуляция

Общее аналитическое выражение для сигнала в частотной манипуляции (frequency shift keying — FSK) имеет следующий вид.

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos(\omega_i t + \phi) \quad 0 \le t \le T \qquad (4.8)$$
$$i = 1, \ldots, M$$

Здесь частота ω_0 может принимать M дискретных значений, а фаза ϕ является произвольной константой. Схематическое изображение сигнала в модуляции FSK дано на рис. 4.5, *б*, где можно наблюдать типичное изменение частоты (тона) в моменты переходов между символами. Такое поведение характерно только для частного случая FSK, называемого частотной манипуляцией без разрыва фазы (continuous-phase FSK — CPFSK); она описана в разделе 9.8. В общем случае многочастотной манипуляции (multiple frequency shift keying — MFSK) переход к другому тону может быть довольно резким, поскольку непрерывность фазы здесь не обязательна. В приведенном примере $M = 3$, что соответствует такому же числу типов сигналов (троичной передаче); отметим, что значение $M = 3$ было выбрано исключительно для демонстрации на рисунке взаимно перпендикулярных осей. На практике M обычно является ненулевой степенью двойки (2, 4, 8, 16, ...), что довольно сложно изобразить графически. Множество сигналов описывается в декартовой системе координат, где каждая координатная ось представляет синусоиду определенной частоты. Как говорилось ранее, множества сигналов, которые описываются подобными взаимно перпендикулярными векторами, называются *ортогональными* (orthogonal). Не все схемы FSK относятся к ортогональным. Чтобы множество сигналов было ортогональным, оно должно удовлетворять критерию, выраженному в формуле (3.69). Этот критерий навязывает определенные условия на взаимное размещение тонов множества. Расстояние по частоте между тонами, необходимое для удовлетворения требования ортогональности, рассмотрено в разделе 4.5.4.

4.2.4. Амплитудная манипуляция

Сигнал в амплитудной манипуляции (amplitude shift keying — ASK), изображенной на рис. 4.5, *в*, описывается выражением

$$s_i(t) = \sqrt{\frac{2E_i(t)}{T}} \cos(\omega_0 t + \phi) \quad 0 \le t \le T \qquad (4.9)$$

$$i = 1, \ldots, M,$$

где амплитудный член $\sqrt{2E_i(t)/T}$ может принимать M дискретных значений, а фазовый член ϕ — это произвольная константа. На рис. 4.5, *в* M выбрано равным 2, что соответствует двум типам сигналов. Изображенный на рисунке сигнал в модуляции ASK может соответствовать радиопередаче с использованием двух сигналов, амплитуды которых равны 0 и $\sqrt{2E/T}$. В векторном представлении использованы те же фазово-амплитудные полярные координаты, что и в примере для модуляции PSK. Правда, в данном случае мы видим один вектор, соответствующий максимальной амплитуде с точкой в начале координат, и второй, соответствующий нулевой амплитуде. Передача сигналов в двухуровневой модуляции ASK — это одна из первых форм цифровой модуляции, изобретенных в начале столетия для беспроводной телеграфии. В настоящее время простая схема ASK в системах цифровой связи уже не используется, поэтому в данной книге мы не будем рассматривать ее подробно.

4.2.5. Амплитудно-фазовая манипуляция

Амплитудно-фазовая манипуляция (amplitude phase keying — APK) — это комбинация схем ASK и PSK. Сигнал в модуляции APK изображен на рис. 4.5, *г* и выражается как

$$s_i(t) = \sqrt{\frac{2E_i(t)}{T}} \cos(\omega_0 t + \phi_i(t)) \quad 0 \le t \le T \qquad (4.10)$$

$$i = 1, \ldots, M$$

с индексированием амплитудного и фазового членов. На рис. 4.5, *г* можно видеть характерные одновременные (в моменты перехода между символами) изменения фазы и амплитуды сигнала в модуляции APK. В приведенном примере $M = 8$, что соответствует 8 сигналам (восьмеричной передаче). Возможный набор из восьми векторов сигналов изображен на графике в координатах "фаза-амплитуда". Четыре показанных вектора имеют одну амплитуду, еще четыре — другую. Векторы ориентированы так, что угол между двумя ближайшими векторами составляет 45°. Если в двухмерном пространстве сигналов между M сигналами набора угол прямой, схема называется квадратурной амплитудной модуляцией (quadrature amplitude modulation — QAM); примеры QAM рассмотрены в главе 9.

Векторные представления модуляций, изображенные на рис. 4.5 (за исключением случая FSK), изображены графиками, *полярные* координаты которых представляют *амплитуду* и *фазу* сигнала. Схема FSK подразумевает ортогональную передачу (см. раздел 4.5.4) и описывается в *декартовой* системе координат, где каждая ось представляет *тон частоты* ($\cos \omega_i t$), совокупность которых формирует M-значный набор ортогональных тонов.

4.2.6. Амплитуда сигнала

Амплитуды сигналов, представленные в формулах (4.7)–(4.10), имеют одинаковый вид $\sqrt{2E/T}$ для всех форматов модуляции. Выведем это. Сигнал описывается формулой

$$s(t) = A \cos \omega t, \tag{4.11}$$

где A — максимальная амплитуда сигнала. Поскольку максимальное значение в $\sqrt{2}$ раза больше его среднеквадратического (root-mean-square — rms) значения, можем записать следующее.

$$s(t) = \sqrt{2} A_{\text{rms}} \cos \omega t =$$
$$= \sqrt{2 A_{\text{rms}}^2} \cos \omega t$$

Предполагается, что сигнал выражен через колебания тока или напряжения, так что A_{rms}^2 представляет среднюю мощность P (нормированную на 1 Ом). Значит, можем записать следующее.

$$s(t) = \sqrt{2P} \cos \omega t \tag{4.12}$$

Заменяя P (единицы измерения — ватт) на E (джоули)/T (секунды), получаем следующее.

$$s(t) = \sqrt{\frac{2E}{T}} \cos \omega t \tag{4.13}$$

Итак, амплитуду сигнала можно записать либо в общем виде, как в формуле (4.11), либо через $\sqrt{2E/T}$, как в формуле (4.13). Поскольку ключевой параметр при определении вероятности ошибки в процессе обнаружения — это *энергия* принятого сигнала, зачастую удобнее использовать запись в форме (4.13), так как в этом случае вероятность ошибки P_E можно получить сразу как функцию энергии сигнала.

4.3. Обнаружение сигнала в гауссовом шуме

Полосовая модель процесса обнаружения, рассмотренная в данной главе, практически идентична узкополосной модели, представленной в главе 3. Дело в том, что принятый полосовой сигнал вначале преобразовывается в узкополосный, после чего наступает этап окончательного обнаружения. Для линейных систем математика процесса обнаружения не зависит от смещения частоты. Фактически *теорему эквивалентности* можно определить следующим образом: выполнение полосовой линейной обработки сигнала с последующим наложением сигнала (превращением полосового сигнала в узкополосный) дает те же результаты, что и наложение сигнала с последующей узкополосной линейной обработкой сигнала. Термин "наложение сигнала" (heterodyning) обозначает *преобразование* частоты или процесс *смешивания*, вызывающий смещение спектра сигнала. Как следствие теоремы эквивалентности, любая линейная модель обработки сигналов может использоваться для узкополосных сигналов (что предпочтительнее с точки зрения простоты) с теми же результатами, что и для полосовых сигналов. Это означает, что производительность большинства цифровых систем связи часто можно описать и проанализировать, считая канал передачи узкополосным.

4.3.1. Области решений

Предположим, что двухмерное пространство сигналов, изображенное на рис. 4.6, — это геометрическое место точек, возмущенных шумом двоичных векторов-прототипов

($\mathbf{s}_1 + \mathbf{n}$) и ($\mathbf{s}_2 + \mathbf{n}$). Вектор шума $\mathbf{n}$ — это случайный вектор с нулевым средним; следовательно, вектор принятого сигнала $\mathbf{r}$ — это случайный вектор со средним значением $\mathbf{s}_1$ или $\mathbf{s}_2$. Задачей детектора после получения $\mathbf{r}$ является принятие решения относительно классификации сигнала, имеющего минимальную вероятность ошибки P_B, хотя возможны и другие стратегии принятия решения [2]. Для случая $M = 2$ с равновероятными сигналами $\mathbf{s}_1$ и $\mathbf{s}_2$ и при шуме AWGN (additive white Gaussian noise — аддитивный белый гауссов шум) использование при принятии решения критерия минимума ошибки равносильно такому выбору класса сигнала, чтобы расстояние $d(\mathbf{r}, \mathbf{s}_i) = \|\mathbf{r} - \mathbf{s}_i\|$ было минимальным, где $\|\mathbf{x}\|$ — норма или абсолютная величина вектора $\mathbf{x}$. Последнее правило часто формулируется в терминах областей решений. Обратимся к рис. 4.6 и рассмотрим формирование областей решений. Итак, вначале необходимо соединить концы векторов-прототипов $\mathbf{s}_1$ и $\mathbf{s}_2$. Затем через середину полученного отрезка проводится плоскость, перпендикулярная к нему. Отметим, что поскольку амплитуды сигналов $\mathbf{s}_1$ и $\mathbf{s}_2$ равны, эта плоскость проходит через начало координат и является биссектрисой угла, образованного векторами-прототипами. Эта биссекторная плоскость, изображенная на рис. 4.6 для случая $M = 2$, является геометрическим местом точек, равноудаленных от векторов $\mathbf{s}_1$ и $\mathbf{s}_2$; следовательно, она является границей между областью решений 1 и областью решений 2. *Правило принятия решения*, используемое детектором, формулируется в терминах *областей решений* следующим образом: если сигнал расположен в области 1 — отнести принятый сигнал к $\mathbf{s}_1$; если в области 2 — выбрать сигнал $\mathbf{s}_2$. Если угол θ (рис. 4.6) равен 180°, набор сигналов $\mathbf{s}_1$ и $\mathbf{s}_2$ описывает модуляцию BPSK. Впрочем, для иллюстрации идеи области решений вообще угол θ на рисунке был заведомо выбран меньшим 180°.

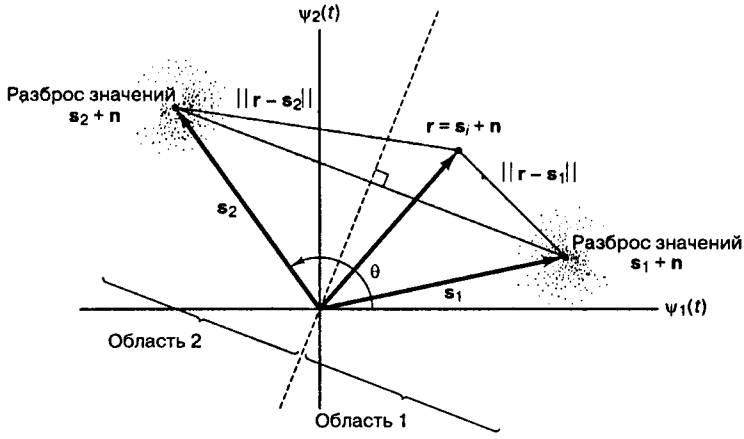


Рис. 4.6. Двухмерное пространство сигналов с равными по модулю произвольными векторами $\mathbf{s}_1$ *и* $\mathbf{s}_2$

4.3.2. Корреляционный приемник

В разделе 3.2 было рассмотрено обнаружение *узкополосных* двоичных сигналов в гауссовом шуме. Поскольку при обнаружении *полосовых* сигналов используются те же понятия, в данном разделе мы просто обобщим ключевые результаты. Ос-

новное внимание будет уделено реализации согласованного фильтра, известного как *коррелятор* (correlator). Помимо двоичного обнаружения будет рассмотрен более общий случай M-арного обнаружения. Предполагается, что сигнал искажается только вследствие шума AWGN. Принятый сигнал будем описывать как сумму переданного сигнала и случайного шума.

$$r(t) = s_i(t) + n(t) \quad 0 \le t \le T \qquad i = 1, \ldots, M \tag{4.14}$$

При наличии подобного принятого сигнала процесс обнаружения, как показано на рис. 3.1, включает *два основных этапа*. На первом этапе принятый сигнал $r(t)$ усекается до *одной случайной переменной* $z(T)$ или до *набора случайных переменных* $z_i(T)$ ($i = 1, \ldots, M$), формируемых на выходе демодулятора и устройства дискретизации в момент времени $t = T$, где T — длительность символа. На втором этапе на основе сравнения $z(T)$ с порогом или согласно критерию максимума $z_i(T)$ принимается решение относительно значения символа. Вообще, этап 1 можно рассматривать как преобразование сигнала в точку в пространстве решений. Эту точку, представляющую собой важнейшую контрольную точку в приемнике, можно назвать *додетекторной* (predetection). Когда мы говорим о мощности принятого сигнала, мощности принятых шумов или отношении E_b/N_0, все эти величины всегда рассматриваются относительно додетекторной точки. Иногда такие параметры определяются относительно *входа приемника* или *принимающей антенны*. Но в подобных случаях всегда подразумевается, что между выбранной и додетекторной точками не происходит снижения отношения сигнал/шум, или E_b/N_0. В каждый момент передачи символа сигнал, доступный в додетекторной точке, является выборкой узкополосного импульса. На данный момент битового значения у нас еще нет. Стоит ли удивляться, что отношение энергии бита к N_0 *определено* там, где еще не существует бита? В действительности, нет, поскольку данная точка является удобной контрольной точкой, где узкополосный импульс — даже до принятия решения на битовом уровне — может давать *эффективное* представление битов. Этап 2 можно рассматривать как определение того, *в какой области решений* расположена данная точка. Для оптимизации детектора (в смысле минимизации вероятности ошибки) необходимо оптимизировать преобразование сигнала в случайную переменную с использованием согласованных фильтров или корреляторов на этапе 1 и оптимизировать критерий принятия решения на этапе 2.

В разделах 3.2.2 и 3.2.3 показывалось, что согласованный фильтр обеспечивает максимальное отношение сигнал/шум на выходе фильтра в момент $t = T$. Как одна из реализаций согласованного фильтра описывался коррелятор. Теперь мы можем определить *корреляционный приемник* (correlation receiver), состоящий, как показано на рис. 4.7, *а*, из M корреляторов, выполняющих преобразование принятого сигнала $r(t)$ в последовательность M чисел или выходов коррелятора, $z_i(T)$ ($i = 1, \ldots, M$). Каждый выход коррелятора описывается следующим интегралом произведения или корреляцией с принятым сигналом.

$$z_i(T) = \int_0^T r(t)s_i(t)\,dt \qquad i = 1, \ldots, M \tag{4.15}$$

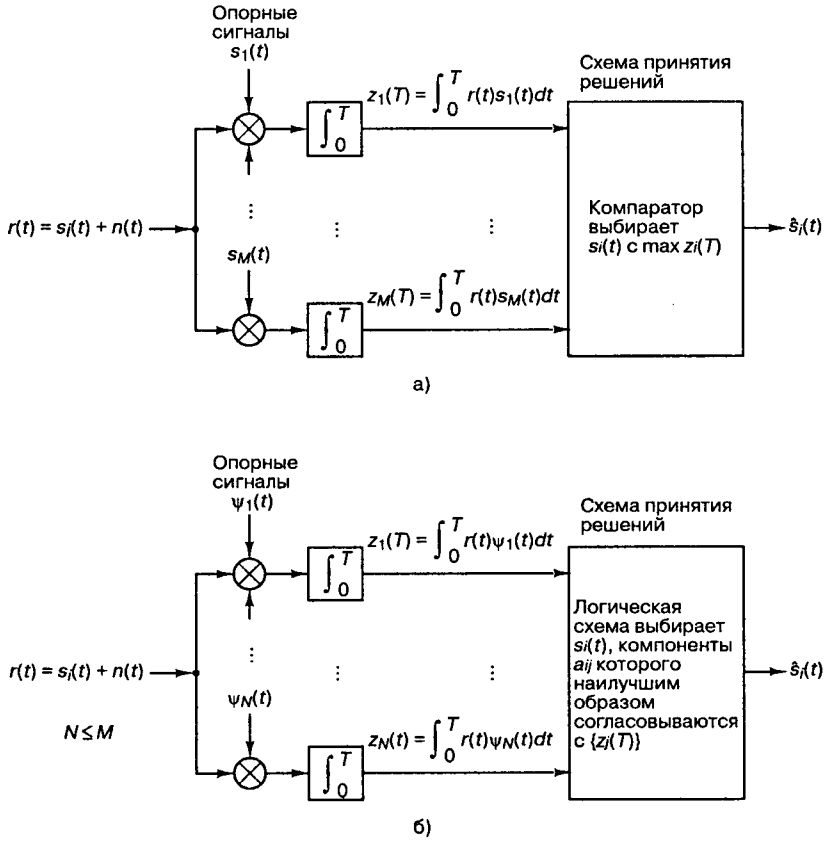


Рис. 4.7. Корреляционный приемник: а) корреляционный приемник с опорными сигналами $\{s_i(t)\}$; б) корреляционный приемник с опорными сигналами $\{\psi_j(t)\}$

Глагол "коррелировать" означает "совпадать", "согласовываться". Корреляторы пытаются найти соответствие принятого сигнала $r(t)$ с каждым возможным сигналом-прототипом $s_i(t)$, известным приемнику априори. Разумное правило принятия решения звучит так: выбирать сигнал $s_i(t)$, *лучше всего согласующийся*, (или *имеющий наибольшую корреляцию*) с $r(t)$. Другими словами, правило принятия решения выглядит следующим образом.

$$\text{Выбрать сигнал } s_i(t), \text{ индекс которого соответствует максимальной } z_i(T) \tag{4.16}$$

Следуя формуле (3.10), любой набор сигналов $\{s_i(t)\}$ $(i = 1, \ldots, M)$ можно выразить через определенный набор базисных функций $\{\psi_j(t)\}$ $(j = 1, \ldots, N)$, где $N \le M$. Таким образом, группу из M корреляторов, изображенную на рис. 4.7, *а*, можно заменить группой из N корреляторов, показанной на рис. 4.7, *б*, где в качестве *опорных сигналов* используется набор базисных функций $\{\psi_j(t)\}$. Для принятия решения с помощью указанных корреляторов необходима логическая схема выбора сигнала $s_i(t)$. Выбор производится на основе определения наилучшего согласования коэффициентов a_{ij}, фигурирующих в формуле (3.10), с набором выходов $\{z_j(T)\}$. Если набор сигналов-прототипов $\{s_i(t)\}$ формирует ортогональное

множество, реализация приемника, показанная на рис. 4.7, *а*, идентична реализации, показанной на рис. 4.7, *б* (могут отличаться масштабом). Если же $\{s_i(t)\}$ *не является* ортогональным множеством, приемник (рис. 4.7, *б*), использующий N корреляторов с опорными сигналами $\{\psi_j(t)\}$ вместо M, представляет более рентабельную реализацию. В разделе 4.4.3 мы рассмотрим применение подобного устройства для обнаружения сигнала в модуляции MPSK (multiple phase shift keying — многофазная манипуляция).

В случае *двоичного обнаружения* корреляционный приемник, как показано на рис. 4.8, *а*, можно построить как согласованный фильтр или интегратор произведений с опорным сигналом, равным разности двоичных сигналов-прототипов $s_1(t) - s_2(t)$. Выход коррелятора $z(T)$ используется непосредственно в процессе принятия решения.

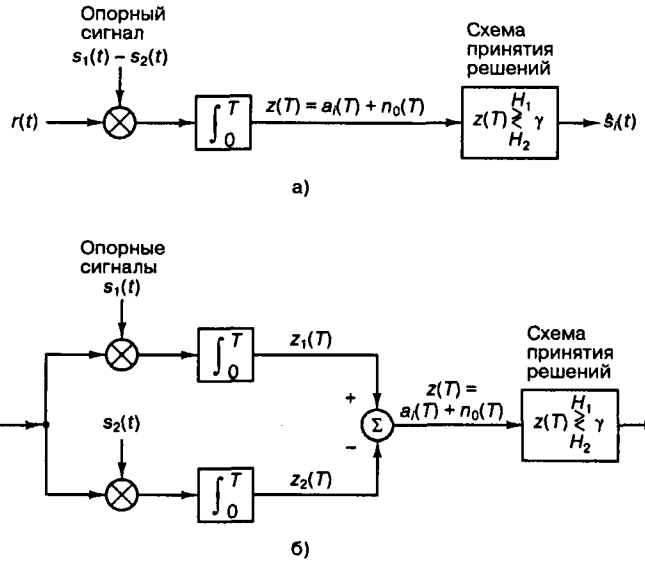


Рис. 4.8. Двоичный корреляционный приемник: а) использование одного коррелятора; б) применение двух корреляторов

При двоичном обнаружении корреляционный приемник можно изобразить как два согласованных фильтра или интегратора произведений, один из которых согласовывается с $s_1(t)$, а второй — с $s_2(t)$ (рис. 4.8, *б*). На этапе принятия решения теперь может использоваться правило, приведенное в формуле (6.16), или же из выхода одного коррелятора можно вычесть выход другого и на этапе принятия решения использовать разность

$$z(T) = z_1(T) - z_2(T), \tag{4.17}$$

как показано на рис. 4.8, *б*. Здесь $z(T)$, называемое *тестовой статистикой* (test statistic), подается в схему принятия решения, как и в случае только одного коррелятора. В *отсутствие шума* на выходе мы получаем $z(T) = a_i(T)$, где $a_i(T)$ — сигнальный компонент. Входной шум $n(T)$ при этом является случайным гауссовым процессом. Поскольку коррелятор — это *линейное* устройство, выходной шум является случайным гауссовым процессом [2]. Таким образом, можно записать выражение с выхода коррелятора в момент взятия выборки $t = T$:

$$z(T) = a_i(T) + n_0(T) \quad i = 1, 2,$$

где $n_0(T)$ — компонент шума. Для сокращения записи мы иногда будем выражать $z(t)$ как $a_i + n_0$. Компонент шума n_0 — это *гауссова случайная переменная* с нулевым средним; следовательно, $z(T)$ — это гауссова случайная переменная со средним a_1 или a_2, в зависимости от того, была передана двоичная единица или двоичный нуль.

4.3.2.1. Порог двоичного решения

На рис. 4.9 для случайной переменной $z(T)$ показаны две плотности условных вероятностей — $p(z|s_1)$ и $p(z|s_2)$ — со средними значениями a_1 и a_2. Эти функции, именуемые *правдоподобием* s_1 и *правдоподобием* s_2, были представлены в разделе 3.1.2. Приведем их повторно.

$$p(z|s_1) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z-a_1}{\sigma_0}\right)^2\right] \tag{4.18,а}$$

и

$$p(z|s_2) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z-a_2}{\sigma_0}\right)^2\right] \tag{4.18,б}$$

Здесь σ_0^2 — дисперсия шума. На рис. 4.9 правое правдоподобие $p(z|s_1)$ иллюстрирует вероятностное распределение сигналов на выходе детектора $z(T)$ при переданном сигнале s_1. Подобным образом левое правдоподобие $p(z|s_2)$ демонстрирует вероятностное распределение сигналов на выходе детектора $z(T)$ при переданном сигнале s_2. Абсцисса $z(T)$ представляет полный диапазон возможных значений выборок на выходе корреляционного приемника, показанного на рис. 4.8.

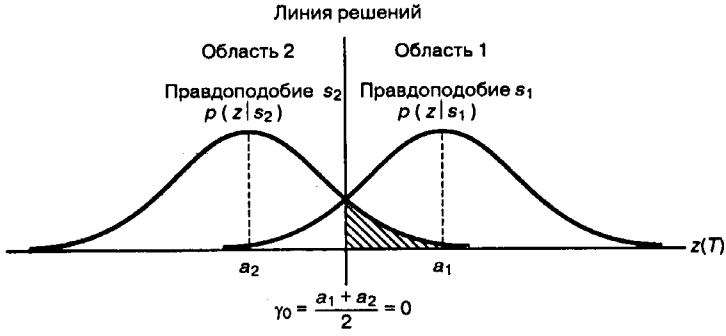


Рис. 4.9. Плотности условных вероятностей $p(z|s_1)$ и $p(z|s_2)$

При рассмотрении задачи оптимизации порога двоичного решения относительно принадлежности принятого сигнала к одной из двух областей, в разделе 3.2.1 было показано, что критерий *минимума ошибки* для равновероятных двоичных сигналов, искаженных гауссовым шумом, можно сформулировать следующим образом.

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} \frac{a_1 + a_2}{2} = \gamma_0 \tag{4.19}$$

Здесь a_1 — сигнальный компонент $z(T)$ при передаче $s_1(t)$, а a_2 — сигнальный компонент $z(T)$ при передаче $s_2(t)$. Порог γ_0, равный $(a_1 + a_2)/2$, — это *оптимальный порог* для минимизации вероятности принятия неверного решения при равновероятных сигналах и симметричных правдоподобиях. Правило принятия решения, приведенное в формуле (4.19), указывает, что гипотеза H_1 (решение, что переданный сигнал — это $s_1(t)$) выбирается при $z(T) > \gamma_0$, а гипотеза H_2 (решение, что переданный сигнал — это $s_1(t)$) — при $z(T) < \gamma_0$. Если $z(T) = \gamma$, решение может быть любым. При равновероятных антиподных сигналах с равными энергиями, где $s_1(t) = -s_2(t)$ и $a_1 = -a_2$, оптимальное правило принятия решения принимает следующий вид.

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} \gamma_0 = 0 \qquad (4.20,а)$$

или

$$\begin{array}{ll} \text{выбрать сигнал } s_1(t), & \text{если } z_1(T) > z_2(T) \\ \text{выбрать сигнал } s_2(t) & \text{в противном случае} \end{array} \qquad (4.20,б)$$

4.4. Когерентное обнаружение

4.4.1. Когерентное обнаружение сигналов PSK

На рис. 4.7 показан детектор, который может использоваться для когерентного обнаружения любого цифрового сигнала. Подобный корреляционный детектор часто называется *детектором, работающим по критерию максимального правдоподобия* (maximum likelihood detector). Рассмотрим следующую бинарную модуляцию PSK (BPSK). Пусть

$$s_1(t) = \sqrt{\frac{2E}{T}} \cos(\omega_0 t + \phi) \quad 0 \le t \le T \qquad (4.21,а)$$

$$s_2(t) = \sqrt{\frac{2E}{T}} \cos(\omega_0 t + \phi + \pi) =$$

$$= -\sqrt{\frac{2E}{T}} \cos(\omega_0 t + \phi) \quad 0 \le t \le T \qquad (4.21,б)$$

и

$n(t)$ — случайный белый гауссов процесс с нулевым средним.

Здесь фазовый член ϕ — произвольная константа, которую мы для удобства положим равной нулю. Параметр E — это энергия сигнала, приходящаяся на символ, а T — длительность символа. Для данного антиподного случая требуется единственная базисная функция. Используя формулы (3.10) и (3.11) и предполагая пространство ортонормированным (т.е. $K_j = 1$), базисную функцию $\psi_1(t)$ можно выразить следующим образом.

$$\psi_1(t) = \sqrt{\frac{2}{T}} \cos \omega_0 t \quad \text{для } 0 \le t \le T \qquad (4.22)$$

Следовательно, переданный сигнал $s_i(t)$ можно выразить через функцию $\psi_1(t)$ и коэффициенты $a_{i1}(t)$.

$$s_i(t) = a_{i1}\psi_1(t) \tag{4.23,а}$$

$$s_1(t) = a_{11}\psi_1(t) = \sqrt{E}\psi_1(t) \tag{4.23,б}$$

$$s_2(t) = a_{21}\psi_1(t) = -\sqrt{E}\psi_1(t) \tag{4.23,в}$$

Предположим, что был передан сигнал $s_1(t)$. Тогда математические ожидания на выходах интеграторов произведений, изображенных на рис. 4.7, *б*, при опорном сигнале $\psi_1(t)$ имеют следующий вид.

$$\mathbf{E}\{z_1|s_1\} = \mathbf{E}\left\{\int_0^T \sqrt{E}\psi_1^2(t) + n(t)\psi_1(t)\,dt\right\} \tag{4.24,а}$$

$$\mathbf{E}\{z_2|s_1\} = \mathbf{E}\left\{\int_0^T -\sqrt{E}\psi_1^2(t) + n(t)\psi_1(t)\,dt\right\} \tag{4.24,б}$$

$$\mathbf{E}\{z_1|s_1\} = \mathbf{E}\left\{\int_0^T \frac{2}{T}\sqrt{E}\cos^2\omega_0 t + n(t)\sqrt{\frac{2}{T}}\cos\omega_0 t\,dt\right\} = \sqrt{E} \tag{4.25,а}$$

и

$$\mathbf{E}\{z_2|s_1\} = \mathbf{E}\left\{\int_0^T -\frac{2}{T}\sqrt{E}\cos^2\omega_0 t + n(t)\sqrt{\frac{2}{T}}\cos\omega_0 t\,dt\right\} = -\sqrt{E} \tag{4.25,б}$$

Здесь $\mathbf{E}\{\cdot\}$ обозначает среднее по ансамблю, так называемое *математическое ожидание* (expected value). В уравнении (4.25) $\mathbf{E}\{n(t)\} = 0$. На этапе принятия решения, путем определения местоположения переданного сигнала в сигнальном пространстве, необходимо определить значение данного сигнала. В приведенном примере, где в качестве базисной функции была взята $\psi_1(t) = \sqrt{2/T}\cos\omega_0 t$, значения $\mathbf{E}\{z_i(T)\}$ равны $\pm\sqrt{E}$. Сигналы-прототипы $\{s_i(t)\}$ аналогичны опорным сигналам $\{\psi_j(t)\}$, с точностью до нормирующего множителя. На этапе принятия решения выбирается сигнал с большим значением $z_i(T)$. Следовательно, в приведенном выше примере принятый сигнал определен как $s_1(t)$. Вероятность ошибки при подобном когерентном обнаружении сигналов BPSK рассмотрена в разделе 4.7.1.

4.4.2. Цифровой согласованный фильтр

В разделе 3.2.2 рассматривалась основная особенность согласованного фильтра — то, что его импульсная характеристика представляет собой запаздывающую версию зеркального отображения (поворота относительно оси $t = 0$) входного сигнала. Таким образом, если сигнал равен $s(t)$, его зеркальное отображение имеет вид $s(-t)$, а зеркальное отображение, запаздывающее на T секунд, имеет вид $s(T - t)$. Следовательно, импульсная характеристика $h(t)$, соответствующая сигналу $s(t)$, будет равна следующему.

$$h(t) = \begin{cases} s(T-t) & 0 \le t \le T \\ 0 & \text{для других } t \end{cases} \tag{4.26}$$

На рис. 4.7 и 4.8 представлена основная функция коррелятора — интегрирование произведения принятого зашумленного сигнала с каждым опорным сигналом и определение наилучшего соответствия. Схемы, показанные на этих рисунках, подразумевают использование аналоговой аппаратуры (умножителей и интеграторов) и непрерывных сигналов. На них не отражена возможность реализации коррелятора или согласованного фильтра (matched filter — MF) с использованием цифровых технологий и дискретных сигналов. Пример подобной реализации приведен на рис. 4.10, где показан согласованный фильтр, использующий цифровую аппаратуру. Входной сигнал $r(t)$ состоит из сигнала-прототипа $s_i(t)$ и шума $n(t)$; ширина полосы сигнала $W = 1/2T$, где T — длительность передачи символа. Таким образом, минимальная частота дискретизации по Найквисту равна $f_s = 2W = 1/T$, а время взятия выборки (T_s) должно быть не больше времени передачи символа. Другими словами, на символ должно приходиться не менее одной выборки. В реальных системах подобная дискретизация производится с частотой, в 4 или более раз превышающей минимальную частоту Найквиста. Платой за это является не увеличение полосы передачи, а увеличение быстродействия процессора. В моменты $t = kT_s$ выборки (как показано на рис. 4.10, *а*) сдвигаются в регистре, так что более ранние из них располагаются правее. При дискретизации (взятии выборки) полученного сигнала непрерывное время t заменяется дискретным kT_s или просто k, что дает право использовать дискретную запись.

$$r(k) = s_i(k) + n(k) \qquad i = 1, 2 \qquad k = 0, 1, \ldots$$

Здесь индекс i определяет символ из M-арного набора (в нашем случае — двоичного), а k — дискретное время. На рис. 4.10 согласованный фильтр аппроксимируется регистром сдвига с весовыми коэффициентами $c_i(n)$, где $n = 0, \ldots, N-1$ — временной индекс весовых коэффициентов и разрядов регистра. В приведенном примере число разрядов регистра и количество выборок на символ равны 4. Итак, суммирование, показанное на рисунке, происходит в моменты времени от $n = 0$ до $n = 3$. Из расположения сумматора на схеме понятно, что решение относительно значения принятого сигнала принимается после заполнения регистра 4 выборками. Отметим, что для простоты в примере на рис. 4.10, *б* выборки $s_i(k)$ могут принимать только три значения (0, ±1). В реальных системах каждая выборка (и весовой коэффициент) — это 6–10 бит. Множеству весовых коэффициентов фильтра $\{c_i(n)\}$ соответствует импульсная характеристика фильтра; согласование весовых коэффициентов с выборками сигнала производится согласно дискретному варианту уравнения (4.26).

$$c_i(n) = s_i[(N-1) - n] = s_i(3 - n) \tag{4.27}$$

Использование дискретной формы *интеграла свертки* из уравнения (А.44,б) позволяет записать выражение с выхода коррелятора в момент времени, соответствующий k-й выборке.

$$z_i(k) = \sum_{n=0}^{N-1} r(k-n)\, c_i(n) \qquad k = 0, 1, \ldots, \text{ по модулю } N \tag{4.28}$$

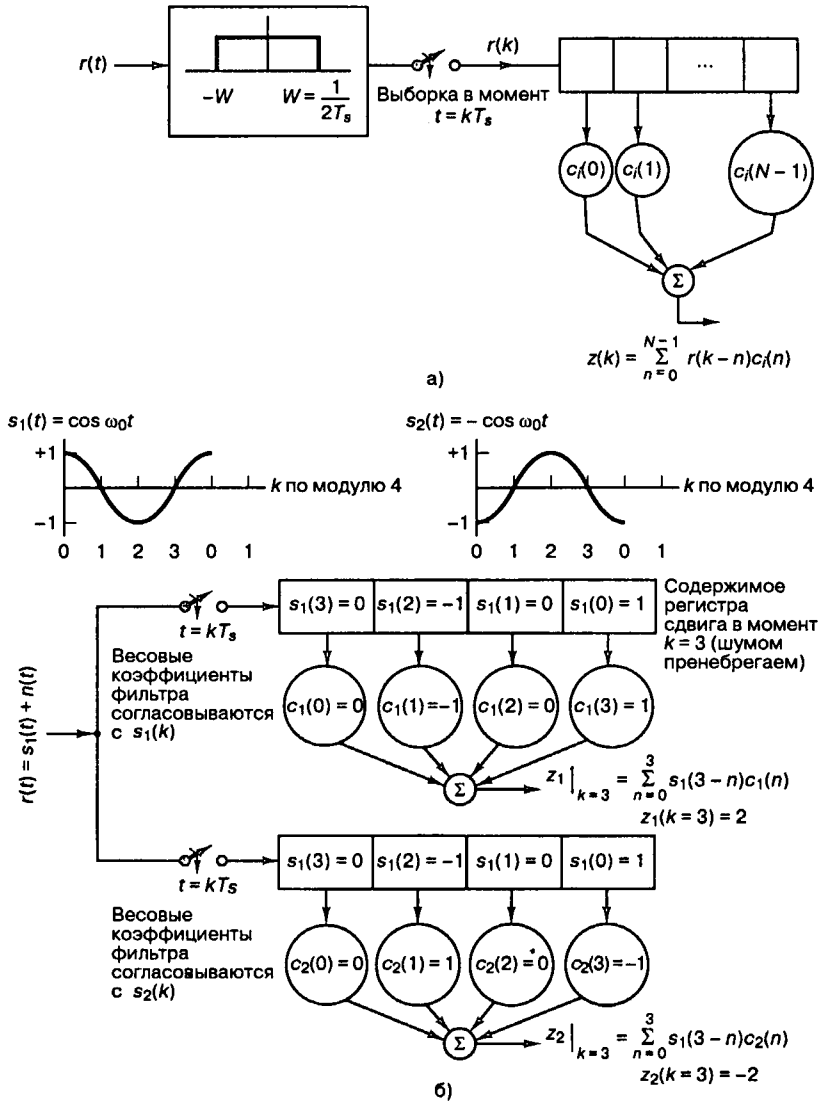


Рис. 4.10. Цифровой согласованный фильтр: а) дискретный согласованный фильтр; б) пример обнаружения с использованием дискретного согласованного фильтра (шумом пренебрегаем)

Здесь x по модулю y — это остаток деления x на y, индекс k — время принятия выборок и выхода фильтра, а n — фиктивная переменная времени. В формуле (4.28) выражение $r(k-n)$ содержит n, которое можно рассматривать как "возраст" выборки (как давно она находится в фильтре). В выражении $c_i(n)$ n удобно рассматривать как адрес весового коэффициента. Предполагается, что система синхронизирована и упорядочение символов во времени известно. Также предполагается, что шум имеет нулевое среднее, так что математическое ожидание принятой выборки равно следующему.

$$\mathrm{E}\{r(k)\} = s_i(k) \qquad i = 1, 2$$

Следовательно, при передаче $s_i(t)$ математическое ожидание выхода согласованного фильтра равно следующему.

$$\mathbf{E}\{z_i(k)\} = \sum_{n=0}^{N-1} s_i(k-n)\, c_i(n) \qquad k = 0, 1, \ldots, \text{по модулю } N \tag{4.29}$$

На рис. 4.10, *б*, где сигналы-прототипы изображены как функции времени, видим, что крайняя слева выборка (амплитуда, равная +1) графика $s_1(t)$ представляет выборку в момент времени $k = 0$. Предполагая, что передан был сигнал $s_1(t)$ и для упрощения записи мы пренебрегли шумом, можем записать принятую выборку $r(k)$ как $s_1(k)$. Выборки заполняют разряды согласованного фильтра, и в конце каждого периода передачи символа в крайнем правом разряде каждого регистра расположена выборка $k = 0$. Отметим, что в формулах (4.28) и (4.29) временны*е* индексы n эталонных весовых коэффициентов расположены в порядке, обратном к временно*м*у индексу $k - n$ выборок, что является ключевой особенностью интеграла свертки. То, что наиболее ранняя выборка теперь соответствует крайнему справа весовому коэффициенту, обеспечивает значащую корреляцию. Даже если действия согласованного фильтра мы математически опишем как *свертку* сигнала с импульсной характеристикой фильтра, конечный результат будет *корреляцией* сигнала с копией самого себя. По этой причине коррелятор можно реализовать как согласованный фильтр.

На рис. 4.10, *б* обнаружение, происходящее после выхода сигнала с согласованного фильтра, осуществляется обычным образом. Для принятия двоичного решения выходы $z_i(k)$ изучаются при каждом значении $k = N - 1$, соответствующем концу символа. При условии передачи $s_1(t)$ и пренебрежении шумом, уравнения (4.27)–(4.29) можно объединить и записать выходы коррелятора в моменты времени $k = N - 1 = 3$.

$$z_1(k=3) = \sum_{n=0}^{3} s_1(3-n)\, c_1(n) = 2 \tag{4.30,а}$$

и

$$z_2(k=3) = \sum_{n=0}^{3} s_1(3-n)\, c_2(n) = -2 \tag{4.30,б}$$

Поскольку $z_1(k=3)$ больше $z_2(k=3)$, детектор принимает решение, что передан был символ $s_1(t)$.

Может возникнуть вопрос: *чем согласованный фильтр на рис. 4.10, б отличается от коррелятора на рис. 4.8.* В случае согласованного фильтра в ответ на каждую новую выборку на входе появляется новое значение на выходе; следовательно, выход представляет собой временной ряд, такой как на рис. 3.7, *б* (последовательность возрастающих положительных и отрицательных корреляций с входной синусоидой). Подобную последовательность на выходе согласованного фильтра можно получить при использовании нескольких корреляторов, работающих на разных начальных точках входящего временно*г*о ряда. Отметим, что за время передачи символа на выходе коррелятора получаем максимальное значение сигнала в момент времени T (см. рис. 3.7, *б*). Если синхронизировать согласованный фильтр и коррелятор, их выходы в конце периода передачи символа бу-

дут идентичными. Важным отличием между согласованным фильтром и коррелятором является то, что поскольку на выходе коррелятора получаем одно значение на символ, он должен использовать дополнительную информацию, например, относительно моментов начала и завершения интегрирования произведения. При наличии ошибок синхронизации дискретный сигнал, подаваемый с коррелятора на детектор, может быть сильно искажен. С другой стороны, поскольку на выходе согласованного фильтра получаем *временной ряд* выходных значений (отражающих смещенные во времени входящие выборки, умноженные на фиксированные весовые коэффициенты), использование дополнительной схемы позволяет определить моменты, наиболее подходящие для дискретизации выхода согласованного фильтра.

Пример 4.1. Цифровой согласованный фильтр

Рассмотрим набор сигналов

$$s_1(t) = At \quad 0 \le t \le kT$$

и

$$s_2(t) = -At \quad 0 \le t \le kT,$$

где $k = 0, 1, 2, 3$.

Опишите, как *цифровой* согласованный фильтр (рис. 4.10) может использоваться для обнаружения принятого сигнала, скажем $s_1(t)$, при отсутствии шума.

Решение

Вначале сигнал $s_1(t)$ преобразуется в набор выборок $\{s_1(k)\}$. Приемник цифрового согласованного фильтра, как показано на рис. 4.10, *б*, представляет собой две ветви. Верхняя ветвь состоит из регистра сдвига и коэффициентов, согласовывающихся с точками дискретизации $\{s_1(k)\}$. Подобным образом нижняя ветвь состоит из регистра сдвига и коэффициентов, согласовывающихся с точками дискретизации $\{s_2(k)\}$. В четырех равномерно расположенных точках выборки ($k = 0, 1, 2, 3$) сигналы $\{s_i(k)\}$ имеют следующие значения.

$$s_1(k=0) = 0 \quad s_1(k=1) = A/4 \quad s_1(k=2) = A/2 \quad s_1(k=3) = 3A/4$$

$$s_2(k=0) = 0 \quad s_2(k=1) = -A/4 \quad s_2(k=2) = -A/2 \quad s_2(k=3) = -3A/4$$

Коэффициенты $c_i(n)$ представляют запаздывающий зеркальный поворот сигнала, с которым согласовывается фильтр. Следовательно, $c_i(n) = s_i(N-1-n)$, где $n = 0, \ldots, N-1$, так что можно записать $c_1(0) = s_1(3)$, $c_1(1) = s_1(2)$, $c_1(2) = s_1(1)$, $c_1(3) = s_1(0)$.

Рассмотрим верхнюю ветвь рис. 4.10, *б*. В момент времени $k = 0$ первая выборка $s_1(k=0) = 0$ поступает в крайний левый разряд каждого регистра. В следующий дискретный момент времени $k = 1$ вторая выборка $s_1(k=1) = A/4$ поступает в крайний левый разряд каждого регистра; в то же время первая выборка сдвигается в ближайший справа разряд каждого регистра и т.д. В момент $k = 3$ в крайний левый разряд поступает выборка $s_1(k=3) = 3A/4$; к этому моменту первая выборка сдвинута к крайнему правому разряду. Четыре выборки сигнала теперь расположены в регистрах в зеркальном порядке по отношению к времени их создания. Таким образом, при данном расположении поступающих выборок сигнала и опорных коэффициентов выход сумматора естественным образом описывается операцией свертки и максимизирует корреляцию в соответствующей ветви.

4.4.3. Когерентное обнаружение сигналов MPSK

На рис. 4.11 показан вид сигнального пространства для набора сигналов в модуляции MPSK (multiple phase-shift keying — многофазная манипуляция); на рисунке представлена четырехуровневая ($M = 4$) фазовая манипуляция, или двукратная фазовая манипуляция (quadriphase shift keying — QPSK). Двоичные цифры в передатчике группиру-

ются по две, и в каждом интервале передачи символов две последовательные цифры определяют, какой из четырех возможных сигналов произведет модулятор. Для типичных когерентных M-уровневых систем PSK (MPSK) сигнал $s_i(t)$ можно выразить следующим образом.

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos\left(\omega_0 t - \frac{2\pi i}{M}\right) \quad 0 \le t \le T \qquad (4.31)$$
$$i = 1, \ldots, M$$

Здесь E — энергия, полученная сигналом за время передачи символа T, а ω_0 — несущая частота. Предполагая пространство ортонормированным и используя формулы (3.10) и (3.11), можно выбрать следующие удобные оси.

$$\psi_1(t) = \sqrt{\frac{2}{T}} \cos\omega_0 t \qquad (4.32,а)$$

и

$$\psi_2(t) = \sqrt{\frac{2}{T}} \sin\omega_0 t \qquad (4.32,б)$$

Здесь, как и в разделе 4.4.1, амплитуда $\sqrt{2/T}$ нормирует ожидаемый выход детектора.

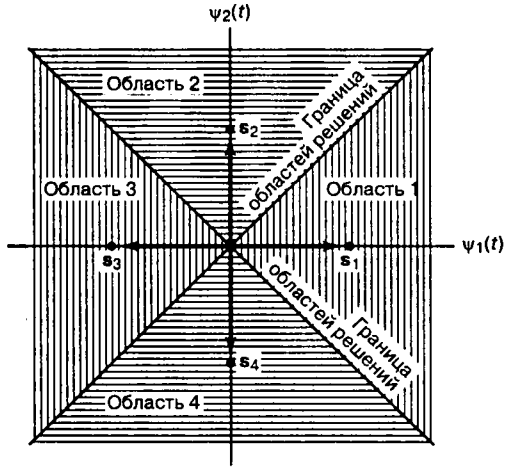


Рис. 4.11. Сигнальное пространство и области решений для системы QPSK

Запишем сигнал $s_i(t)$ через выбранные ортонормированные координаты.

$$s_i(t) = a_{i1}\psi_1(t) + a_{i2}\psi_2(t) \quad 0 \le t \le T \qquad (4.33,а)$$
$$i = 1, \ldots, M$$

$$= \sqrt{E} \cos\left(\frac{2\pi i}{M}\right) \psi_1(t) + \sqrt{E} \sin\left(\frac{2\pi i}{M}\right) \psi_2(t) \qquad (4.33,б)$$

Отметим, что формула (4.33) выражает набор M многофазных сигналов (в общем случае не ортогональный) всего через два ортогональных несущих компонента. Случай $M=4$ (QPSK) является уникальным среди множества сигналов MPSK в том смысле, что сигналы QPSK представляются комбинацией антиподных и ортогональных членов. Границы областей решений разбивают сигнальное пространство на $M=4$ области; процедура разбития подобна описанной в разделе 4.3.1 и изображенной на рис. 4.6 для $M=2$. Правило принятия решения для детектора (рис. 4.11) звучит следующим образом: если вектор принятого сигнала попадает в область 1 — отнести его к $s_1(t)$; если вектор принятого сигнала попадает в область 2 — выбрать сигнал $s_2(t)$ и т.д. Другими словами, правило принятия решения заключается в выборе i-го сигнала, если $z_i(T)$ является наибольшим из выходов корреляторов (см. рис. 4.7).

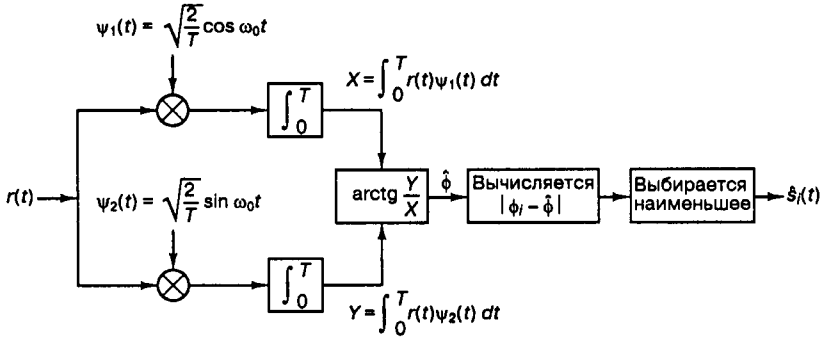


Рис. 4.12. Демодулятор сигналов MPSK

Структура коррелятора, изображенного на рис. 4.7, *а*, подразумевает использование для демодуляции сигналов MPSK M корреляторов произведений. Также предполагается, что для каждой из M ветвей был соответствующим образом выбран опорный сигнал (т.е. сигнал, имеющий требуемый сдвиг фаз). Стоит отметить, что на практике реализация демодулятора MPSK, согласно схеме на рис. 4.7, *б*, требует всего $N=2$ интеграторов произведений, вне зависимости от размера множества сигналов M. Такая экономия позволительна вследствие того, что, как показано в разделе 3.1.3, любой произвольный интегрируемый набор сигналов можно выразить в виде линейной комбинации ортогональных сигналов. Пример подобного демодулятора приведен на рис. 4.12. Объединив формулы (4.32) и (4.33), можно записать принятый сигнал $r(t)$ следующим образом.

$$r(t) = \sqrt{\frac{2E}{T}}(\cos\phi_i \cos\omega_0 t + \sin\phi_i \sin\omega_0 t) + n(t) \quad 0 \le t \le T \qquad (4.34)$$

$$i = 1, \ldots, M$$

Здесь $\phi_i = 2\pi i/M$, а $n(t)$ — гауссов процесс шума с нулевым средним. Отметим, что на рис. 4.12 изображены только два опорных сигнала (или две базисные функции) — $\psi_1(t) = \sqrt{2/T}\cos\omega_0 t$ для верхнего коррелятора и $\psi_2(t) = \sqrt{2/T}\sin\omega_0 t$ для нижнего. Верхний коррелятор вычисляет функцию

$$X = \int_0^T r(t)\psi_1(t)\,dt, \qquad (4.35)$$

а нижний — функцию

$$Y = \int_0^T r(t)\psi_2(t)\,dt. \tag{4.36}$$

На рис. 4.13 показано, что определение фазы принятого сигнала $\hat{\phi}$ производится путем вычисления арктангенса Y/X, где X — синфазный, Y — квадратурный компонент принятого сигнала, а $\hat{\phi}$ — зашумленная оценка переданной фазы ϕ_i. Другими словами, с верхнего коррелятора (рис. 4.12) поступает на выход X, значение синфазной проекции вектора **r**, а с нижнего — Y, значение квадратурной проекции вектора **r**, где **r** — векторное представление $r(t)$. Сигналы X и Y с корреляторов поступают в блок "arctg (Y/X)". Полученное значение фазы $\hat{\phi}$ сравнивается с каждой фазой-прототипом ϕ_i. Далее демодулятор выбирает фазу ϕ_i, ближайшую к $\hat{\phi}$. Другими словами, демодулятор вычисляет $|\phi_i - \hat{\phi}|$ для каждого прототипа ϕ_i и выбирает ϕ_i, дающую наименьший выход.

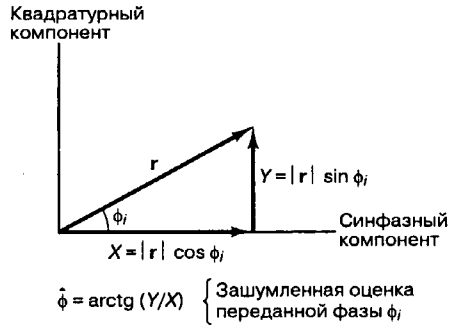


Рис. 4.13. Синфазный и квадратурный компоненты вектора принятого сигнала **r**

4.4.4. Когерентное обнаружение сигналов FSK

При использовании схемы FSK информация модулируется частотой несущей. Типичный вид набора сигналов FSK выражается формулой (4.8)

$$s_i(t) = \sqrt{\frac{2E}{T}}(\cos\omega_i t + \phi) \quad 0 \le t \le T$$

$$i = 1, \ldots, M,$$

где E — энергия, переданная сигналу $s_i(t)$ в течение времени передачи символа T; кроме того, $(\omega_{i+1} - \omega_i)$ обычно выбирается кратным π/T. Фазовый член ϕ — это произвольная константа, которую можно положить равной нулю. Предполагая, что базисные функции $\psi_1(t)$, $\psi_2(t)$, ..., $\psi_M(t)$ формируют ортонормированное множество, можно получить более удобное выражение для $\{\psi_i(t)\}$.

$$\psi_j(t) = \sqrt{\frac{2}{T}} \cos \omega_j t \quad j = 1, \ldots, N \tag{4.37}$$

Здесь, как и выше, амплитуда $\sqrt{2/T}$ нормирует ожидаемый выход согласованного фильтра. Используя уравнение (3.11), можно записать следующее.

$$a_{ij} = \int_0^T \sqrt{\frac{2E}{T}} \cos(\omega_i t) \sqrt{\frac{2}{T}} \cos(\omega_j t)\, dt \tag{4.38}$$

Следовательно,

$$a_{ij} = \begin{cases} \sqrt{E} & \text{для } i = j \\ 0 & \text{для других } i, j \end{cases} \tag{4.39}$$

Другими словами, i-й вектор сигнала-прототипа расположен на i-й координатной оси на расстоянии $\sqrt{E}$ от начала координат сигнального пространства. В этой схеме, при данном числе уровней M и данной E, расстояние между любыми двумя векторами сигналов-прототипов $\mathbf{s}_i$ и $\mathbf{s}_j$ является постоянным.

$$d(\mathbf{s}_i, \mathbf{s}_j) = \|\mathbf{s}_i - \mathbf{s}_j\| = \sqrt{2E} \text{ для } i \neq j \tag{4.40}$$

На рис. 4.14 показаны векторы сигналов-прототипов и области решений для троичной ($M = 3$) ортогональной модуляции FSK с когерентным обнаружением. Как правило, естественным выбором размера M сигнального множества является степень двойки. Причина неортодоксального выбора $M = 3$ состоит в том, что мы желаем исследовать сигнальное множество, большее чем бинарное, а визуальное представление сигнального пространства лучше всего выглядит при использовании взаимно перпендикулярных осей. Наибольшим числом перпендикулярных осей, которые можно аккуратно изобразить визуально, является 3. Как и при использовании модуляции PSK, сигнальное пространство разбивается на M различных областей, каждая из которых содержит один вектор сигнала-прототипа; в нашем примере, где области решений являются трехмерными, границы областей являются уже не линиями, а плоскостями. Оптимальное правило принятия решения состоит в следующем: определить сигнал к тому классу, индекс которого соответствует области нахождения принятого сигнала. На рис. 4.14 вектор принятого сигнала $\mathbf{r}$ изображен в области 2. Согласно приведенному выше правилу принятия решений, детектор классифицирует $\mathbf{r}$ как сигнал $\mathbf{s}_2$. Поскольку шум изображается гауссовым случайным вектором, существует отличная от нуля вероятность того, что вектор $\mathbf{r}$ даст сигнал, отличный от $\mathbf{s}_2$. Например, если передатчик послал сигнал $\mathbf{s}_2$, вектор $\mathbf{r}$ будет суммой сигнала и шума $\mathbf{s}_2 + \mathbf{n}_a$, а решение о выборе $\mathbf{s}_2$ будет справедливым; в то же время, если передатчик в действительности послал сигнал $\mathbf{s}_2$, вектор $\mathbf{r}$ будет суммой сигнала и шума $\mathbf{s}_3 + \mathbf{n}_b$, а решение относительно выбора $\mathbf{s}_2$ будет ошибочным. Вопросы вероятности возникновения ошибки при когерентном обнаружении сигналов в модуляции FSK подробно рассмотрены в разделе 4.7.3.

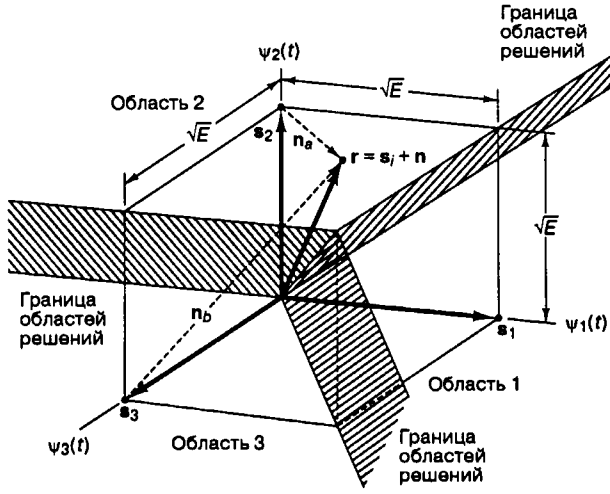


Рис. 4.14. Разбиение сигнального пространства для троичного сигнала FSK

Пример 4.2. Принятая фаза как функция задержки распространения

а) Из схемы, приведенной на рис. 4.8, непонятно, откуда берутся опорные сигналы коррелятора. Кто-то может подумать, что они известны всегда и хранятся в памяти, пока не понадобятся. При некоторых обстоятельствах приемник действительно может, в разумных пределах, предсказывать некоторое ожидаемое значение амплитуды поступающего сигнала или его частоты. Но существует один параметр, который нельзя оценить без специальной помощи, — это фаза принятого сигнала. Наиболее популярным способом получения оценки фазы является использование схемы, называемой *контуром фазовой автоподстройки частоты* (ФАПЧ, phase-locked loop — PLL). Схема восстановления несущей захватывает прибывающую несущую волну (или воссоздает ее) и оценивает ее фазу. Чтобы показать, как иногда нереально предсказать фазу без использования ФАПЧ, рассмотрим канал радиосвязи, изображенный на рис. 4.15. Здесь мобильный пользователь расположен в точке A на расстоянии d от центральной станции, а задержка распространения сигнала равна T_d. Используя комплексную форму записи, можем описать сигнал, излучаемый передатчиком, как $s(t) = \exp(2\pi i f_0 t)$. Пусть частота f_0 равна 1 ГГц. Если пренебречь шумом, сигнал, принятый центральной станцией, можно записать как $r(t) = \exp[2\pi i f_0(t + T_d)]$. Определите, на какое минимальное расстояние d (рис. 4.15) должен переместиться мобильный пользователь, чтобы это привело к изменению фазы принятого сигнала на 2π.

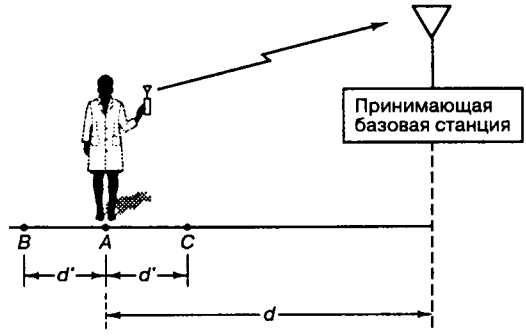


Рис. 4.15. Канал радиосвязи

б) Действительно ли нас волнует изменение фазы на 2π? Разумеется, нет, поскольку в этом случае вектор принятого сигнала будет находиться в той же точке, что и ранее, когда пользователь находился в точке A. Но зададимся вопросом, чему равно минимальное расстояние, изменяющее фазу на $\pi/2$ (скажем, дающее запаздывание на $\pi/2$)? Приемник должен отнести вектор, соответствующий $r(t)$, к той же группе, что и в п. а, но запаздывание приводит к тому, что принятый сигнал уже имеет вид $r(t) = \exp[2\pi i f_0(t + T_d) - \pi/2]$, и коррелятор, используемый в процессе обнаружения, дает нулевой выход.

$$\int_0^T \cos\omega_0 t \cos\left(\omega_0 t - \frac{\pi}{2}\right) dt = \int_0^T \cos\omega_0 t \sin\omega_0 t \, dt = 0$$

Определите минимальное расстояние перемещения пользователя, приводящее к изменению фазы на $\pi/2$.

Решение

а) Пусть в начальный момент времени $t = 0$ мобильный пользователь находится в точке A, так что вектор, принятый центральной станцией, дается выражением $r(t) = \exp(2\pi i f_0 T_d)$. Затем, после перемещения пользователя в точку B, принятый (еще сильнее запаздывающий) вектор $r_d(t = T_d + T_d')$ можно записать в виде $r_d(t) = \exp[2\pi i f_0(T_d + T_d')]$. Минимальное время задержки T_d', соответствующее повороту вектора на 2π, равно $T_d' = 1/f_0 = 10^{-9}$ секунд. Следовательно, минимальное расстояние для такого поворота (предполагая идеальное электромагнитное распространение со скоростью света) равно следующему.

$$d' = \frac{c}{f_0} = 3 \times 10^8 \text{ м/с} \times 10^{-9} \text{ с} = 0{,}3 \text{ м}$$

б) Используя предыдущий результат, получаем следующее расстояние для поворота вектора на $\pi/2$.

$$d'' = \frac{d'}{4} = \frac{0{,}3 \text{ м}}{4} = 7{,}5 \text{ см}$$

Очевидно, что даже если передатчик и приемник жестко установлены на стационарных башнях, небольшое смещение, вызванное ветром, может привести к абсолютной неопределенности относительно значения фазы. Если предположить, что используемая частота равна не 1 ГГц, а 10 ГГц, то минимальное расстояние изменяется с 7,5 см до 0,75 см. На практике зачастую желательно избегать приемников, использующих ФАПЧ. Вычисления, выполненные в данном примере, могут породить вопрос, как изменится вероятность ошибки, если в процессе обнаружения не будет использоваться информация о фазе? Другими словами, чем заплатит система, если обнаружение будет выполнено некогерентно? Этот и другие подобные вопросы рассматриваются в следующем разделе.

4.5. Некогерентное обнаружение

4.5.1. Обнаружение сигналов в дифференциальной модуляции PSK

Название *дифференциальная фазовая манипуляция* (differential phase-shift keying — DPSK) иногда требует некоторого пояснения, поскольку со словом "дифференциальный" связано два различных аспекта процесса модуляции/демодуляции: процедура кодирования и процедура обнаружения. Термин "дифференциальное кодирование" употребляется тогда, когда кодировка двоичных символов определяется не их значением (т.е. нуль или единица), а тем, совпадает ли символ с предыдущим или отличается от него. Термин "дифференциаль-

ное когерентное обнаружение" сигналов в дифференциальной модуляции PSK (именно в этом значении обычно используется название DPSK) связан со схемой обнаружения, которая зачастую относится к некогерентным схемам, поскольку не требует согласования по фазе с принятой несущей. Стоит отметить, что дифференциально кодированные сигналы PSK иногда обнаруживаются *когерентно*. Эта возможность будет рассмотрена в разделе 4.7.2.

В некогерентных системах не предпринимаются попытки определить действительное значение фазы поступающего сигнала. Следовательно, если переданный сигнал имеет вид

$$s_i(t) = \sqrt{\frac{2E}{T}}(\cos\omega_0 t + \phi) \quad 0 \le t \le T$$

$$i = 1, \ldots, M,$$

то принятый сигнал можно описать следующим образом.

$$r(t) = \sqrt{\frac{2E}{T}}\cos[\omega_0 t + \theta_i(t) + \alpha] + n(t) \quad 0 \le t \le T \qquad (4.41)$$

$$i = 1, \ldots, M$$

Здесь α — произвольная константа, обычно предполагаемая случайной переменной, равномерно распределенной между нулем и 2π, а $n(t)$ — процесс AWGN.

Для когерентного обнаружения используются согласованные фильтры (или их эквиваленты); для некогерентного обнаружения подобное невозможно, поскольку в этом случае выход согласованного фильтра будет зависеть от неизвестного угла α. Но если предположить, что α меняется медленно относительно интервала в два периода (2*T*), то разность фаз между двумя последовательными сигналами $\theta_j(T_1)$ и $\theta_k(T_2)$ не будет зависеть от α.

$$[\theta_k(T_2) + \alpha] - [\theta_j(T_1) + \alpha] = \theta_k(T_2) - \theta_j(T_2) = \phi_i(T_2) \qquad (4.42)$$

Основа дифференциального когерентного обнаружения сигналов в дифференциальной модуляции PSK (DPSK) состоит в следующем. В процессе демодуляции в качестве опорной фазы может применяться фаза несущей предыдущего интервала передачи символа. Ее использование требует *дифференциального кодирования* последовательности сообщений в передатчике, поскольку информация кодируется разностью фаз между двумя последовательными импульсами. Для передачи i-го сообщения ($i = 1, 2, \ldots, M$) фаза текущего сигнала должна быть смещена на $\phi_i = 2\pi i/M$ радиан относительно фазы предыдущего сигнала. Вообще, детектор вычисляет координаты поступающего сигнала путем определения его корреляции с локально генерируемыми сигналами $\sqrt{2/T}\cos\omega_0 t$ и $\sqrt{2/T}\sin\omega_0 t$. Затем, как показано на рис. 4.16, детектор измеряет угол между вектором текущего принятого сигнала и вектором предыдущего сигнала.

Вообще, схема DPSK менее эффективна, чем PSK, поскольку в первом случае, вследствие корреляции между сигналами, ошибки имеют тенденцию к распространению (на соседние времена передачи символов). Стоит помнить, что схемы PSK и DPSK отличаются тем, что в первом случае сравнивается принятый сигнал с идеальным опорным, а во втором — два зашумленных сигнала. Отметим, что модуляция DPSK дает вдвое больший шум, чем модуляция PSK. Следовательно, при использовании DPSK следует ожидать вдвое (на 3 дБ) большей вероятности ошибки, чем в случае PSK; ухудшение качества передачи происходит довольно быстро с уменьшением

отношения сигнал/шум (вопрос достоверности передачи при использовании модуляции DPSK рассмотрен в разделе 4.7.5). Преимуществом схемы DPSK можно назвать меньшую сложность системы.

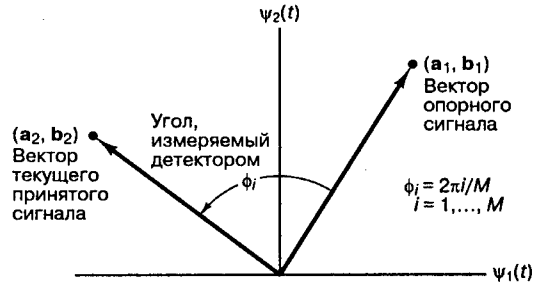


Рис. 4.16. Сигнальное пространство для схемы DPSK

4.5.2. Пример бинарной модуляции DPSK

Суть дифференциального когерентного обнаружения в схеме DPSK состоит в том, что информация из сигнала извлекается путем изменения фазы от символа к символу. Следовательно, переданный сигнал требуется вначале закодировать. На рис. 4.17, *а* представлено дифференциальное кодирование двоичного потока сообщений $m(k)$, где k — индекс дискретизации. Дифференциальное кодирование начинается (третья строка на рисунке) с произвольного выбора первого бита кодовой последовательности $c(k=0)$ (в данном случае выбрана единица). Затем последовательность закодированных битов $c(k)$ может, в общем случае, кодироваться одним из двух способов.

$$c(k) = c(k-1) \oplus m(k) \tag{4.43}$$

или

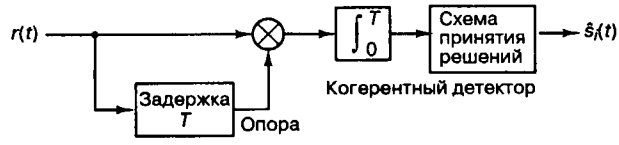
$$c(k) = \overline{c(k-1) \oplus m(k)} \tag{4.44}$$

Здесь символ "⊕" представляет сложение по модулю 2 (определенное в разделе 2.9.3), а черта над выражением означает его дополнение. На рис. 4.17, *а* дифференциальное кодирование сообщения было выполнено с помощью уравнения (4.44). Другими словами, текущий бит кода $c(k)$ равен единице, если бит сообщения $m(k)$ совпадает с предыдущим закодированным битом $c(k-1)$, в противном случае — $c(k)=0$. В четвертой строке рисунка кодированная последовательность битов $c(k)$ преобразовывается в последовательность сдвигов фаз $\theta(k)$, где единица представляется сдвигом фазы на 180°, а нуль — нулевым сдвигом фазы.

На рис. 4.17, *б* в виде блочной диаграммы представлена схема обнаружения бинарных сигналов в модуляции DPSK. Отметим, что основным элементом демодулятора на рис. 4.7 является интегратор произведений; как и при когерентном обнаружении сигналов PSK, мы пытаемся определить корреляцию принятого сигнала с опорным. (Опорный сигнал — это просто запаздывающая версия принятого сигнала.) Другими словами, в течение каждого интервала передачи символа мы согласовываем принятый символ с предыдущим на предмет корреляции или антикорреляции (отличия в фазе на 180°).

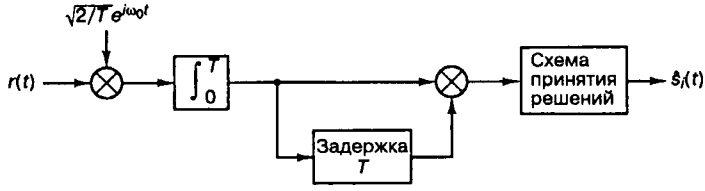
Индекс дискретизации, k	0	1	2	3	4	5	6	7	8	9	10
Информационное сообщение, $m(k)$		1	1	0	1	0	1	1	0	0	1
Сообщение в дифференциальной кодировке (первый бит произвольный), $c(k)$	1	1	1	0	0	1	1	1	0	1	1
Соответствующий сдвиг фаз, $\theta(k)$	π	π	π	0	0	π	π	π	0	π	π

а)



Обнаруженное сообщение, $\hat{m}(k)$ 1 1 0 1 0 1 1 0 0 1

б)



в)

Рис. 4.17. Дифференциальная фазовая манипуляция (DPSK): а) дифференциальное кодирование; б) дифференциальное когерентное обнаружение; в) оптимальное дифференциальное когерентное обнаружение

Пусть при отсутствии шума принятый сигнал с последовательностью сдвигов фаз $\theta(k)$ поступает в коррелятор, изображенный на рис. 4.17, *б*. Фаза $\theta(k=1)$ совпадает с $\theta(k=0)$; обе имеют одинаковое значение, π. Следовательно, первый бит обнаруженного выхода — $\hat{m}(k=1)=1$. Далее $\theta(k=2)$ совпадает с $\theta(k=1)$, и снова имеем то же значение и $\hat{m}(k=2)=1$. Затем $\theta(k=3)$ отличается от $\theta(k=2)$, так что $\hat{m}(k=3)=0$, и т.д.

Необходимо отметить, что детектор, изображенный на рис. 4.17, *б*, является близким к оптимальному [3] в смысле вероятности ошибки. Оптимальный дифференциальный детектор для схемы DPSK требует согласования опорной несущей с принятой несущей по частоте, но не обязательно по фазе. Отсюда — вид оптимального дифференциального детектора, приведенного на рис. 4.17, *в* [4]. Достоверность передачи при использовании такого детектора рассмотрена в разделе 4.7.5. Обратите внимание на то, что опорный сигнал (рис. 4.17, *в*) приведен в комплексной форме записи ($\sqrt{2/T}e^{j\omega_0 t}$); это показывает необходимость квадратурной реализации, использующей квадратурный и синфазный компоненты (см. раздел 4.6.1).

4.5.3. Некогерентное обнаружение сигналов FSK

Детектор, выполняющий *некогерентное* обнаружение сигналов в модуляции FSK, описываемых уравнением (4.8), можно реализовать с помощью корреляторов, подобных показанным на рис. 4.7. При этом оборудование приема следует настроить как *детектор энергии* без измерения фазы. По этой причине некогерентный детектор обычно требует вдвое большего числа ветвей-каналов, чем когерентный. На рис. 4.18 показаны синфазный (I) и квадратурный (Q) каналы, используемые для некогерентного обнаружения набора сигналов в бинарной модуляции FSK (BFSK). Отметим, что две верхние ветви настроены на обнаружение сигнала с частотой ω_1; для синфазной ветви опорный сигнал имеет вид $\sqrt{2/T}\cos\omega_1 t$, а для квадратурной — $\sqrt{2/T}\sin\omega_1 t$. Подобным образом две нижние ветви настроены на обнаружение сигнала с частотой ω_2; для синфазной ветви опорный сигнал имеет вид $\sqrt{2/T}\cos\omega_2 t$, а для квадратурной — $\sqrt{2/T}\sin\omega_2 t$. Предположим, что принятый сигнал $r(t)$ имеет вид точно $\cos\omega_1 t + n(t)$, т.е. фаза точно равна нулю. Следовательно, сигнальный компонент принятого сигнала точно соответствует (по частоте и фазе) опорному сигналу верхней ветви. В такой ситуации максимальный выход должен дать интегратор произведений верхней ветви. Вторая ветвь должна дать нулевой выход (проинтегрированный шум с нулевым средним), поскольку ее опорный сигнал $\sqrt{2/T}\sin\omega_1 t$ ортогонален сигнальному компоненту сигнала $r(t)$. При ортогональной передаче сигналов (см. раздел 4.5.4) третья и четвертая ветви также должны дать выходы порядка нуля, поскольку их опорные сигналы также ортогональны сигнальному компоненту сигнала $r(t)$.

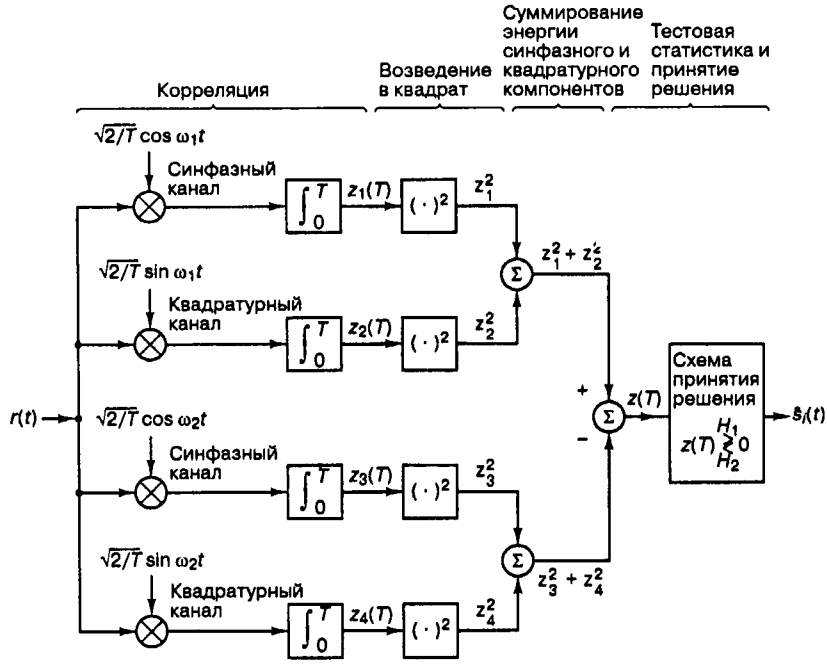


Рис. 4.18. Квадратурный приемник

Рассмотрим теперь другую возможность. Пусть принятый сигнал $r(t)$ имеет вид $\sin \omega_1 t + n(t)$. В этом случае максимальный выход должна дать вторая ветвь схемы (рис. 4.18), а выходы других ветвей должны быть порядка нуля. В реальной системе сигнал $r(t)$ скорее всего описывается выражением $\cos(\omega_1 t + \phi) + n(t)$, т.е. входящий сигнал будет частично коррелировать с опорным сигналом $\cos \omega_1 t$ и частично — с сигналом $\sin \omega_1 t$. Поэтому некогерентный квадратурный приемник ортогональных сигналов и требует синфазной и квадратурной ветви для каждого возможного сигнала набора. Блоки, показанные на рис. 4.18 после интеграторов произведений, выполняют операцию возведения в квадрат, что предотвращает появление возможных отрицательных значений. Затем для каждого класса сигналов набора (в бинарном случае — для двух) складываются величины z_1^2 из синфазного канала и z_2^2 из квадратурного канала. На конечном этапе формируется тестовая статистика $z(T)$ и выбирается сигнал с частотой ω_1 или ω_2, в зависимости от того, какая пара детекторов энергии дала максимальный выход.

Существует еще одна возможная реализация некогерентного обнаружения сигналов FSK. В этом случае используются полосовые фильтры, центрированные на частоте $f_i = \omega_i/2\pi$ с полосой $W_f = 1/T$, за которыми, как показано на рис. 4.19, следуют *детекторы огибающей* (envelope detector). Детектор огибающей состоит из выпрямителя и фильтра нижних частот. Детекторы согласовываются с *огибающими сигнала*, а не с самими сигналами. При определении огибающей фаза несущей не имеет значения. При бинарной FSK решение относительно значения переданного символа принимается путем определения, какой из двух детекторов огибающей дает большую амплитуду на момент измерения. Подобным образом для системы, использующей многочастотную фазовую манипуляцию (multiple frequency shift-keying — MFSK), решение относительно принадлежности переданного символа к одному из M возможных принимается путем определения, какой из M детекторов огибающей дает максимальный выход.

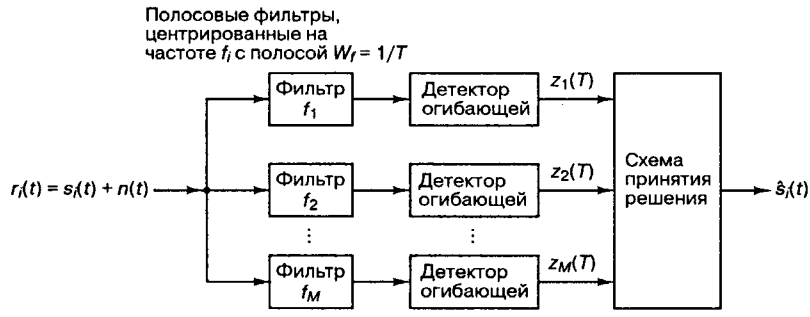


Рис. 4.19. Некогерентное обнаружение сигналов FSK с использованием детекторов огибающей

Детектор огибающей, изображенный на блочной диаграмме рис. 4.19, кажется проще квадратурного приемника, показанного на рис. 4.18, но не стоит забывать, что использование (аналоговых) фильтров обычно приводит к большей массе и стоимости детекторов огибающей по сравнению с квадратурным приемником. Поскольку квадратурные приемники могут реализовываться цифровым образом, с появлением больших интегральных схем их использование в качестве некогерентных детекторов стало предпочтительнее. Детектор, показанный на рис. 4.19, может реализовываться цифровым образом, использование аналоговых фильтров заменяется выполнением дискрет-

ного преобразования Фурье. Подобная структура обычно сложнее цифровой реализации квадратурного приемника.

4.5.4. Расстояние между тонами для некогерентной ортогональной передачи сигналов FSK

Частотная манипуляция (frequency shift keying — FSK) обычно реализуется как ортогональная передача сигналов, хотя ортогональными являются не все схемы FSK. Что мы подразумеваем под ортогональностью, когда речь идет о тонах сигнального множества? Предположим, что мы используем два тона $f_1 = 10\,000$ Гц и $f_2 = 11\,000$ Гц. Ортогональны ли они между собой? Другими словами, удовлетворяют ли они критерию ортогональности (уравнение (3.39)) и не коррелируют ли в течение периода передачи символа T? Пока у нас недостаточно информации, чтобы ответить на этот вопрос. Вообще, тоны f_1 и f_2 являются ортогональными, если при переданном тоне f_1 дискретная огибающая на выходе принимающего фильтра, согласованного с f_2, дает нуль (т.е. отсутствуют перекрестные помехи). Подобная ортогональность между тонами сигнального множества FSK обеспечивается, если любая пара тонов множества разделена по частоте расстоянием, кратным $1/T$ Гц. (Это доказывается ниже, в примере 4.3.) Тон с частотой f_i, который включается на время передачи символа (T с) и после этого выключается (такой, как тон FSK, приведенный в выражении (4.8)), аналитически можно описать следующим образом.

$$s_i(t) = (\cos 2\pi f_i t)\ \text{rect}\ (t/T),$$

где

$$\text{rect}\ (t/T) = \begin{cases} 1 & \text{для } -T/2 \le t \le T/2 \\ 0 & \text{для } |t| > T/2 \end{cases}$$

Из табл. А.1 находим Фурье-образ $s_i(t)$.

$$\mathfrak{F}\{s_i(t)\} = T\ \text{sinc}\ (f - f_i)T$$

Здесь функция sinc определена выражением (1.39). Спектры подобных соседствующих тонов — тона 1 с частотой f_1 и тона 2 с частотой f_2 — показаны на рис. 4.20.

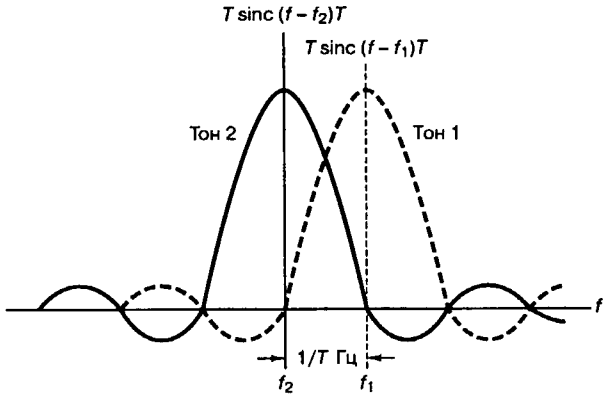


Рис. 4.20. Минимальное расстояние между тонами для ортогональной передачи сигналов FSK с некогерентным обнаружением

4.5.4.1. Минимальное расстояние между тонами и ширина полосы

Для того чтобы некогерентно обнаруживаемый тон давал максимальный сигнал на выходе "своего" фильтра и нулевой сигнал — на выходе любого соседнего фильтра (схема на рис. 4.19), максимум спектра тона 1 должен совпадать с одним из переходов через нуль спектра тона 2, а максимум спектра тона 2 должен приходиться на один из переходов через нуль спектра тона 1. Расстояние по частоте между центром спектрального главного лепестка и первым переходом через нуль является *минимальным необходимым расстоянием между тонами*. При некогерентном обнаружении это соответствует минимальному расстоянию между тонами, которое, как показано на рис. 4.20, равно $1/T$ Гц. Несмотря на то что использование схемы FSK подразумевает передачу в течение каждого интервала передачи символа всего одного однополосного тона, когда мы говорим о ширине полосы сигнала, подразумеваем спектр, достаточный для всех тонов M-арного множества. Следовательно, для модуляции FSK требования к полосе связаны со спектральным расстоянием между тонами. Можно считать, что с каждым из группы соседствующих тонов связан спектр, простирающийся в обе стороны от максимального значения на величину, равную половине расстояния между тонами. Следовательно, для бинарной модуляции FSK, изображенной на рис. 4.20, ширина полосы передачи равна спектру, находящемуся между тонами, плюс области слева и справа, ширина которых равна половине расстояния между тонами. Общий спектр, таким образом, равен удвоенному расстоянию между тонами. Экстраполируя этот результат на M-арный случай, получаем, что ширина полосы сигнала в ортогональной модуляции MFSK с некогерентным обнаружением равна M/T.

До сих пор мы рассматривали только некогерентное обнаружение сигналов в ортогональной модуляции FSK. Будет ли отличаться критерий минимального расстояния между тонами (и, как следствие, ширина полосы) при когерентном обнаружении? Разумеется, да. Как будет показано ниже, в примере 4.3, при использовании когерентного обнаружения минимальное расстояние между тонами снижается до $1/2T$.

4.5.4.2. Дуальные соотношения

Инженерную концепцию дуальности можно определить следующим образом. Два процесса (функции, элемента или системы) *дуальны* друг другу, если описывающие их математические соотношения идентичны, с точностью до фигурирующих в них переменных (например, время и частота). Рассмотрим передачу сигналов FSK, где, как показано на рис. 4.20, модулированные сигналы имеют вид функций sinc (fT). Данная длительность тона определяет минимальное расстояние по частоте между тонами, необходимое для получения ортогональности. Это соотношение в частотной области имеет дуальное ему во временной области — передачу импульсов (рис. 3.16, *б*), где прямоугольным участкам полосы соответствуют импульсы вида sinc (t/T). Данная ширина полосы определяет минимальное расстояние (на временной оси) между импульсами, необходимое для получения нулевой межсимвольной интерференции.

Пример 4.3. Минимальное расстояние между тонами для ортогональной FSK

Рассмотрим два сигнала $\cos(2\pi f_1 t + \phi)$ и $\cos(2\pi f_2 t)$, используемые для некогерентной передачи сигналов FSK, где $f_1 > f_2$. Скорость передачи символов равна $1/T$ символов/с, где T — длительность символа, а ϕ — произвольный постоянный угол между 0 и 2π.

а) Докажите, что минимальное расстояние между тонами для ортогональной передачи сигналов FSK с *некогерентным обнаружением* равно $1/T$.

б) Чему равно минимальное расстояние между тонами для ортогональной передачи сигналов FSK с *когерентным обнаружением*?

Решение

а) Чтобы два сигнала были ортогональными, они должны удовлетворять условию ортогональности, которое дается выражением (3.69).

$$\int_0^T \cos(2\pi f_1 t + \phi)\cos 2\pi f_2 t\, dt = 0 \tag{4.45}$$

Используя основные тригонометрические соотношения, приведенные в формулах (Г.6) и (Г.1)–(Г.3), можно переписать выражение (4.45) в виде

$$\cos\phi \int_0^T \cos 2\pi f_1 t \cos 2\pi f_2 t\, dt - \sin\phi \int_0^T \sin 2\pi f_1 t \cos 2\pi f_2 t\, dt = 0, \tag{4.46}$$

так что

$$\begin{aligned}\cos\phi \int_0^T [\cos 2\pi(f_1 + f_2)t + \cos 2\pi(f_1 - f_2)t]\,dt - \\ - \sin\phi \int_0^T [\sin 2\pi(f_1 + f_2)t + \sin 2\pi(f_1 - f_2 t)]dt = 0\end{aligned} \tag{4.47}$$

что дает

$$\begin{aligned}\cos\phi \left[\frac{\sin 2\pi(f_1 + f_2)t}{2\pi(f_1 + f_2)} + \frac{\sin 2\pi(f_1 - f_2)t}{2\pi(f_1 - f_2)}\right]_0^T + \\ + \sin\phi \left[\frac{\cos 2\pi(f_1 + f_2)t}{2\pi(f_1 + f_2)} + \frac{\cos 2\pi(f_1 - f_2)t}{2\pi(f_1 - f_2)}\right]_0^T = 0\end{aligned} \tag{4.48}$$

или

$$\begin{aligned}\cos\phi \left[\frac{\sin 2\pi(f_1+f_2)T}{2\pi(f_1+f_2)} + \frac{\sin 2\pi(f_1-f_2)T}{2\pi(f_1-f_2)}\right] + \\ + \sin\phi \left[\frac{\cos 2\pi(f_1+f_2)T - 1}{2\pi(f_1+f_2)} + \frac{\cos 2\pi(f_1-f_2)T - 1}{2\pi(f_1-f_2)}\right] = 0.\end{aligned} \tag{4.49}$$

Далее можно предположить, что $f_1 + f_2 >> 1$; это позволяет записать следующее.

$$\frac{\sin 2\pi(f_1 + f_2)T}{2\pi(f_1 + f_2)} \approx \frac{\cos 2\pi(f_1 + f_2)T}{2\pi(f_1 + f_2)} \approx 0 \tag{4.50}$$

Затем, объединяя выражения (4.49) и (4.50), можем записать следующее.

$$\cos\phi\ \sin 2\pi(f_1 - f_2) + \sin\phi\,[\cos 2\pi(f_1 - f_2)T - 1] \approx 0 \tag{4.51}$$

Отметим, что при произвольной фазе ϕ выражение (4.51) всегда справедливо, только если $\sin 2\pi(f_1 - f_2)T = 0$ и при этом $\cos 2\pi(f_1 - f_2)T = 1$.
Поскольку

$$\sin x = 0 \qquad \text{при } x = n\pi$$

и

$$\sin x = 1 \qquad \text{при } x = 2k\pi,$$

где n и k — целые, условия $\sin x = 0$ и $\cos x = 1$ удовлетворяются одновременно при $n = 2k$. Следовательно, из формулы (4.51) для произвольного ϕ можем записать следующее.

$$2\pi(f_1 - f_2)T = 2k\pi$$

или

$$f_1 - f_2 = k/T \tag{4.52}$$

Минимальное расстояние между тонами для ортогональной передачи сигналов FSK с *некогерентным обнаружением* получаем при $k = 1$, при этом

$$f_1 - f_2 = 1/T \tag{4.53}$$

Напомним вопрос, сформулированный выше. Имея два тона f_1 = 10 000 Гц и f_2 = 11 000 Гц, мы спрашивали, являются ли они ортогональными? Теперь у нас достаточно информации для ответа на поставленный вопрос. Ответ связан со скоростью передачи сигналов FSK. Если манипуляция сигналами (переключение сигналов) происходит со скоростью 1 000 символов/с и используется некогерентное обнаружение, то сигналы ортогональны. Если манипуляция происходит быстрее, скажем со скоростью 10 000 символов/с, сигналы не ортогональны.

б) При некогерентном обнаружении, рассмотренном в п. а, расстояние между тонами, превращающее сигналы в ортогональные, было найдено посредством выполнения уравнения (4.45) для любой произвольной фазы. В случае когерентного обнаружения расстояние между тонами находится, если положить $\phi = 0$. Причина в том, что мы знаем фазу принятого сигнала (ее дает контур ФАПЧ). Этот принятый сигнал будет коррелировать с каждым опорным сигналом, причем в качестве фазы опорного сигнала используется фаза принятого сигнала. Уравнение (4.51) можно теперь переписать с учетом $\phi = 0$.

$$\sin 2\pi(f_1 - f_2)T = 0 \tag{4.54}$$

или

$$f_1 - f_2 = n/2T \tag{4.55}$$

Минимальное расстояние между тонами для ортогональной передачи сигналов FSK с *когерентным обнаружением* получаем при $k = 1$, при этом

$$f_1 - f_2 = 1/2T \tag{4.56}$$

Следовательно, при одинаковых скоростях передачи символов когерентное обнаружение требует меньшей ширины полосы, чем некогерентное, обеспечивая при этом ортогональную передачу сигналов. Можно сказать, что передача сигналов FSK с *когерентным обнаружением* более *эффективно использует полосу.* (Вопрос эффективности использования полосы подробно рассмотрен в главе 9.)

При когерентном обнаружении тоны расположены более плотно, чем при некогерентном, поскольку, если расположить два периодических сигнала так, чтобы их начальные фазы совпадали, ортогональность будет получена автоматически в силу симметрии (четности и нечетности) соответствующих сигналов в течение одного периода передачи символа. Это является отличием от способа получения ортогональности в п. а, где мы не уделяли внимания фазе. В случае когерентного обнаружения регулировка фазы в разрядах коррелятора означает, что мы можем расположить тоны ближе (по частоте) друг к другу, при этом по-прежнему поддерживая ортогональность в наборе тонов FSK. Вы можете доказать это самостоятельно, изобразив две синусоиды (или косинусоиды, или последовательности прямоугольных импульсов). Начальная фаза всех сигналов должна быть одинаковой (удобнее всего взять ее равной 0 радиан). Используя миллиметровку,

выберите удобную временную шкалу для представления одного периода передачи символа T. Изобразите тон с периодом T, а затем изобразите другой тон, имеющий такую же начальную фазу, как и предыдущий, и период $2/3T$. Выполните численное суммирование произведений тонов (смещенных относительно друг друга на $1/2T$) и докажите, что они действительно являются ортогональными.

4.6. Комплексная огибающая

Описание реальных модуляторов и демодуляторов облегчается при использовании комплексной формы записи, введенной в разделе 4.2.1. Любой реальный полосовой сигнал $s(t)$ можно представить в комплексной форме как

$$s(t) = \mathrm{Re}\left\{g(t)e^{i\omega_0 t}\right\}, \tag{4.57}$$

где $g(t)$ — *комплексная огибающая* (complex envelop), которую можно записать следующим образом.

$$g(t) = x(t) + iy(t) = |g(t)|e^{i\theta(t)} = R(t)e^{i\theta(t)} \tag{4.58}$$

Амплитуда комплексной огибающей выражается как

$$R(t) = |g(t)| = \sqrt{x^2(t) + y^2(t)}\,, \tag{4.59}$$

а фаза определяется следующим образом.

$$\theta(t) = \mathrm{arctg}\frac{y(t)}{x(t)} \tag{4.60}$$

В формуле (4.57) $g(t)$ можно называть полосовым сообщением или данными в комплексной форме, а $e^{i\omega_0 t}$ — несущей в комплексной форме. Произведение этих двух величин представляет операцию модулирования, а $s(t)$, действительная часть произведения, — это переданный сигнал. Следовательно, используя формулы (4.4), (4.57) и (4.58), $s(t)$ можно выразить следующим образом.

$$\begin{aligned} s(t) &= \mathrm{Re}\{[x(t) + iy(t)][\cos\omega_0 t + i\sin\omega_0 t]\} = \\ &= x(t)\cos\omega_0 t - y(t)\sin\omega_0 t \end{aligned} \tag{4.61}$$

Отметим, что модулирование сигналов, выраженное в общей форме $(a + ib)$, умноженное на $(c + id)$, дает сигнал с переменой знака (в квадратурном члене несущей волны) вида $ac - bd$.

4.6.1. Квадратурная реализация модулятора

Рассмотрим узкополосный сигнал $g(t)$, который представлен последовательностью идеальных импульсов $x(t)$ и $y(t)$, передаваемых в дискретные моменты времени $k = 1, 2, \ldots$. Таким образом, $g(t)$, $x(t)$ и $y(t)$ в уравнении (4.58) можно записывать как g_k, x_k и y_k. Пусть значения амплитуд импульсов равны $x_k = y_k = 0{,}707A$. При этом комплексную огибающую можно выразить в дискретной форме следующим образом.

$$g_k = x_k + iy_k = 0{,}707A + i0{,}707A \tag{4.62}$$

Из комплексной алгебры знаем, что $i = \sqrt{-1}$, но с практической точки зрения i можно рассматривать как "метку", напоминающую, что мы не можем использовать обычное сложение при группировке членов в формуле (4.62). Далее мы будем рассматривать синфазную и квадратурную модуляции, x_k и y_k, как упорядоченную пару. Модулятор, реализованный по квадратурному принципу, показан на рис. 4.21, где можно видеть, что импульс x_k умножается на $\cos \omega_0 t$ (синфазный компонент несущей), а импульс y_k — на $\sin \omega_0 t$ (квадратурный компонент несущей). Процесс модулирования можно кратко описать как умножение комплексной огибающей на $e^{i\omega_0 t}$ с последующей передачей действительной части произведения. Итак, записываем следующее.

$$\begin{aligned} s(t) &= \mathrm{Re}\{g_k e^{i\omega_0 t}\} = \\ &= \mathrm{Re}\{(x_k + iy_k)(\cos \omega_0 t + i \sin \omega_0 t)\} = \\ &= x_k \cos \omega_0 t - y_k \sin \omega_0 t = \\ &= 0{,}707A \cos \omega_0 t - 0{,}707A \sin \omega_0 t = \\ &= A \cos\left(\omega_0 t + \frac{\pi}{4}\right) \end{aligned} \tag{4.63}$$

Снова напомним, что квадратурный член несущей включает перемену знака в процессе модуляции. Если в качестве опорного сигнала использовать $0{,}707A \cos \omega_0 t$, то при передаче сигнала $s(t)$ (уравнение (4.63)) происходит сдвиг опоры на $\pi/4$. Если же в качестве опорного сигнала применить $-0{,}707A \sin \omega_0 t$, то переданный сигнал $s(t)$ в уравнении (4.63) приводит к запаздыванию опоры на $\pi/4$. Графическая иллюстрация сказанного приведена на рис. 4.22

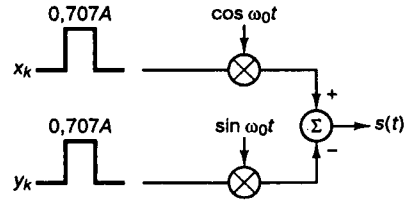


Рис. 4.21. Модулятор, работающий по квадратурному принципу

4.6.2. Пример модулятора D8PSK

На рис. 4.23 изображена квадратурная реализация модулятора дифференциальной восьмифазной манипуляции (differential 8-PSK — D8PSK). Поскольку модуляция является 8-ричной, каждому информационному вектору ϕ_k, который можно записать как

$$\phi_k = \Delta\phi_k + \phi_{k-1}, \tag{4.64}$$

присваивается 3-битовое сообщение (x_k, y_k, z_k).

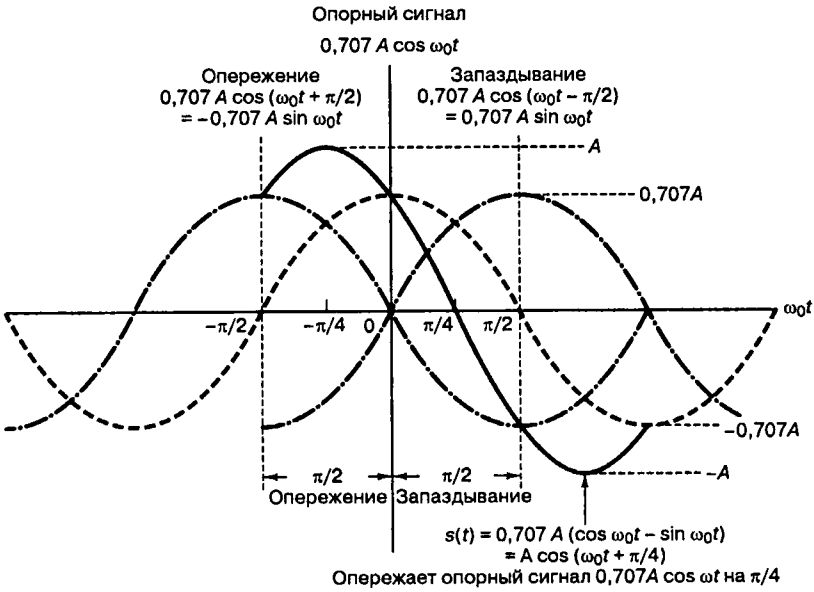
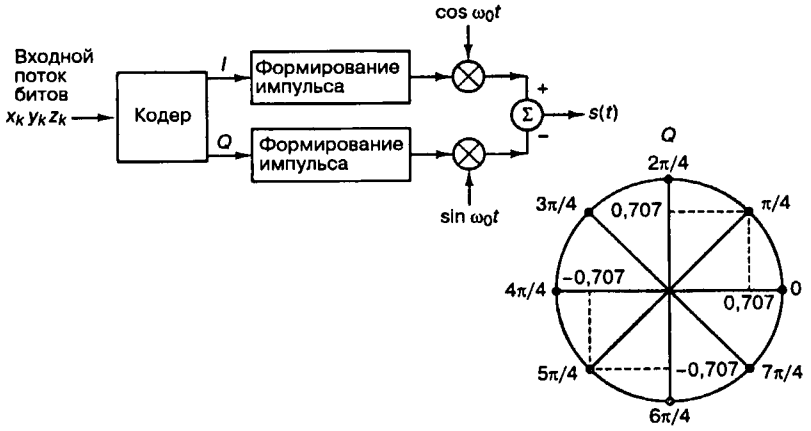


Рис. 4.22. Опережение/запаздывание синусоид



Кодирование данных

x_k	y_k	z_k	Δf_k
0	0	0	0
0	0	1	π/4
0	1	1	2π/4
0	1	0	3π/4
1	1	0	4π/4
1	1	1	5π/4
1	0	1	6π/4
1	0	0	7π/4

Дифференциальный информационный вектор

$\phi_k = \phi_{k-1} + \Delta\phi_k$

Положим $\phi_0 = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$x_k y_k z_k$	110	001	110	010
$\Delta\phi_k$:	π	π/4	π	3π/4
ϕ_k:	π	5π/4	π/4	π
I:	−1	−0,707	0,707	−1
Q:	0	−0,707	0,707	0

Рис. 4.23. Квадратурная реализация модулятора D8PSK

Сложение текущего кодируемого сообщения, выраженного разностью фаз $\Delta\phi_k$, с предыдущей фазой ϕ_{k-1} обеспечивает дифференциальное кодирование сообщений. Последовательность векторов, созданная с использованием уравнения (4.64), подобна результатам дифференциального кодирования, полученного с помощью процедуры, описанной в разделе 4.5.2. Можно заметить (рис. 4.23), что в результате кодирования 3-битовых последовательностей сообщений разностями фаз $\Delta\phi_k$ получаем не двоичную последовательность от 000 до 111, а специальный код, называемый *кодом Грея* (Gray code). (Преимущества использования подобного кода приведены в разделе 4.9.4.)

Пусть на вход модулятора, изображенного на рис. 4.23, в моменты времени $k = 1$, 2, 3, 4 поступают информационные последовательности 110, 001, 110, 010. Далее используем таблицу кодирования данных, приведенную на рис. 4.23, формулу (4.64) и, кроме того, положим начальную фазу (момент времени $k = 0$) равной нулю: $\phi_0 = 0$. В момент времени $k = 1$ дифференциальная информационная фаза, соответствующая набору $x_1y_1z_1 = 110$, равна $\phi_1 = 4\pi/4 = \pi$. Считая амплитуду вращающегося вектора единичной, синфазный (I) и квадратурный (Q) узкополосные импульсы равны –1 и 0. Как показано на рис. 4.23, форму этих импульсов обычно задает фильтр (такой, как фильтр с характеристикой типа приподнятого косинуса).

Для момента $k = 2$ таблица на рис. 4.23 показывает, что сообщение 001 кодируется сдвигом фаз $\Delta\phi_2 = \pi/4$. Следовательно, согласно формуле (4.64), вторая дифференциальная информационная фаза равна $\phi_2 = \pi + \pi/4 = 5\pi/4$, и в момент $k = 2$ синфазный и квадратурный узкополосные импульсы равны, соответственно, $x_k = -0{,}707$ и $y_k = -0{,}707$. Переданный сигнал имеет вид, приведенный в формуле (4.61).

$$s(t) = \mathrm{Re}\{(x_k + iy_k)(\cos\omega_0 t + i\sin\omega_0 t)\} = \tag{4.65}$$
$$= x_k\cos\omega_0 t - y_k\sin\omega_0 t$$

Для сигнального множества, которое может представляться в координатах "фаза-амплитуда", такого как MPSK или MQAM, уравнение (4.65) позволяет сделать интересное наблюдение. Из него видно, что квадратурная реализация передатчика сводит все типы передачи сигналов к единственной амплитудной модуляции. Каждый вектор на плоскости передается посредством амплитудной модуляции его синфазной и квадратурной проекций на синусоидный и косинусоидный компоненты его несущей. В каждом случае процесс формирования импульса считается идеальным, т.е. предполагается, что информационные импульсы имеют идеальные прямоугольные формы. Таким образом, используя уравнение (4.65) для момента $k = 2$, при $x_k = -0{,}707$ и $y_k = -0{,}707$, можно записать переданный сигнал $s(t)$ следующим образом.

$$s(t) = -0{,}707\cos\omega_0 t - 0{,}707\sin\omega_0 t \tag{4.66}$$
$$= \sin\left(\omega_0 t - \frac{\pi}{4}\right)$$

4.6.3. Пример демодулятора D8PSK

В предыдущем разделе описание квадратурной реализации модулятора начиналось с умножения комплексной огибающей (узкополосного сообщения) на $e^{i\omega_0 t}$ с последующей передачей действительной части произведения $s(t)$, описанного в форму-

ле (4.63). Демодулятор подобной схемы включает обратный процесс, т.е. умножение принятого полосового сигнала на $e^{-i\omega_0 t}$ с целью восстановления узкополосного сигнала. В левой части рис. 4.24 в упрощенном виде показан модулятор, изображенный на рис. 4.23, и сигнал $s(t) = \sin(\omega_0 t - \pi/4)$, переданный в момент времени $k = 2$ (продолжаем использовать пример, описанный в предыдущем разделе). В правой части рис. 4.24 показана квадратурная реализация демодулятора.

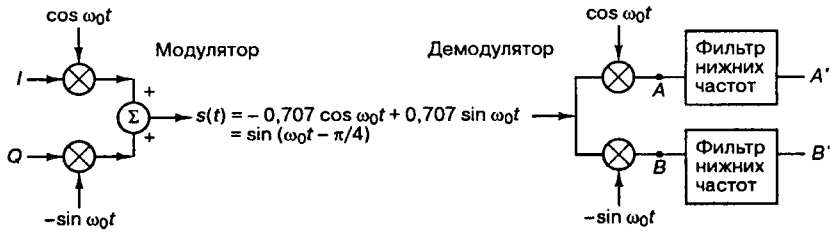


Рис. 4.24. Пример модулятора/демодулятора

Отметим тонкое отличие между членом $-\sin\omega_0 t$ в модуляторе и демодуляторе. В модуляторе знак "минус" появляется при определении действительной части комплексного сигнала (произведения комплексной огибающей и комплексной несущей). В демодуляторе член $-\sin\omega_0 t$ появляется при умножении полосового сигнала на сопряженное $e^{-i\omega_0 t}$ несущей модулятора. Демодуляция является когерентной, если фаза восстанавливается. Для упрощения записи основных соотношений процесса мы пренебрегаем шумом. Итак, после синфазного умножения в демодуляторе на $\cos\omega_0 t$ в точке A получаем следующий сигнал.

$$A = (-0{,}707\cos\omega_0 t + 0{,}707\sin\omega_0 t)\cos\omega_0 t = \tag{4.67}$$

$$= -0{,}707\cos^2\omega_0 t + 0{,}707\sin\omega_0 t\cos\omega_0 t$$

Используя тригонометрические соотношения, приведенные в формулах (Г.7) и (Г.9), получаем следующее.

$$A = \frac{-0{,}707}{2}(1+\cos 2\omega_0 t) + \frac{0{,}707}{2}\sin 2\omega_0 t \tag{4.68}$$

После фильтрации с использованием фильтра нижних частот (low-pass filter — LPF) в точке A' восстанавливается идеальный отрицательный импульс.

$$A' = -0{,}707 \text{ (с точностью до масштабного коэффициента)} \tag{4.69}$$

Подобным образом после квадратурного умножения в демодуляторе на $-\sin\omega_0 t$ в точке B получаем сигнал.

$$B = (-0{,}707\cos\omega_0 t + 0{,}707\sin\omega_0 t)(-\sin\omega_0 t) = \tag{4.70}$$

$$= \frac{0{,}707}{2}\sin 2\omega_0 t - \frac{0{,}707}{2}(1-\cos 2\omega_0 t)$$

После прохождения сигналом фильтра нижних частот в точке B' восстанавливается идеальный отрицательный импульс.

$$B' = -0{,}707 \text{ (с точностью до масштабного коэффициента)} \tag{4.71}$$

Таким образом, видим, что в точках A' и B' (идеальные) дифференциальные информационные импульсы для синфазного и квадратурного каналов равны –0,707. Поскольку модулятор/демодулятор является дифференциальным, для нашего примера $k = 2$ получаем следующее.

$$\Delta\phi_{k=2} = \phi_{k=2} - \phi_{k=1} \tag{4.72}$$

Будем считать, что в предыдущий момент времени $k = 1$ демодулятор правильно определил, что фаза сигнала равна π. Тогда из формулы (4.72) можем получить следующее.

$$\Delta\phi_{k=2} = 5\pi/4 - \pi = \pi/4 \tag{4.73}$$

Вернувшись к таблице модуляции на рис. 4.23, видим, что данной фазе соответствует информационная последовательность $x_2y_2z_2 = 001$, что совпадает с данными, посланными в момент времени $k = 2$.

4.7. Вероятность ошибки в бинарных системах

4.7.1. Вероятность появления ошибочного бита при когерентном обнаружении сигнала BPSK

Важной мерой производительности, используемой для сравнения цифровых схем модуляции, является вероятность ошибки, P_E. Для коррелятора или согласованного фильтра вычисление P_E можно представить геометрически (см. рис. 4.6). Расчет P_E включает нахождение вероятности того, что при данном векторе переданного сигнала, скажем $\mathbf{s}_1$, вектор шума $\mathbf{n}$ выведет сигнал из области 1. Вероятность принятия детектором неверного решения называется *вероятностью символьной ошибки*, P_E. Несмотря на то что решения принимаются на символьном уровне, производительность системы часто удобнее задавать через вероятность битовой ошибки (P_B). Связь P_B и P_E рассмотрена в разделе 4.9.3 для ортогональной передачи сигналов и в разделе 4.9.4 для многофазной передачи сигналов.

Для удобства изложения в данном разделе мы ограничимся когерентным обнаружением сигналов BPSK. В этом случае вероятность символьной ошибки — это то же самое, что и вероятность битовой ошибки. Предположим, что сигналы равновероятны. Допустим также, что при передаче сигнала $s_i(t)$ ($i = 1, 2$) принятый сигнал $r(t)$ равен $s_i(t) + n(t)$, где $n(t)$ — процесс AWGN; кроме того, мы пренебрегаем ухудшением качества вследствие введенной каналом или схемой межсимвольной интерференции. Как показывалось в разделе 4.4.1, антиподные сигналы $s_1(t)$ и $s_2(t)$ можно описать в одномерном сигнальном пространстве, где

и

$$\left.\begin{aligned} s_1(t) &= \sqrt{E}\psi_1(t) \\ s_2(t) &= -\sqrt{E}\psi_1(t) \end{aligned}\right\} 0 \le t \le T. \tag{4.74}$$

Детектор выбирает $s_i(t)$ с наибольшим выходом коррелятора $z_i(T)$; или, в нашем случае антиподных сигналов с равными энергиями, детектор, используя формулу (4.20), принимает решение следующего вида.

и

$$\begin{aligned} &s_1(t), \quad \text{если } z(T) > \gamma_0 = 0 \\ &s_2(t) \quad \text{при других } z(T) \end{aligned} \tag{4.75}$$

Как видно из рис. 4.9, возможны ошибки двух типов: шум так искажает переданный сигнал $s_1(t)$, что измерения в детекторе дают отрицательную величину $z(T)$, и детектор выбирает гипотезу H_2, что был послан сигнал $s_2(t)$. Возможна также обратная ситуация: шум искажает переданный сигнал $s_2(t)$, измерения в детекторе дают положительную величину $z(T)$, и детектор выбирает гипотезу H_1, соответствующую предположению о передаче сигнала s_1.

В разделе 3.2.1.1 была выведена формула (3.42), описывающая вероятность битовой ошибки P_B для детектора, работающего по принципу *минимальной вероятности ошибки*.

$$P_B = \int_{(a_1 - a_2)/2\sigma_0}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du = Q\left(\frac{a_1 - a_2}{2\sigma_0}\right) \tag{4.76}$$

Здесь σ_0 — среднеквадратическое отклонение шума вне коррелятора. Функция $Q(x)$, называемая *гауссовым интегралом ошибок*, определяется следующим образом.

$$Q(X) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} \exp\left(-\frac{u^2}{2}\right) du \tag{4.77}$$

Эта функция подробно описывается в разделах 3.2 и Б.3.2.

Для передачи антиподных сигналов с равными энергиями, таких как сигналы в формате BPSK, приведенные в выражении (4.74), на выход приемника поступают следующие компоненты: $a_1 = \sqrt{E_b}$, при переданном сигнале $s_1(t)$, и $a_2 = -\sqrt{E_b}$, при переданном сигнале $s_2(t)$, где E_b — энергия сигнала, приходящаяся на двоичный символ. Для процесса AWGN дисперсию шума σ_0^2 вне коррелятора можно заменить $N_0/2$ (см. приложение В), так что формулу (4.76) можно переписать следующим образом.

$$P_B = \int_{\sqrt{2E_b/N_0}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du = \tag{4.78}$$

$$= Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \tag{4.79}$$

Данный результат для полосовой передачи антиподных сигналов BPSK совпадает с полученными ранее формулами для обнаружения антиподных сигналов с использованием согласованного фильтра (формула (3.70)) и обнаружения узкополосных антиподных сигналов с применением согласованного фильтра (формула (3.76)). Это является примером описанной ранее *теоремы эквивалентности*. Для линейных систем теорема эквивалентности утверждает, что на математическое описание процесса обнаружения не влияет сдвиг частоты. Как следствие, использование согласованных фильтров или корреляторов для обнаружения полосовых сигналов (рассмотренное в данной главе) дает те же соотношения, что были выведены ранее для сопоставимых узкополосных сигналов.

Пример 4.4. Вероятность битовой ошибки при передаче сигналов BPSK

Найдите вероятность появления ошибочного бита в системе, использующей схему BPSK и скорость 1 Мбит/с. Принятые сигналы $s_1(t) = A\cos\omega_0 t$ и $s_2(t) = -A\cos\omega_0 t$ обнаруживаются когерентно с использованием согласованного фильтра. Величина A равна 10 мВ. Однополосную спектральную плотность шума считать равной $N_0 = 10^{-11}$ Вт/Гц, а мощность сигнала и энергию на бит — нормированными на 1 Ом.

Решение

$$A = \sqrt{\frac{2E_b}{T}} = 10^{-2}\,\text{В} \qquad T = \frac{1}{R} = 10^{-6}\,\text{с}$$

Следовательно,

$$E_b = \frac{A^2}{2}T = 5\times 10^{-11}\,\text{Дж} \quad \text{и} \quad \sqrt{\frac{2E_b}{N_0}} = 3{,}16$$

$$P_B = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) = Q(3{,}16)$$

Используя табл. Б.1 или формулу (3.44), получаем следующее.

$$P_B = 8\times 10^{-4}$$

4.7.2. Вероятность появления ошибочного бита при когерентном обнаружении сигнала в дифференциальной модуляции BPSK

Сигналы в канале иногда инвертируются; например, при использовании когерентного опорного сигнала, генерируемого контуром ФАПЧ, фаза может быть неоднозначной. Если фаза несущей была инвертирована при использовании схемы DPSK, как это скажется на сообщении? Поскольку информация сообщения кодируется подобием или отличием соседних символов, единственным следствием может быть ошибка в бите, который инвертируется, или в бите, непосредственно следующим за инвертированным. Точность определения подобия или отличия символов не меняется при инвертировании несущей. Иногда сообщения (и кодирующие их сигналы) *дифференциально кодируются* и *когерентно обнаруживаются*, чтобы просто избежать неопределенности в определении фазы.

Вероятность появления ошибочного бита при когерентном обнаружении сигналов в дифференциальной модуляции PSK (DPSK) дается выражением [5].

$$P_B = 2Q\left(\sqrt{\frac{2E_b}{N_0}}\right)\left[1 - Q\left(\sqrt{\frac{2E_b}{N_0}}\right)\right] \tag{4.80}$$

Это соотношение изображено на рис. 4.25. Отметим, что существует незначительное ухудшение достоверности обнаружения по сравнению с когерентным обнаружением сигналов в модуляции PSK. Это вызвано дифференциальным кодированием, поскольку любая отдельная ошибка обнаружения обычно приводит к принятию двух ошибочных решений. Подробно вероятность ошибки при использовании наиболее популярной схемы — когерентного обнаружения сигналов в модуляции DPSK — рассмотрена в разделе 4.7.5.

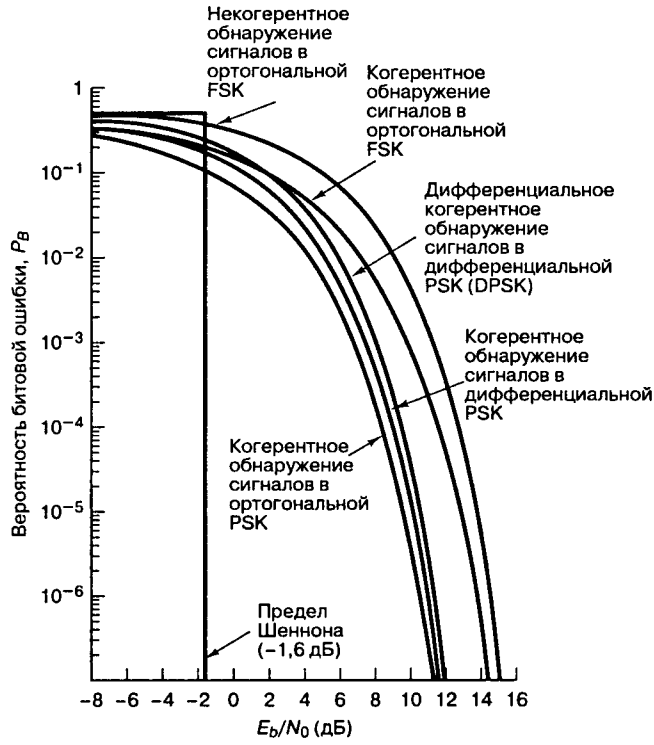


Рис. 4.25. Вероятность появления ошибочного бита для бинарных систем нескольких типов

4.7.3. Вероятность появления ошибочного бита при когерентном обнаружении сигнала в бинарной ортогональной модуляции FSK

Формулы (4.78) и (4.79) описывают вероятность появления ошибочного бита для когерентного обнаружения антиподных сигналов. Более общую трактовку для когерентного обнаружения бинарных сигналов (не ограничивающихся антиподными сигналами) дает следующее выражение для P_B [6].

$$P_B = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{(1-\rho)E_b/N_0}}^{\infty} \exp\left(-\frac{u^2}{2}\right) du \tag{4.81}$$

Из формулы (3.64,б) $\rho = \cos\theta$ — временной коэффициент взаимной корреляции между $s_1(t)$ и $s_2(t)$, где θ — угол между векторами сигналов $\mathbf{s}_1$ и $\mathbf{s}_2$ (см. рис. 4.6). Для антиподных сигналов, таких как сигналы BPSK, $\theta = \pi$, поэтому $\rho = -1$.

Для ортогональных сигналов, таких как сигналы бинарной FSK (BFSK), $\theta = \pi/2$, поскольку векторы $\mathbf{s}_1$ и $\mathbf{s}_2$ перпендикулярны; следовательно, $\rho = 0$, что можно доказать с помощью формулы (3.64,а), поэтому выражение (4.81) можно переписать следующим образом.

$$P_B = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{E_b/N_0}}^{\infty} \exp\left(-\frac{u^2}{2}\right) du = Q\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{4.82}$$

Здесь $Q(x)$ — дополнительная функция ошибок, подробно описанная в разделах 3.2 и Б.3.2. Зависимость (4.82) для когерентного обнаружения ортогональных сигналов BFSK, показанная на рис. 4.25, аналогична зависимости, полученной для обнаружения ортогональных сигналов с помощью согласованного фильтра (формула (3.71)) и узкополосных ортогональных сигналов (униполярных импульсов) с использованием согласованного фильтра (формула (3.73)). В данной книге мы не рассматриваем *амплитудную манипуляцию OOK* (on-off keying), но соотношение (4.82) применимо к обнаружению с помощью согласованного фильтра сигналов OOK, так же как и к когерентному обнаружению любых ортогональных сигналов.

Справедливость соотношения (4.82) подтверждает и то, что разность энергий между ортогональными векторами сигналов $\mathbf{s}_1$ и $\mathbf{s}_2$ с амплитудой $\sqrt{E}$, как показано на рис. 3.10, *б*, равна квадрату расстояния между концами ортогональных векторов $E_d = 2E_b$. Подстановка этого результата в формулу (3.63) также дает формулу (4.82). Сравнивая формулы (4.82) и (4.79), видим, что, по сравнению со схемой BPSK, схема BFSK требует на 3 дБ большего отношения E_b/N_0 для обеспечения аналогичной достоверности передачи. Этот результат не должен быть неожиданным, поскольку при данной мощности сигнала квадрат расстояния между ортогональными векторами вдвое (на 3 дБ) больше квадрата расстояния между антиподными векторами.

4.7.4. Вероятность появления ошибочного бита при некогерентном обнаружении сигнала в бинарной ортогональной модуляции FSK

Рассмотрим бинарное ортогональное множество равновероятных сигналов FSK $\{s_i(t)\}$, определенное формулой (4.8).

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos(\omega_i t + \phi) \qquad 0 \le t \le T, \quad i = 1, 2$$

Фаза ϕ неизвестна и предполагается постоянной. Детектор описывается $M = 2$ каналами, состоящими, как показано на рис. 4.19, из полосовых фильтров и детекторов огибающей. На вход детектора поступает принятый сигнал $r(t) = s_i(t) + n(t)$, где $n(t)$ — гауссов шум с двусторонней спектральной плотностью мощности $N_0/2$. Предположим, что $s_1(t)$ и $s_2(t)$ достаточно разнесены по частоте, чтобы их перекрытием можно было пренебречь. Вычисление вероятности появления ошибочного бита для равновероятных сигналов $s_1(t)$ и $s_2(t)$ начнем, как и в случае узкополосной передачи, с уравнения (3.38).

$$\begin{aligned} P_B &= \tfrac{1}{2} P(H_2 \mid s_1) + \tfrac{1}{2} P(H_1 \mid s_2) = \\ &= \frac{1}{2} \int_{-\infty}^{0} p(z|s_1)\, dz + \frac{1}{2} \int_{0}^{\infty} p(z|s_2)\, dz \end{aligned} \tag{4.83}$$

Для бинарного случая тестовая статистика $z(T)$ определена как $z_1(T) - z_2(T)$. Предположим, что полоса фильтра W_f равна $1/T$, так что огибающая сигнала FSK (приблизительно) сохраняется на выходе фильтра. При отсутствии шума в приемнике значение $z(T)$ равно $\sqrt{2E/T}$

при передаче $s_1(t)$ и $-\sqrt{2E/T}$ — при передаче $s_2(t)$. Вследствие такой симметрии оптимальный порог $\gamma_0 = 0$. Плотность вероятности $p(z|s_1)$ подобна плотности вероятности $p(z|s_2)$.

$$p(z|s_1) = p(-z|s_2) \tag{4.84}$$

Таким образом, можем записать

$$P_B = \int_0^\infty p(z|s_2)\, dz \tag{4.85}$$

или

$$P_B = P(z_1 > z_2|s_2), \tag{4.86}$$

где z_1 и z_2 обозначают выходы $z_1(T)$ и $z_2(T)$ детекторов огибающей, показанных на рис. 4.19. При передаче тона $s_2(t) = \cos \omega_2 t$, т.е. когда $r(t) = s_2(t) + n(t)$, выход $z_1(T)$ состоит *исключительно из случайной переменной гауссового шума*; он не содержит сигнального компонента. Распределение Гаусса в *нелинейном детекторе огибающей* дает распределение Релея на выходе [6], так что

$$p(z_1|s_2) = \begin{cases} \dfrac{z_1}{\sigma_0^2} \exp\left(-\dfrac{z_1^2}{2\sigma_0^2}\right) & z_1 \geq 0 \\ 0 & z_1 < 0 \end{cases}, \tag{4.87}$$

где σ_0^2 — шум на выходе фильтра. С другой стороны, $z_2(T)$ имеет распределение Раиса, поскольку на вход нижнего детектора огибающей подается синусоида плюс шум [6]. Плотность вероятности $p(z_2|s_2)$ записывается как

$$p(z_2|s_2) = \begin{cases} \dfrac{z_2}{\sigma_0^2} \exp\left(-\dfrac{(z_2^2 + A^2)}{2\sigma_0^2}\right) I_0\left(\dfrac{z_2 A}{\sigma_0^2}\right) & z_2 \geq 0 \\ 0 & z_1 < 0 \end{cases}, \tag{4.88}$$

где $A = \sqrt{2E/T}$ и, как и ранее, σ_0^2 — шум на выходе фильтра. Функция $I_0(x)$, известная как модифицированная функция Бесселя первого рода нулевого порядка [7], определяется следующим образом.

$$I_0(x) = \frac{1}{2\pi} \int_0^{2\pi} \exp(x \cos\theta)\, d\theta \tag{4.89}$$

Ошибка при передаче $s_2(t)$ происходит, если выборка огибающей $z_1(T)$, полученная из верхнего канала (по которому проходит шум), больше выборки огибающей $z_2(T)$, полученной из нижнего канала (по которому проходит сигнал и шум). Таким образом, вероятность этой ошибки можно получить, проинтегрировав $p(z_1|s_2)$ по z_1 от z_2 до бесконечности с последующим усреднением результата по всем возможным z_2.

$$P_B = P(z_1 > z_2|s_2) = \tag{4.90}$$
$$= \int_0^\infty p(z_2 \mid s_2) \left[\int_{z_2}^\infty p(z_1 \mid s_2) dz_1 \right] dz_2 =$$

$$= \int_0^\infty \frac{z_2}{\sigma_0^2} \exp\left[-\frac{(z_2^2 + A^2)}{2\sigma_0^2}\right] I_0\left(\frac{z_2 A}{2\sigma_0^2}\right) \left[\int_{z_2}^\infty \frac{z_1}{\sigma_0^2} \exp\left(-\frac{z_1^2}{2\sigma_0^2}\right) dz_1\right] dz_2 \quad (4.91)$$

Здесь $A = \sqrt{2E/T}$, внутренний интеграл — условная вероятность ошибки при фиксированном значении z_2, если был передан сигнал $s_2(t)$, а внешний интеграл усредняет условную вероятность по всем возможным значениям z_2. Данный интеграл можно вычислить аналитически [8], и его значение равно следующему.

$$P_B = \frac{1}{2} \exp\left(-\frac{A^2}{4\sigma_0^2}\right) \quad (4.92)$$

С помощью формулы (1.19) шум на выходе фильтра можно выразить как

$$\sigma_0^2 = 2\left(\frac{N_0}{2}\right) W_f, \quad (4.93)$$

где $G_n(f) = N_0/2$, а W_f — ширина полосы фильтра. Таким образом, формула (4.92) приобретает следующий вид.

$$P_B = \frac{1}{2} \exp\left(-\frac{A^2}{4N_0 W_f}\right) \quad (4.94)$$

Выражение (4.94) показывает, что вероятность ошибки зависит от ширины полосы полосового фильтра и P_B уменьшается при снижении W_f. Результат справедлив только при пренебрежении межсимвольной интерференцией (intersymbol interference — ISI). Минимальная разрешенная W_f (т.е. не дающая межсимвольной интерференции) получается из уравнения (3.81) при коэффициенте сглаживания $r = 0$. Следовательно, $W_f = R$ бит/с $= 1/T$, и выражение (4.94) можно переписать следующим образом.

$$P_B = \frac{1}{2} \exp\left(-\frac{A^2 T}{4N_0}\right) = \quad (4.95)$$

$$= \frac{1}{2} \exp\left(-\frac{E_b}{2N_0}\right) \quad (4.96)$$

Здесь $E_b = (1/2)A^2 T$ — энергия одного бита. Если сравнить вероятность ошибки схем некогерентной и когерентной FSK (см. рис. 4.25), можно заметить, что при равных P_B некогерентная FSK требует приблизительно на 1 дБ большего отношения E_b/N_0, чем когерентная FSK (для $P_B \leq 10^{-4}$). При этом некогерентный приемник легче реализуется, поскольку не требуется генерировать когерентные опорные сигналы. По этой причине практически все приемники FSK используют некогерентное обнаружение. В следующем разделе будет показано, что при сравнении когерентной ортогональной схемы FSK с некогерентной схемой DPSK имеет место та же разница в 3 дБ, что и при сравнении когерентной ортогональной FSK и когерентной PSK.

Как указывалось ранее, в данной книге не рассматривается амплитудная манипуляция OOK (on-off keying). Все же отметим, что вероятность появления ошибочного бита P_B, выраженная в формуле (4.96), идентична P_B для некогерентного обнаружения сигналов OOK.

4.7.5. Вероятность появления ошибочного бита для бинарной модуляции DPSK

Определим набор сигналов BPSK следующим образом.

$$\begin{aligned} x_1(t) &= \sqrt{\frac{2E}{T}}\cos(\omega_0 t + \phi) \qquad 0 \le t \le T \\ x_2(t) &= \sqrt{\frac{2E}{T}}\cos(\omega_0 t + \phi \pm \pi) \qquad 0 \le t \le T \end{aligned} \tag{4.97}$$

Особенностью схемы DPSK является отсутствие в сигнальном пространстве четко определенных областей решений. В данном случае решение основывается на разности фаз между принятыми сигналами. Таким образом, при передаче сигналов DPSK каждый бит в действительности передается парой двоичных сигналов.

$$\text{и} \quad \begin{aligned} s_1(t) &= (x_1, x_1) \quad \text{или} \quad (x_2, x_2) \quad 0 \le t \le 2T \\ s_2(t) &= (x_1, x_2) \quad \text{или} \quad (x_2, x_1) \quad 0 \le t \le 2T \end{aligned} \tag{4.98}$$

Здесь (x_i, x_j) $(i, j = 1, 2)$ обозначает сигнал $x_i(t)$, за которым следует сигнал $x_j(t)$. Первые T секунд каждого сигнала — это в действительности последние T секунд предыдущего. Отметим, что оба сигнала $s_1(t)$ и $s_2(t)$ могут принимать любую из возможных форм и что $x_1(t)$ и $x_2(t)$ — это антиподные сигналы. Таким образом, корреляцию между $s_1(t)$ и $s_2(t)$ для *любой комбинации* сигналов можно записать следующим образом.

$$\begin{aligned} z(2T) &= \int_0^{2T} s_1(t)\, s_2(t)\, dt \\ &= \int_0^T [x_1(t)]^2 dt - \int_0^T [x_1(t)]^2 dt = 0 \end{aligned} \tag{4.99}$$

Следовательно, каждую пару сигналов DPSK можно представить как ортогональный сигнал длительностью $2T$ секунд. Обнаружение может соответствовать некогерентному обнаружению огибающей с помощью четырех каналов, согласованных с каждым возможным выходом огибающей, как показано на рис. 4.26. Поскольку два детектора огибающей, представляющих каждый символ, обратны друг другу, выборки их огибающих будут совпадать. Значит, мы можем реализовать детектор как один канал для $s_1(t)$, согласовывающегося с (x_1, x_1) или (x_2, x_2), и один канал для $s_2(t)$, согласовывающегося с (x_1, x_2) или (x_2, x_1), как показано на рис. 4.26. Следовательно, детектор DPSK сокращается до стандартного двухканального некогерентного детектора. В действительности фильтр может согласовываться с разностным сигналом; так что необходимым является всего один канал. На рис. 4.26 показаны фильтры, которые согласовываются с огибающими сигнала (в течение двух периодов передачи символа). Что это означает, если вспомнить, что DPSK — это схема передачи сигналов с постоянной

огибающей? Это означает, что нам требуется реализовать детектор энергии, подобный квадратурному приемнику на рис. 4.18, где каждый сигнал в течение периода ($0 \le t \le 2T$) представляется синфазным и квадратурным опорными сигналами.

синфазный опорный сигнал $s_1(t)$: $\sqrt{2/T}\cos\omega_0 t$, $\sqrt{2/T}\cos\omega_0 t$

квадратурный опорный сигнал $s_1(t)$: $\sqrt{2/T}\sin\omega_0 t$, $\sqrt{2/T}\sin\omega_0 t$

синфазный опорный сигнал $s_2(t)$: $\sqrt{2/T}\cos\omega_0 t$, $-\sqrt{2/T}\cos\omega_0 t$

квадратурный опорный сигнал $s_2(t)$: $\sqrt{2/T}\sin\omega_0 t$, $-\sqrt{2/T}\sin\omega_0 t$

Поскольку пары сигналов DPSK ортогональны, вероятность ошибки при подобном некогерентном обнаружении дается выражением (4.96). Впрочем, поскольку сигналы DPSK длятся $2T$ секунд, энергия сигналов $s_i(t)$, определенных в формуле (4.98), равна удвоенной энергии сигнала, определенного в течение одного периода передачи символа.

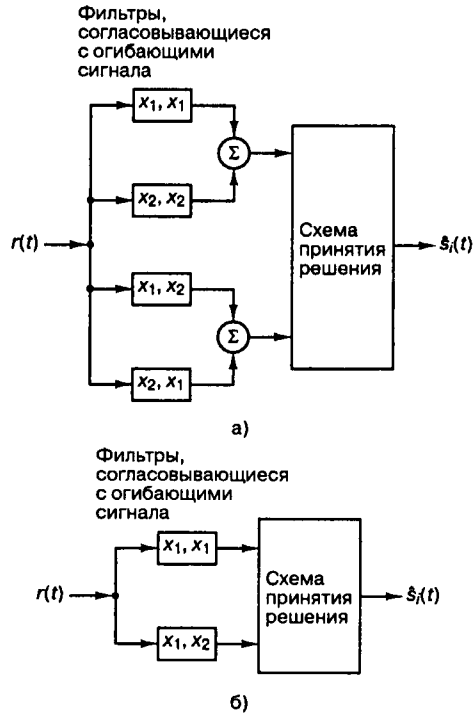


Рис. 4.26. Обнаружение в схеме DPSK: а) четырехканальное дифференциально-когерентное обнаружение сигналов в бинарной модуляции DPSK; б) эквивалентный двухканальный детектор сигналов в бинарной модуляции DPSK

Таким образом, P_B можно записать в следующем виде.

$$P_B = \frac{1}{2}\exp\left(-\frac{E_b}{N_0}\right) \tag{4.100}$$

Зависимость (4.100), изображенная на рис. 4.25, представляет собой дифференциальное когерентное обнаружение сигналов в дифференциальной модуляции PSK, или просто DPSK. Выражение справедливо для оптимального детектора DPSK (рис. 4.17, *в*). Для детектора, показанного на рис. 4.17, *б*, вероятность ошибки будет несколько выше приведенной в выражении (4.100) [3]. Если сравнить вероятность ошибки, приведенную в формуле (4.100), с вероятностью ошибки когерентной схемы PSK (см. рис. 4.25), видно, что при равных P_B схема DPSK требует приблизительно на 1 дБ большего отношения E_b/N_0, чем схема BPSK (для $P_B \leq 10^{-4}$). Систему DPSK реализовать легче, чем систему PSK, поскольку приемник DPSK не требует фазовой синхронизации. По этой причине иногда предпочтительнее использовать менее эффективную схему DPSK, чем более сложную схему PSK.

4.7.6. Вероятность ошибки для различных модуляций

В табл. 4.1 и на рис. 4.25 приведены аналитические выражения и графики P_B для наиболее распространенных схем модуляции, описанных выше. Для $P_B = 10^{-4}$ можно видеть, что разница между лучшей (когерентной PSK) и худшей (некогерентной ортогональной FSK) из рассмотренных схем равна приблизительно 4 дБ. В некоторых случаях 4 дБ — это небольшая цена за простоту реализации, увеличивающуюся от когерентной схемы PSK до некогерентной FSK (рис. 4.25); впрочем, в других случаях ценным является даже выигрыш в 1 дБ. Помимо сложности реализации и вероятности P_B существуют и другие факторы, влияющие на выбор модуляции; например, в некоторых случаях (в каналах со случайным затуханием) желательными являются некогерентные системы, поскольку иногда когерентные опорные сигналы затруднительно определять и использовать. В военных и космических приложениях весьма желательны сигналы, которые могут противостоять значительному ухудшению качества, сохраняя возможность обнаружения.

Таблица 4.1. Вероятность ошибки для различных бинарных модуляций

Модуляция	P_B
PSK (когерентное обнаружение)	$Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$
DPSK (дифференциальное когерентное обнаружение)	$\frac{1}{2}\exp\left(-\frac{E_b}{N_0}\right)$
Ортогональная FSK (когерентное обнаружение)	$Q\left(\sqrt{\frac{E_b}{N_0}}\right)$
Ортогональная FSK (некогерентное обнаружение)	$\frac{1}{2}\exp\left(-\frac{1}{2}\frac{E_b}{N_0}\right)$

4.8. *M*-арная передача сигналов и производительность

4.8.1. Идеальная достоверность передачи

На рис. 3.6 приводился характерный, "водопадоподобный" график зависимости вероятности ошибки от отношения E_b/N_0. Как видно из рис. 4.25, вероятность появления ошибочного бита (P_B) для различных бинарных схем модуляции при наличии AWGN также имеет подобную форму. А на что будет похож график зависимости *идеальной* P_B от E_b/N_0? Ответ, в виде *предела Шеннона*, приведен на рис. 4.27. Этот предел представляет порог E_b/N_0, ниже которого поддержание достоверной связи невозможно. Подробно работа Шеннона рассмотрена в главе 9.

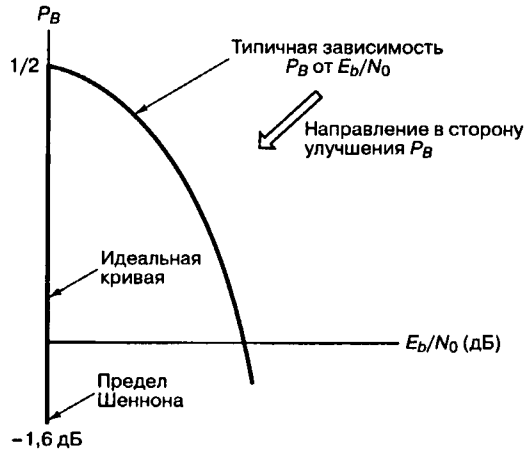


Рис. 4.27. Зависимость идеальной P_B от E_b/N_0

Идеальную кривую на рис. 4.27 можно описать следующим образом. Для всех значений E_b/N_0, находящихся выше предела Шеннона (–1,6 дБ), P_B равно нулю. Как только E_b/N_0 падает ниже предела Шеннона, P_B в худшем случае возрастает до 1/2. (Отметим, что $P_B = 1$ — это не самый неблагоприятный вариант для бинарной передачи сигналов, поскольку это значение аналогично $P_B = 0$; если вероятность появления ошибочного бита равна 100%, то для восстановления точной информации поток битов просто можно инвертировать.) На рис. 4.27 большой стрелкой показано направление повышения достоверности передачи от типичной к идеальной вероятности P_B.

4.8.2. *M*-арная передача сигналов

Рассмотрим *M*-арную передачу сигналов. В каждый момент времени процессор рассматривает k бит. Он указывает модулятору произвести один из $M = 2^k$ сигналов; частным случаем $k = 1$ является бинарная передача сигналов. Как увеличение k влияет на достоверность передачи — снижает или повышает ее? (Не спешите отвечать — вопрос с подвохом.) На рис. 4.28 показана зависимость вероятности появления ошибочного бита $P_B(M)$ от E_b/N_0 для *ортогональной* *M*-уровневой передачи сигналов по каналу с гауссовым шумом при использовании когерентного обнаружения. На рис. 4.29 подобные графики приведены для *многофазной* передачи по каналу с гауссовым шумом при применении когерентного обнаружения. В каком направлении движется график при увеличении k (или M)? Из рис. 4.27 мы

знаем, как изменяется кривая при увеличении и уменьшении вероятности ошибки. Поэтому можем сказать, что на рис. 4.28 по мере роста k график перемещается в направлении уменьшения вероятности ошибки. На рис. 4.29 рост k приводит к увеличению вероятности ошибки. Подобное передвижение свидетельствует, что M-арная передача сигналов уменьшает вероятность ошибки при ортогональной передаче сигналов и увеличивает — при многофазной передаче. Справедливо ли это? Почему вообще используют многофазную модуляцию PSK, если она приводит к высокой вероятности ошибки по сравнению с бинарной PSK? Сказанное действительно справедливо, и во многих системах действительно применяется многофазная передача сигналов. Подвох был в формулировке вопроса: там подразумевалось, что зависимость вероятности ошибки от E_b/N_0 является *единственным* критерием качества. На самом деле существует множество других характеристик (например, ширина полосы, пропускная способность, сложность, стоимость), но на рис. 4.28 и 4.29 явно показана только вероятность ошибки.

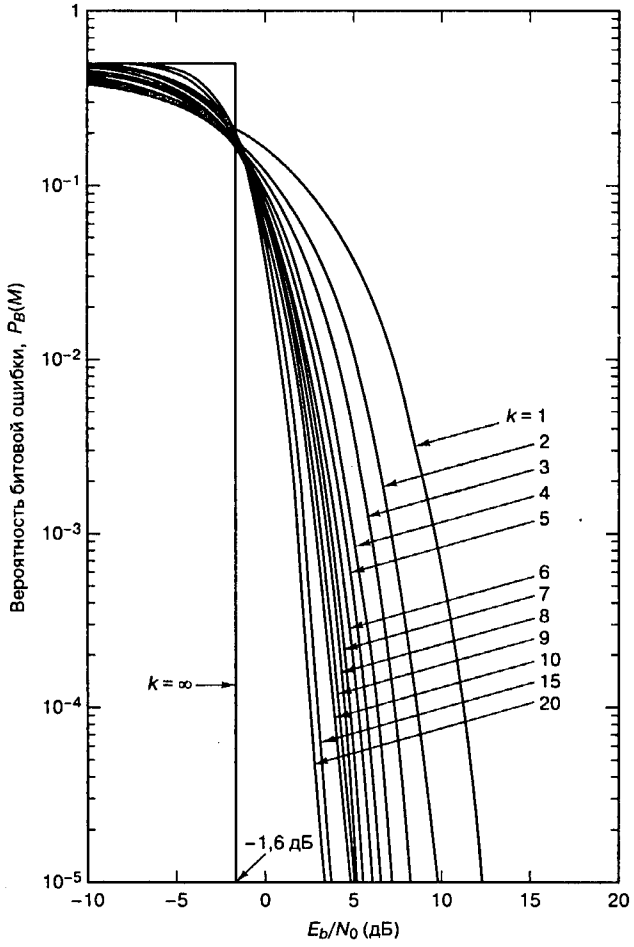


Рис. 4.28. Зависимость $P_B(M)$ от E_b/N_0 для ортогональной M-арной передачи сигналов по каналу с гауссовым шумом при использовании когерентного обнаружения. (Перепечатано с разрешения авторов из работы W. C. Lindsey and M. K. Simon. Telecommunication Systems Engineering. *Prentice Hall, Inc., Englewood Cliffs, N. J.)*

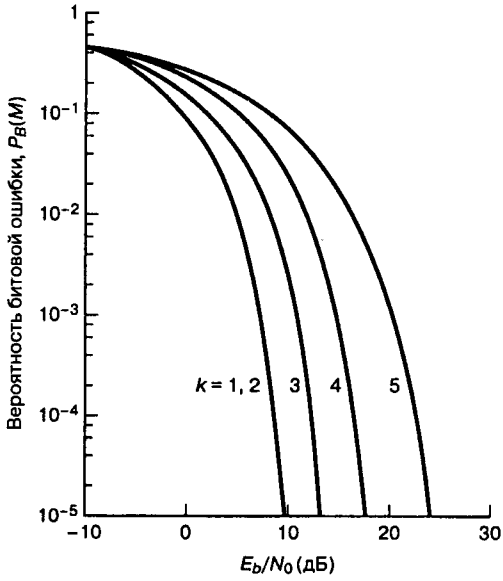


Рис. 4.29. Зависимость $P_B(M)$ от E_b/N_0 для ортогональной многофазной передачи сигналов по каналу с гауссовым шумом при использовании когерентного обнаружения

Одной из рабочих характеристик, не представленных на рис. 4.28 и 4.29 явно, является необходимая ширина полосы. Для графиков на рис. 4.28 повышение значений k подразумевает увеличение требуемой ширины полосы. Для M-арных многофазных кривых, приведенных на рис. 4.29, рост величины k позволяет получать большую скорость передачи битов при той же ширине полосы. Другими словами, при фиксированной скорости передачи данных уменьшается необходимая полоса. Следовательно, графики вероятности ошибки *и* при ортогональной, *и* при многофазной передаче показывают, что M-арная передача сигналов представляет средство реализации компромиссов между параметрами системы. При ортогональной передаче сигналов повышение достоверности передачи может быть получено за счет расширения полосы. В случае многофазной передачи эффективность использования полосы может быть получена за счет вероятности ошибки. Подробнее о компромиссах между полосой и вероятностью ошибки рассказывается в главе 9.

4.8.3. Векторное представление сигналов MPSK

На рис. 4.30 показаны наборы сигналов MPSK для $M = 2$, 4, 8 и 16. На рис. 4.30, *а* видим бинарные ($k = 1$, $M = 2$) антиподные векторы $\mathbf{s}_1$ и $\mathbf{s}_2$, угол между которыми равен 180°. Граница областей решений разделяет сигнальное пространство на две области. На рисунке также показан вектор шума $\mathbf{n}$, равный по амплитуде сигналу $\mathbf{s}_1$. При указанных направлении и амплитуде энергия вектора шума является минимальной, и детектор может допустить символьную ошибку.

На рис. 4.30, *б* видим 4-арные ($k = 2$, $M = 4$) векторы, расположенные друг к другу под углом 90°. Границы областей решений (на рисунке изображена только одна) делят сигнальное пространство на четыре области.

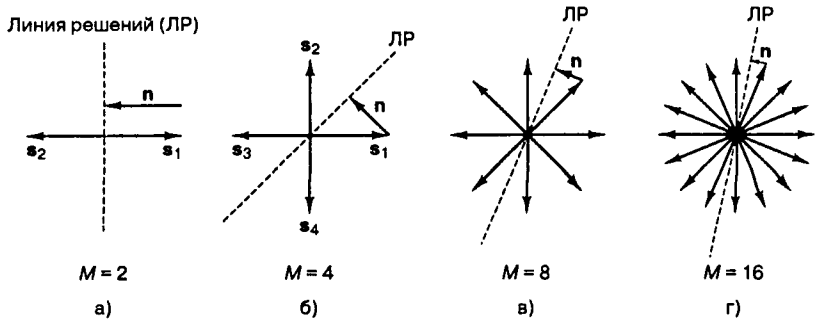


Рис. 4.30. Наборы сигналов MPSK для M = 2, 4, 8, 16

Здесь также изображен вектор шума **n** (начало — в вершине вектора сигнала, направление перпендикулярно ближайшей границе областей решений), являющийся вектором минимальной энергии, достаточной, чтобы детектор допустил символьную ошибку. Отметим, что вектор шума минимальной энергии на рис. 4.30, *б* меньше вектора шума на рис. 4.30, *а*, что свидетельствует о большей уязвимости 4-арной системы к шуму, по сравнению с бинарной (энергии сигналов в обоих случаях взяты равными). Изучая рис. 4.30, *в*, *г*, можно отметить следующую закономерность. При многофазной передаче сигналов по мере роста величины M на сигнальную плоскость помещается все больше сигнальных векторов. По мере того как векторы располагаются плотнее, для появления ошибки вследствие шума требуется все меньше энергии.

С помощью рис. 4.30 можно лучше понять поведение зависимости вероятности P_B от E_b/N_0, изображенной на рис. 4.29, при росте k. Кроме того, рисунок позволяет взглянуть на природу компромиссов при многофазной передаче сигналов. Размещение большего числа векторов сигналов в сигнальном пространстве эквивалентно повышению скорости передачи данных без увеличения системной ширины полосы (все векторы ограничиваются одной и той же плоскостью). Другими словами, мы повысили использование полосы за счет вероятности ошибки. Рассмотрим рис. 4.30, *г*, где из приведенных вариантов вероятность ошибки является наивысшей. Чем мы может заплатить, чтобы "выкупить" возросшую вероятность ошибки? Иными словами, чем мы можем поступиться, чтобы расстояние между соседними векторами сигналов на рис. 4.30, *д* стало таким же, как на рис. 4.30, *а*? Мы можем увеличивать интенсивность сигнала (сделать векторы сигналов длиннее), пока минимальное расстояние от вершины вектора сигнала до линии решений не станет равным размеру вектора шума на рис. 4.30, *а*. Таким образом, для многофазной системы по мере роста M мы можем увеличивать производительность полосы либо за счет повышения вероятности ошибки, либо за счет увеличения отношения E_b/N_0.

Отметим, что на схемах, изображенных на рис. 4.30, *а* для различных значений M, все векторы имеют одинаковую амплитуду. Это равносильно утверждению, что сопоставление различных схем выполняется при фиксированном отношении E_s/N_0, где E_s — энергия символа. Сравнительные схемы можно сделать и при фиксированном отношении E_b/N_0, в этом случае амплитуды векторов будут увеличиваться с ростом M. При $M = 4$, 8 и 16 амплитуды векторов будут, соответственно, в $\sqrt{2}$, $\sqrt{3}$ и 2 раза больше векторов для случая $M = 2$. Как и в предыдущем случае, с ростом M будет усиливаться восприимчивость к шуму, но она не будет такой явной, как на рис. 4.30.

4.8.4. Схемы BPSK и QPSK имеют одинаковые вероятности ошибки

В уравнении (3.30) было получено следующее соотношение между E_b/N_0 и S/N.

$$\frac{E_b}{N_0} = \frac{S}{N}\left(\frac{W}{R}\right) \tag{4.101}$$

Здесь S — средняя мощность сигнала, а R — скорость передачи битов. Вероятность ошибки в сигнале BPSK с отношением E_b/N_0, найденным из уравнения (4.101), определяется из кривой на рис. 4.29, соответствующей $k = 1$. Схему QPSK можно описать с помощью двух ортогональных каналов BPSK. Поток битов QPSK обычно разбивается на четный и нечетный (синфазный и квадратурный) потоки; каждый новый поток модулирует ортогональный компонент несущей со скоростью, вдвое меньшей скорости исходного потока. Синфазный поток модулирует член $\cos \omega_0 t$, а квадратурный — член $\sin \omega_0 t$. Если амплитуда исходного вектора QPSK была равна A, то амплитуды векторов синфазного и квадратурного компонентов равны, как показано на рис. 4.31, $A/\sqrt{2}$. Следовательно, на каждый квадратурный сигнал BPSK приходится половина средней мощности исходного сигнала QPSK. Значит, если исходный сигнал QPSK имел скорость R бит/с и среднюю мощность S Вт, квадратурное разбиение приводит к тому, что каждый сигнал BPSK имеет скорость передачи $R/2$ бит/с и среднюю мощность $S/2$ Вт.

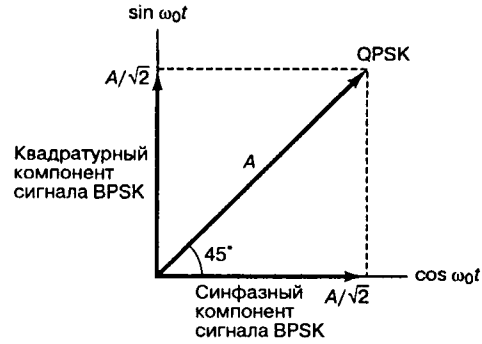


Рис. 4.31. Синфазный и квадратурный компоненты (модуляция BPSK) вектора QPSK

Следовательно, отношение E_b/N_0, характеризующее оба ортогональных канала BPSK, создающих сигнал QPSK, эквивалентно отношению E_b/N_0 в уравнении (4.101), поскольку его можно записать точно так же.

$$\frac{E_b}{N_0} = \frac{S/2}{N_0}\left(\frac{W}{R/2}\right) = \frac{S}{N_0}\left(\frac{W}{R}\right) \tag{4.102}$$

Таким образом, каждый из ортогональных каналов BPSK, а следовательно, и составной сигнал QPSK характеризуются одним отношением E_b/N_0, а значит — такой же вероятностью P_B, что и сигнал BPSK. Ортогональность (разность фаз 90°) соседних символов QPSK приводит к равным вероятностям появления *ошибочного бита* для схем BPSK и QPSK. Следует отметить, что вероятности появления *ошибочного символа* для этих схем *не равны*. Связь этих двух вероятностей рассмотрена

в разделах 4.9.3 и 4.9.4. Там будет показано, что схема QPSK эквивалентна двум квадратурным каналам BPSK. Этот результат будет расширен на все симметричные схемы передачи сигналов с модуляцией амплитуды/фазы, подобные квадратурной амплитудной модуляции (quadrature amplitude modulation — QAM), описанной в разделе 9.8.3.

4.8.5. Векторное представление сигналов MFSK

В разделе 4.8.3 мы исследовали рис. 4.30, что позволило получить представление о причинах роста вероятности ошибки при увеличении числа k (или M) в схеме MPSK. Полезно будет рассмотреть подобную векторную иллюстрацию для схемы MFSK, которая позволит лучше понять графики на рис. 4.28. Поскольку сигнальное пространство MFSK описывается M взаимно перпендикулярными осями, мы без труда можем проиллюстрировать случаи $M = 2$ и $M = 3$. Итак, на рис. 4.32, *а* видим бинарные ортогональные векторы $\mathbf{s}_1$ и $\mathbf{s}_2$. Граница областей решений разбивает сигнальное пространство на две области. На рисунке также показан вектор шума $\mathbf{n}$, представляющий минимальный вектор, который может привести к принятию неправильного решения.

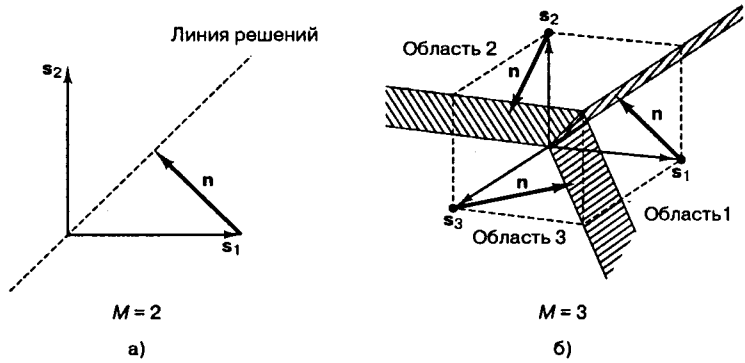


Рис. 4.32. Наборы сигналов MFSK для M = 2, 3

На рис. 4.32, *б* показано трехмерное сигнальное пространство со взаимно перпендикулярными координатными осями. В этом случае плоскости решений разбивают пространство на три области. Показано, как к каждому сигнальному вектору $\mathbf{s}_1$, $\mathbf{s}_2$ и $\mathbf{s}_3$ прибавляется вектор шума $\mathbf{n}$, представляющий минимальный вектор, который может привести к принятию неправильного решения. Векторы шума на рис. 4.32, *б* имеют тот же модуль, что и вектор шума, показанный на рис. 4.32, *а*. В разделе 4.4.4 мы утверждали, что при данном уровне принятой энергии расстояние между любыми двумя векторами сигналов-прототипов $\mathbf{s}_i$ и $\mathbf{s}_j$ M-мерного ортогонального пространства является константой. Отсюда следует, что минимальное расстояние между вектором сигнала-прототипа и любой границей решений не меняется с изменением M. В отличие от модуляции MPSK, когда добавление нового сигнала к сигнальному множеству делало сигналы более уязвимыми к меньшим векторам шума, при MFSK такого *не происходит*.

Для иллюстрации этого момента можно было бы нарисовать ортогональные пространства высших размерностей, но, к сожалению, это затруднительно. Мы можем использовать только наш "мысленный взгляд", чтобы понять, что увели-

чение сигнального множества M — путем введения дополнительных осей, причем каждая новая ось перпендикулярна всем существующим — не приводит к его уплотнению. Следовательно, переданный сигнал, принадлежащий ортогональному набору, *не* становится более уязвимым к шуму при увеличении размерности. Фактически, как можно видеть из рис. 4.28, при увеличении k вероятность появления ошибочного бита даже уменьшается.

Пониманию улучшения надежности при ортогональной передаче сигналов, показанного на рис. 4.28, способствует сравнение зависимости вероятности символьной ошибки (P_E) от ненормированного отношения сигнал/шум (signal-to-noise ratio — SNR) с зависимостью P_E от E_b/N_0. На рис. 4.33 для когерентной передачи сигналов FSK представлено несколько зависимостей P_E от нормированного SNR. Видим, что P_E падает с ростом M. Можем ли мы сказать, что сигнал из ортогонального набора *не становится* более уязвимым к данному шуму при увеличении размерности ортогонального набора? Для ортогональной передачи сигналов справедливо утверждение, что при данном SNR вектора шума фиксированного размера достаточно для перевода переданного сигнала в область ошибок; следовательно, сигналы не становятся более уязвимыми к меньшим векторам шума при увеличении M. В то же время при росте M вводится большее число окрестных областей решений; следовательно, увеличивается число возможностей для появления символьной ошибки, всего существует $(M-1)$ возможностей допустить ошибку. На рис. 4.33 отражено ухудшение P_E в зависимости от ненормированного SNR при увеличении M. Стоит отметить, что изучение зависимости достоверности передачи от M при фиксированном SNR не является лучшим направлением в цифровой связи. Фиксированное SNR означает фиксированный объем энергии на символ; следовательно, при увеличении M этот объем энергии необходимо распределять уже между большим числом битов, т.е. на каждый бит приходится меньше энергии. В этой связи наиболее удобным способом сравнения различных цифровых систем является использование в качестве критерия *отношения сигнал/шум, нормированного на бит*, или E_b/N_0. Повышение достоверности передачи с увеличением M (см. рис. 4.28) проявляется только в том случае, если вероятность ошибки изображается как зависимость от E_b/N_0. В этом случае при увеличении M отношение E_b/N_0, требуемое для получения заданной вероятности ошибки, снижается при фиксированном SNR; следовательно, нам нужен новый график, подобный показанному на рис. 4.28, где ось абсцисс представляет не SNR, а E_b/N_0. На рис. 4.34 показано, как зависимость от SNR отображается в зависимость от E_b/N_0; видно, как графики, демонстрирующие ухудшение P_E с увеличением M (подобно представленному на рис. 4.33), преобразуются в графики, показывающие улучшение P_E с увеличением M. Само преобразование выполняется согласно соотношению, приведенному в формуле (4.101).

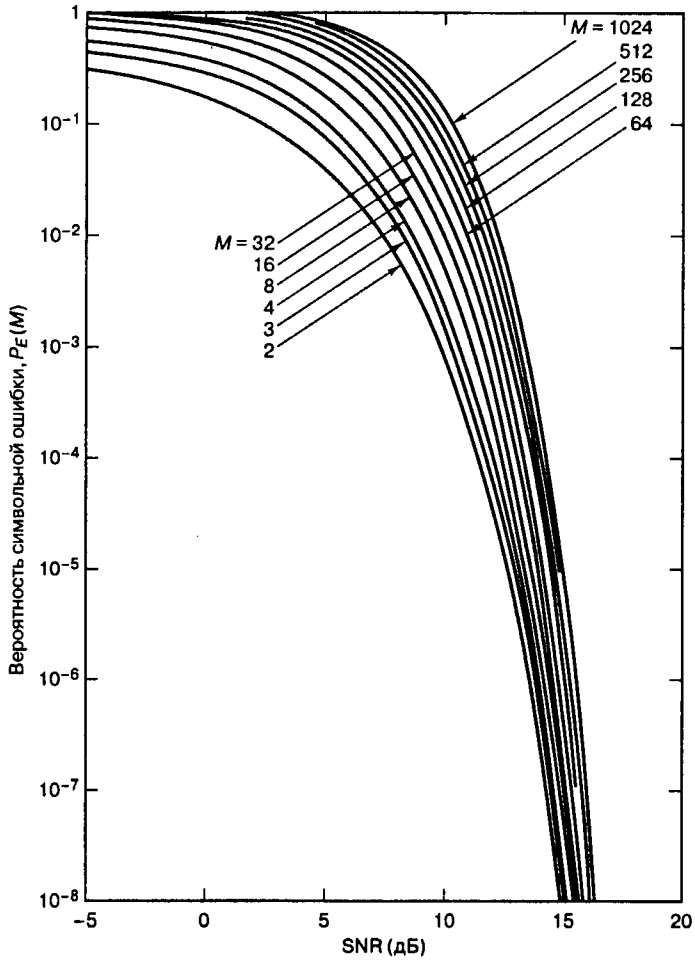


Рис. 4.33. Зависимость вероятности символьной ошибки от SNR для когерентной передачи сигналов FSK. (Из документа Bureau of Standards. Technical Note 167, *March, 1963; перепечатано с разрешения National Bureau of Standards из* Central Radio Propagation Laboratory Technical Note 167, *March, 25, 1963, Fig. 1, p. 2.)*

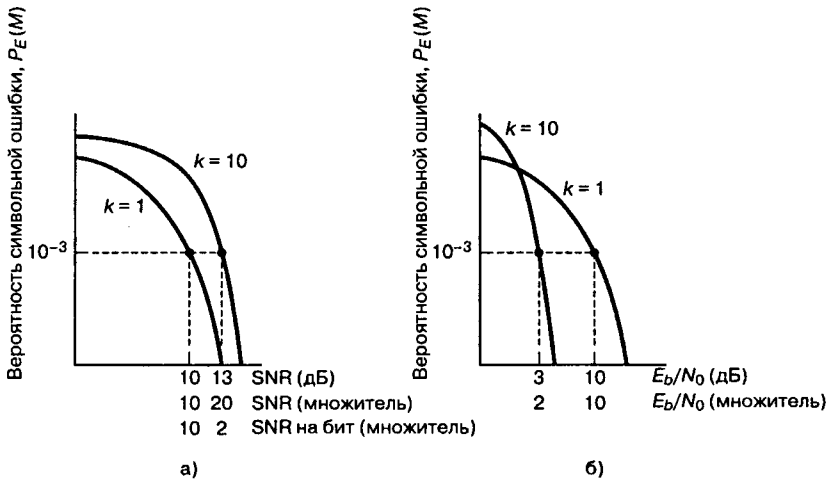


Рис. 4.34. Отображение зависимости P_E от SNR в зависимость P_E от E_b/N_0 для ортогональной передачи сигналов: а) ненормированная зависимость; б) нормированная зависимость

$$\frac{E_b}{N_0} = \frac{S}{N}\left(\frac{W}{R}\right)$$

Здесь W — ширина полосы обнаружения. Поскольку

$$R = \frac{\log_2 M}{T} = \frac{k}{T},$$

где T — длительность символа, можем записать следующее.

$$\frac{E_b}{N_0} = \frac{S}{N}\left(\frac{WT}{\log_2 M}\right) = \frac{S}{N}\left(\frac{WT}{k}\right) \tag{4.103}$$

При передаче сигналов FSK ширина полосы обнаружения W (в герцах) обычно равна скорости передачи символов $1/T$; другими словами, $TW \approx 1$. Следовательно,

$$\frac{E_b}{N_0} \approx \frac{S}{N}\left(\frac{1}{k}\right) \tag{4.104}$$

На рис. 4.34 представлено отображение зависимости P_E от SNR в зависимость P_E от E_b/N_0 для M-мерной ортогональной передачи сигналов с когерентным обнаружением; на осях показано сопоставление величин разных размерностей. На рис. 4.34, *а* выбрана рабочая точка, соответствующая отношению сигнал/шум = 10 дБ схемы с $k = 1$, при данной вероятности ошибки $P_E = 10^{-3}$. В той же системе координат приведен график схемы с $k = 10$; рабочая точка, соответствующая той же величине $P_E = 10^{-3}$, теперь соответствует отношению сигнал/шум, равному 13 дБ (приблизительное значение, полученное из рис. 4.33). Из приведенных графиков явно видно снижение достоверности при увеличении k. Чтобы понять, как улучшается производительность, преобразуем масштаб оси абсцисс из нелинейного (отношение сигнал/шум в децибелах) в

линейный (SNR как коэффициент). На рис. 4.34, *а* показано, как соотносятся значения SNR в децибелах (10 и 13) со значениями, представленными как коэффициент (10 и 20), для случаев $k = 1$ и $k = 10$. Далее преобразуем масштаб оси абсцисс, чтобы единицами измерения служило отношение сигнал/шум, нормированное на бит (также выраженное как коэффициент). Этому случаю на рис. 4.34, *а* соответствуют величины 10 и 2 для $k = 1$ и $k = 10$. Вообще, удобно не различать 1024-ричный символ или сигнал (случай $k = 10$) и его 10-битовое значение. При таком подходе, если символ требует 20 единиц SNR, то 10 бит, кодирующих этот символ, требуют тех же 20 единиц; другими словами, каждый бит требует двух единиц отношения сигнал/шум.

Вместо подобного сравнения, можно просто отобразить рассматриваемые случаи $k = 1$ и $k = 10$ графиками, изображенными на рис. 4.34, *б* и представляющими зависимости P_E от E_b/N_0. Случай $k = 1$ соответствует представленному на рис. 4.34, *а*. Но для случая $k = 10$ наблюдаем разительные отличия. Видим, что при $k = 10$ передача 10-битового символа требует всего 2 единиц (3 дБ) отношения E_b/N_0 по сравнению с 10 единицами (10 дБ) для бинарного символа. Действительно, из формулы (4.104) получаем значение отношения $E_b/N_0 = 20\,(1/10) = 2$ (или 3 дБ), т.е. имеем повышение достоверности при увеличении k. В системах цифровой связи достоверность передачи (или вероятность ошибки) всегда выражается через E_b/N_0, поскольку такой подход позволяет выполнять сравнение производительности различных систем. Графики, приведенные на рис. 4.33 и 4.34, *а*, на практике встречаются крайне редко.

Хотя изображенные на рис. 4.33 зависимости и не используются на практике часто, все же с помощью этого рисунка мы можем понять, почему ортогональная передача сигналов приводит к повышению достоверности при увеличении M или k. Рассмотрим аналогию — приобретение товара, скажем прессованного творога высшего качества. Выбор качества соответствует выбору точки на оси P_E рис. 4.33, скажем 10^{-3}. Проведем из этой точки горизонтальную линию через все кривые (от $M = 2$ до $M = 1024$). В бакалейно-гастрономическом отделе мы покупаем самую маленькую упаковку прессованного творога, которая содержит 2 унции и стоит \$1. Обращаясь к рис. 4.33, можем сказать, что такая покупка соответствует пересечению проведенной горизонтальной линии с графиком для $M = 2$. Смотрим вниз на соответствующее значение параметра SNR и называем пересечение с этой осью ценой \$1. При следующем походе за покупками мы решаем, что в прошлый раз стоимость творога была высокой — по 50 центов за унцию. Поэтому решаем купить большую упаковку (8 унций) за \$2. Обращаемся к рис. 4.33 и видим, что данная покупка соответствует пересечению горизонтальной линии с кривой $M = 8$. Смотрим вниз и называем соответствующее значение SNR ценой \$2. Замечаем, что хотя мы и купили большую емкость, заплатив за нее большую цену, все же стоимость одной унции упала (и составляет теперь всего 25 центов). Эту аналогию можно продолжать; мы можем приобретать все большие и большие упаковки, при этом их цена (SNR) будет расти, а стоимость за унцию будет падать. Вообще, это известно давно и называется *эффектом масштаба*: приобретение за раз большого количества товара соответствует закупкам по оптовым ценам; при этом цена единицы товара падает. Подобным образом при использовании ортогональной передачи сигналов с символами, содержащими большее число бит, нам требуется большая мощность (большее отношение SNR), а требования относительно бита (E_b/N_0) при этом снижаются.

4.9. Вероятность символьной ошибки для *M*-арных систем (*M* > 2)

4.9.1. Вероятность символьной ошибки для модуляции MPSK

Для больших отношений сигнал/шум вероятность символьной ошибки $P_E(M)$ для равновероятных сигналов в M-арной модуляции PSK с когерентным обнаружением можно выразить как [7]

$$P_E(M) \approx 2Q\left(\sqrt{\frac{2E_s}{N_0}}\sin\frac{\pi}{M}\right), \tag{4.105}$$

где $P_E(M)$ — вероятность символьной ошибки, $E_s = E_b(\log_2 M)$ — энергия, приходящаяся на символ, а $M = 2^k$ — размер множества символов. Зависимость $P_E(M)$ от E_b/N_0 для передачи сигналов MPSK с когерентным обнаружением показана на рис. 4.35.

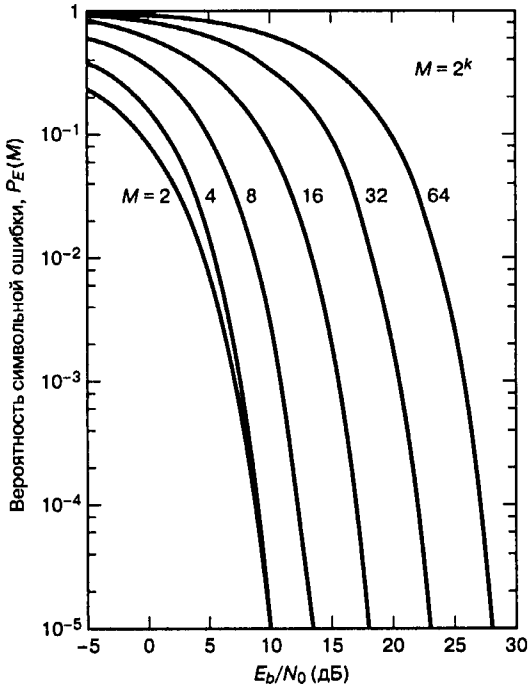


Рис. 4.35. Вероятность символьной ошибки для многофазной передачи сигналов с когерентным обнаружением. (Перепечатано с разрешения авторов из W. C. Lindsey and M. K. Simon. Telecommunication Systems Engineering. *Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973.)*

Вероятность символьной ошибки для дифференциального когерентного обнаружения M-арной схемы DPSK (для больших значений E_s/N_0) выражается подобно тому, как это было приведено выше [7].

$$P_E(M) \approx 2Q\left(\sqrt{\frac{2E_s}{N_0}} \sin\frac{\pi}{\sqrt{2}M}\right) \tag{4.106}$$

4.9.2. Вероятность символьной ошибки для модуляции MFSK

Вероятность символьной ошибки $P_E(M)$ для равновероятных ортогональных сигналов с *когерентным* обнаружением можно выразить как [5]

$$P_E(M) \leq (M-1)Q\left(\sqrt{\frac{E_s}{N_0}}\right), \tag{4.107}$$

где $E_s = E_b(\log_2 M)$ — энергия, приходящаяся на символ, а M — размер множества символов. Зависимость $P_E(M)$ от E_b/N_0 для M-арных ортогональных сигналов с когерентным обнаружением показана на рис. 4.36.

Вероятность символьной ошибки для равновероятных M-арных ортогональных сигналов с *некогерентным* обнаружением дается следующим выражением [9].

$$P_E(M) = \frac{1}{M}\exp\left(-\frac{E_s}{N_0}\right)\sum_{j=2}^{M}(-1)^j\binom{M}{j}\exp\left(\frac{E_s}{jN_0}\right), \tag{4.108}$$

где

$$\binom{M}{j} = \frac{M!}{j!(M-j)!} \tag{4.109}$$

является стандартным биномиальным коэффициентом, выражающим число способов выбора j ошибочных символов из M возможных. Отметим, что для бинарного случая формула (4.108) сокращается до

$$P_B = \frac{1}{2}\exp\left(-\frac{E_b}{2N_0}\right), \tag{4.110}$$

что совпадает с результатом, полученным в выражении (4.96). Кривая зависимости $P_E(M)$ от E_b/N_0 для M-арной передачи сигналов с некогерентным обнаружением изображена на рис. 4.37. При сравнении данных графиков с приведенными на рис. 4.6 и соответствующими когерентному обнаружению можно заметить, что для $k > 7$ различием уже можно пренебрегать. В заключение отметим, что для когерентного и некогерентного приема ортогональных сигналов верхний предел вероятности ошибки дается выражением [9].

$$P_E(M) < \frac{M-1}{2}\exp\left(-\frac{E_s}{2N_0}\right) \tag{4.111}$$

Здесь E_s — энергия на символ, а M — размер алфавита символов.

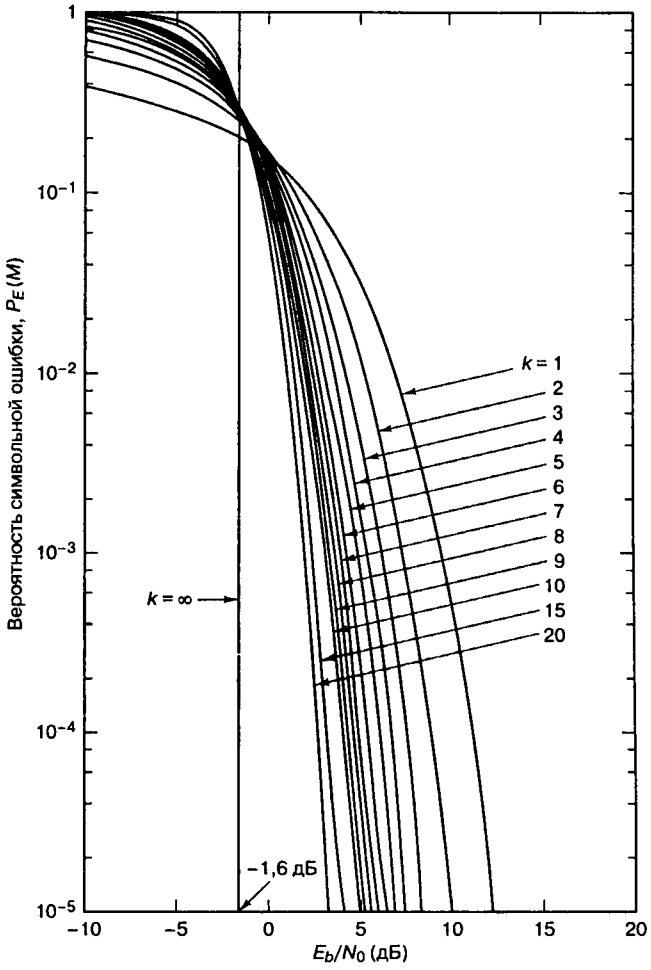


Рис. 4.36. Вероятность символьной ошибки для M-арной ортогональной передачи сигналов с когерентным обнаружением. (Перепечатано с разрешения авторов из W. C. Lindsey and M. K. Simon. Telecommunication Systems Engineering. *Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973.)*

4.9.3. Зависимость вероятности битовой ошибки от вероятности символьной ошибки для ортогональных сигналов

Можно показать [9], что соотношение между вероятностью битовой ошибки (P_B) и вероятностью символьной ошибки (P_E) для ортогональных M-арных сигналов дается следующим выражением.

$$\frac{P_B}{P_E} = \frac{2^{k-1}}{2^k - 1} = \frac{M/2}{M-1} \tag{4.112}$$

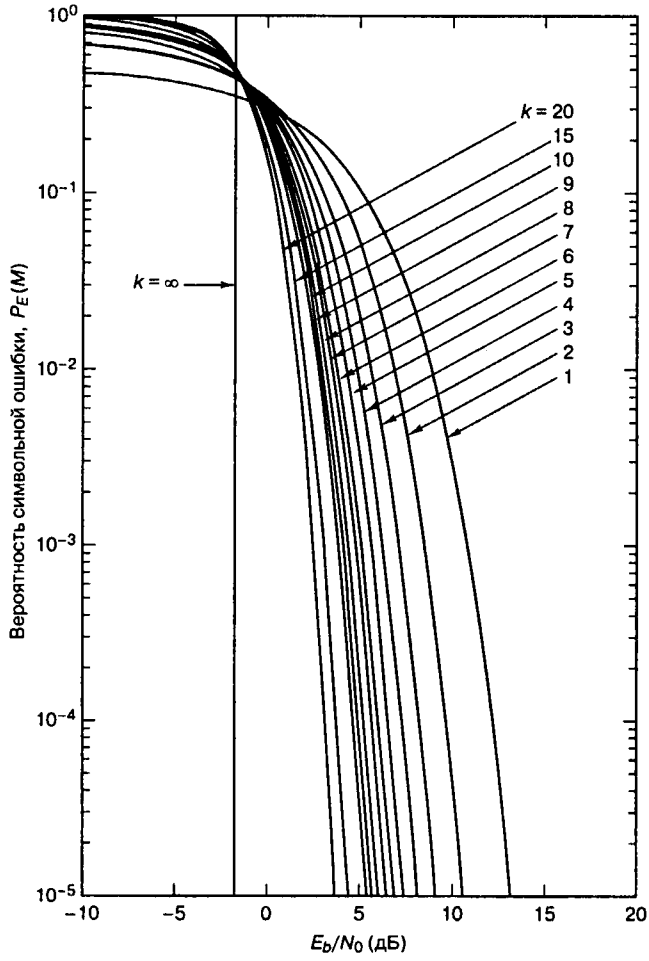


Рис. 4.37. Вероятность символьной ошибки для M-арной ортогональной передачи сигналов с некогерентным обнаружением. (Перепечатано с разрешения авторов из W. C. Lindsey and M. K. Simon. Telecommunication Systems Engineering. *Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973.)*

В пределе при увеличении k получаем следующее.

$$\lim_{k \to \infty} \frac{P_B}{P_E} = \frac{1}{2}$$

Понять формулу (4.112) позволяет простой пример. На рис. 4.38 показан восьмеричный набор символов сообщения. Эти символы (предполагаемые равновероятными) передаются с помощью ортогональных сигналов, таких как сигналы FSK. При использовании ортогональной передачи ошибка принятия решения равновероятно преобразует верный сигнал в один из $(M-1)$ неверных. Пример на рисунке демонстрирует передачу символа, состоящего из битов 0 1 1. Ошибка с равной вероятностью

может перевести данный символ в любой из оставшихся $2^k - 1 = 7$ символов. Отметим, что наличие ошибки еще не означает, что *все* биты символа являются ошибочными. Если (рис. 4.38) приемник решит, что переданным символом является нижний из указанных, состоящий из битов 1 1 1, два из трех переданных битов будут верными. Должно быть очевидно, что для недвоичной передачи P_B всегда будет меньше P_E (P_B и P_E — *средние* частоты появления ошибок).

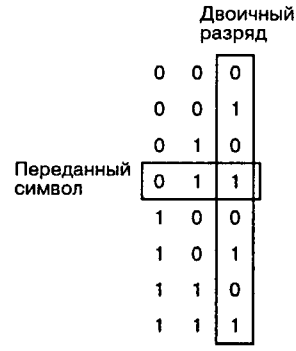


Рис. 4.38. Пример зависимости P_B от P_E

Рассмотрим любой из столбцов битов на рис. 4.38. Каждая битовая позиция на 50% заполнена нулями и на 50% — единицами. Рассмотрим первый бит переданного символа (правый столбец). Сколько существует возможностей появления ошибочного бита 1? Всего существует $2^k - 1 = 4$ возможности (нули в столбце появляются в четырех местах) появления битовой ошибки; то же значение получаем для каждого столбца. Окончательное соотношение P_B/P_E для ортогональной передачи сигналов в формуле (4.112) получается следующим образом: число возможностей появления битовой ошибки (2^{k-1}) делится на число возможностей появления символьной ошибки ($2^k - 1$). Для случая, изображенного на рис. 4.38, $P_B/P_E = 4/7$.

4.9.4. Зависимость вероятности битовой ошибки от вероятности символьной ошибки для многофазных сигналов

При передаче сигналов MPSK значение P_B меньше или равно P_E, так же как и при передаче сигналов MFSK. В то же время имеется и существенное отличие. Для ортогональной передачи сигналов выбор одного из $(M - 1)$ ошибочных символов равновероятен. При передаче в модуляции MPSK каждый сигнальный вектор не является равноудаленным от всех остальных. На рис. 4.39, *а* показано восьмеричное пространство решений, где области решений обозначены 8-ричными символами в двоичной записи. При передаче символа (0 1 1) и появлении в нем ошибки наиболее вероятными являются ближайшие соседние символы, (0 1 0) и (1 0 0). Вероятность превращения символа (0 1 1) вследствие ошибки в символ (1 1 1) относительно мала. Если биты распределяются по символам согласно двоичной последовательности, показанной на рис. 4.39, *а*, то некоторые символьные ошибки всегда будут давать две (или более) битовые ошибки, даже при значительном отношении сигнал/шум.

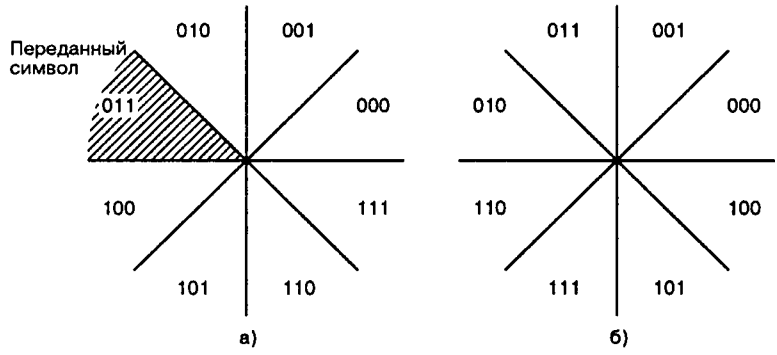


Рис. 4.39. Области решения в сигнальном пространстве MPSK: а) в бинарной кодировке; б) в кодировке Грея

Для неортогональных схем, таких как MPSK, часто используется код преобразования бинарных символов в M-арные, такие, что двоичные последовательности, соответствующие соседним символам (сдвигам фаз), отличаются единственной битовой позицией; таким образом, при появлении ошибки в M-арном символе высока вероятность того, что ошибочным является только один из k прибывших битов. Кодом, обеспечивающим подобное свойство, является код Грея (Gray code) [7]; на рис. 4.39, *б* для восьмеричной схемы PSK показано распределение битов по символам с использованием кода Грея. Можно видеть, что соседние символы отличаются одним двоичным разрядом. Следовательно, вероятность появления многобитовой ошибки при данной символьной ошибке значительно меньше по сравнению с некодированным распределением битов, показанным на рис. 4.39, *а*. Реализация подобного кода Грея представляет один из редких случаев в цифровой связи, когда определенная выгода может быть получена без сопутствующих недостатков. Код Грея — это просто присвоение, не требующее специальных или дополнительных схем. Можно показать [5], что при использовании кода Грея вероятность ошибки будет следующей.

$$P_B \approx \frac{P_E}{\log_2 M} \qquad (\text{для } P_E << 1) \tag{4.113}$$

Напомним из раздела 4.8.4, что передача сигналов BPSK и QPSK имеет одинаковую вероятность битовой ошибки. Формула (4.113) доказывает, что вероятности символьных ошибок этих схем отличаются. Для модуляции BPSK $P_E = P_B$, а для QPSK $P_E \approx 2P_B$.

Точное аналитическое выражение вероятности битовой ошибки P_B в восьмеричной схеме PSK, а также довольно точные аппроксимации верхнего и нижнего пределов P_B для M-арной PSK при больших M можно найти в работе [10].

4.9.5. Влияние межсимвольной интерференции

В предыдущем разделе и в главе 3 обнаружение сигналов рассматривалось при наличии шума AWGN в предположении, что межсимвольная интерференция (intersymbol interference — ISI) отсутствует. Это упростило анализ, поскольку процесс AWGN с нулевым средним описывается единственным параметром — дисперсией. На практике обычно оказывается, что межсимвольная интерференция — это второй (после теплового шума) источник помех, которому необходимо уделять пристальное внимание. Как объяснялось в разделе 3.3, ISI может возникать вследствие использования узко-

полосных фильтров на выходе передатчика, в канале или на входе приемника. Результатом этой дополнительной интерференции является ухудшение достоверности передачи как для когерентного, так и некогерентного приема. Вычисление вероятности ошибки при ISI (помимо AWGN) является значительно более сложной задачей, поскольку в вычислениях будет фигурировать импульсная характеристика канала. Этот вопрос мы не рассматриваем; впрочем, для читателей, интересующихся данной темой, можно порекомендовать работы [11–16].

4.10. Резюме

В данной главе систематизированы некоторые основные форматы полосовой цифровой модуляции, в частности фазовая манипуляция (phase shift keying — PSK) и частотная манипуляция (frequency shift keying — FSK). Здесь рассмотрено геометрическое представление векторов сигналов и шумов, в частности антиподных и ортогональных множеств сигналов. Данное геометрическое представление позволило рассмотреть проблему обнаружения в ортогональном сигнальном пространстве и областях сигналов. Это представление и графическое изображение воздействия векторов шума, способных перевести переданные сигналы в ложную область, способствуют пониманию проблемы обнаружения и достоверности различных методов модуляции/демодуляции. В главе 9 вопрос модуляции и демодуляции будет рассмотрен повторно; также будут исследованы некоторые методы модуляции, повышающие эффективность использования полосы.

Литература

1. Schwartz M. *Information, Transmission, Modulation, and Noise.* McGraw-Hill Book Company, New York, 1970.
2. Van Trees H. L. *Detection, Estimation, and Modulation Theory.* Part 1, John Wiley & Sons, Inc., New York, 1968.
3. Park J. H., Jr. *On Binary DPSK Detection.* IEEE Trans. Commun., vol. COM26, n. 4, April, 1978, pp.484–486.
4. Ziemer R. E. and Peterson R. L. *Digital Communications and Spread Spectrum systems.* Macmillan Publishing Company, Inc., New York, 1985.
5. Lindsey W. C. and Simon M. K. *Telecommunication Systems Engineering.* Prentice-Hall, Inc. Englewood Cliffs, N. J., 1973.
6. Whalen A. D. *Detection of Signals in Noise.* Academic Press, Inc., New York, 1971.
7. Korn I. *Digital Communications.* Van Nostrand Reinhold Company, Inc., New York, 1985.
8. Couch L. W. II. *Digital and Analog Communication Systems.* Macmillan Publishing Company, New York, 1983.
9. Viterbi A. J. *Principles of Coherent Communications.* McGraw-Hill Book Company, New York, 1966.
10. Lee P. J. *Computation of the Bit Error Rate of Coherent M-ary PSK with Gray Code Bit Mapping.* IEEE Trans. Commun., vol. COM34, n. 5, May, 1986, pp. 488–491.
11. Hoo E. Y. and Yeh Y. S. *A New Approach for Evaluating the Error Probability in the Presence of the Intersymbol Interference and Additive Gaussian Noise.* Bell Syst. Tech. J., vol. 49, November, 1970, pp. 2249–2266.
12. Shimbo O., Fang R. J. and Celebiler M. *Performance of M-ary PSK Systems on Gaussian Noise and Intersymbol Interference.* IEEE Trans. Inf. Theory, vol. IT19, January, 1973, pp. 44–58.
13. Prabhu V. K. *Error Probability Performance of M-ary CPSK Systems with Intersymbol Interference.* IEEE Trans. Commun., vol. COM21, February, 1973, pp. 97–109.
14. Yao K. and Tobin R. M. *Moment Space Upper and Lower Error Bounds for Digital Systems with Intersymbol Interference.* IEEE Trans. Inf. Theory, vol. IT22, January, 1976, pp. 65–74.

15. King M. A., Jr. *Three Dimensional Geometric Moment Bounding Techniques.* J. Franklin Inst., vol. 309, n. 4, April, 1980, pp. 195–213.
16. Prabhu V. K. and Salz J. *On the Performance of Phase-Shift Keying Systems.* Bell Syst. Tech. J., vol. 60, December, 1981, pp. 2307–2343.

Задачи

4.1. Определите точное число битовых ошибок, сделанных за сутки когерентным приемником, использующим схему BPSK. Скорость передачи данных равна 5000 бит/с. Входящими цифровыми сигналами являются: $s_1(t) = A \cos \omega_0 t$ и $s_2(t) = -A \cos \omega_0 t$, где $A = 1$ мВ, а односторонняя спектральная плотность мощности шума равна $N_0 = 10^{-11}$ Вт/Гц. Считайте, что мощность сигнала и энергия, приходящаяся на бит, нормированы на нагрузку с сопротивлением 1 Ом.

4.2. Непрерывно работающая когерентная система BPSK совершает ошибки со средней частотой 100 ошибок в сутки. Скорость передачи данных 1000 бит/с. Односторонняя спектральная плотность мощности равна $N_0 = 10^{-10}$ Вт/Гц.

а) Чему равна средняя вероятность ошибки, если система является эргодической?

б) Если значение средней мощности принятого сигнала равно 10^{-6} Вт, будет ли ее достаточно для поддержания вероятности ошибки, найденной в п. а?

4.3. Если основным критерием производительности системы является вероятность битовой ошибки, какую из следующих двух схем следует выбрать для канала с шумом AWGN? Приведите соответствующие вычисления.

Бинарная некогерентная ортогональная схема FSK с $E_b/N_0 = 13$ дБ
Бинарная когерентная схема PSK с $E_b/N_0 = 8$ дБ

4.4. Поток битов

1 0 1 0 1 0 1 1 1 1 0 1 0 1 0 1 0 0 0 0 0 1 1 1 1

передается с использованием модуляции DPSK. Покажите четыре различные дифференциально-кодированные последовательности, которые могут представлять данное сообщение, и объясните алгоритм генерации каждой из них.

4.5. а) Вычислите минимальную требуемую полосу для некогерентного обнаружения символов в ортогональной бинарной модуляции FSK. Сигнальный тон наивысшей частоты *равен 1 МГц, а длительность символа равна 1 мс.*

б) Чему равна минимальная требуемая полоса для некогерентной системы MFSK с той же продолжительностью символа?

4.6. Рассмотрим систему BPSK с равновероятными сигналами $s_1(t) = \cos \omega_0 t$ и $s_2(t) = -\cos \omega_0 t$. Будем считать, что отношение сигнал/шум в приемнике равно $E_b/N_0 = 9{,}6$ и при идеальной синхронизации вероятность битовой ошибки равна 10^{-5}. Допустим, восстановление несущей с использованием контура ФАПЧ вносит некоторую фиксированную ошибку ϕ, связанную с оценкой фазы, так что опорные сигналы выражаются как $\cos(\omega_0 t + \phi)$ и $-\cos(\omega_0 t + \phi)$. Отметим, что эффект ухудшения достоверности вследствие известного фиксированного смещения можно вычислить, используя аналитические выражения, данные в тексте главы. В то же время, если ошибка фазы будет включать случайное смещение, вычисление его воздействия потребует стохастического рассмотрения (см. главу 10).

а) Насколько возрастет вероятность битовой ошибки при $\phi = 25°$?

б) Какая ошибка в определении фазы приведет к росту вероятности битовой ошибки до 10^{-3}?

4.7. Определите вероятность появления ошибочного бита P_B для когерентного обнаружения с использованием согласованного фильтра равновероятных сигналов FSK.

$$s_1(t) = 0{,}5 \cos 2000\pi t$$

и

$$s_2(t) = 0{,}5 \cos 2020\pi t$$

Здесь двусторонняя спектральная плотность мощности шума AWGN равна $N_0/2 = 0{,}0001$. Длительность символа считать равной $T = 0{,}01$ с.

4.8. Определите оптимальный (дающий минимальную вероятность ошибки) порог γ_0 для обнаружения равновероятных сигналов $s_1(t) = \sqrt{2E/T}\cos\omega_0 t$ и $s_2(t) = \sqrt{\frac{1}{2}E/T}\cos(\omega_0 t + \pi)$ в шуме AWGN при использовании корреляционного приемника, изображенного на рис. 4.7, *б*. В качестве опорного возьмите сигнал $\psi_1(t) = \sqrt{2/T}\cos\omega_0 t$.

4.9. Система обнаружения с помощью согласованного фильтра равновероятных сигналов $s_1(t) = \sqrt{2E/T}\cos\omega_0 t$ и $s_2(t) = \sqrt{2E/T}\cos(\omega_0 t + \pi)$ работает при шуме AWGN при отношении $E_b/N_0 = 6{,}8$ дБ. Считать, что $\mathbf{E}\{z(T)\} = \pm\sqrt{E}$.

а) Найдите минимальную вероятность ошибки P_B для данного отношения E_b/N_0 и данного множества сигналов.

б) Найдите P_B, если порог принятия решения равен $\gamma = 0{,}1\sqrt{E}$.

в) Порог $\gamma = 0{,}1\sqrt{E}$ является оптимальным для определенного множества априорных вероятностей $P(s_1)$ и $P(s_2)$. Найдите значения этих вероятностей (используйте раздел Б.2).

4.10. а) Опишите импульсную характеристику согласованного фильтра, используемого для обнаружения дискретного сигнала, изображенного на рис. 34.1. Какой сигнал на выходе фильтра получится при подаче данного сигнала на вход? Воздействием шума можно пренебречь. Чему равно максимальное значение на выходе?

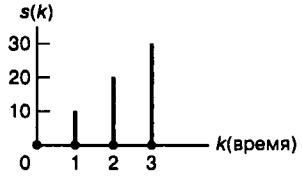


Рис. 34.1

б) В согласованном фильтре сигнал сворачивается с обращенной во времени функцией сигнала (импульсной характеристикой согласованного фильтра). Свертка еще раз обращает функцию; таким образом, согласованный фильтр выдает корреляцию сигнала и его копии (несмотря на то что работа согласованного фильтра описывается операцией свертки). Предположим, что при реализации согласованного фильтра вы случайно соединили каналы так, что фильтр дает корреляцию сигнала и его обращенной во времени копии. Покажите выход как функцию времени. Чему равно максимальное значение на выходе? Отметим, что при данных условиях максимальное значение на выходе появляется в другой момент времени, чем в п. а.

в) С помощью значений на выходе неверного фильтра, описанного в п. б, по сравнению с корректными значениями из п. а, можно ли найти ключ, который поможет предсказать, появляется ли некоторая последовательность с выхода правильного или неправильного фильтра?

г) Пусть к сигналу добавлен шум. Сравните отношение SNR на выходе коррелятора и устройства свертки. Пусть выход состоит исключительно из шума. Сравните выходы коррелятора и устройства свертки.

4.11. Двоичный источник с равновероятными символами управляет положением коммутатора приемника, работающего в канале с шумом AWGN (рис. 34.2) Двусторонняя спектральная плотность шума равна $N_0/2$. Пусть передаются антиподные сигналы длительностью T секунд с энергией E Дж. Системная схема синхронизации каждые T секунд генерирует синхронизирующие импульсы, а скорость передачи двоичного источника равна $1/T$ бит/с. При *нормальной* работе ключ находится в положении "вверх", когда двоичный нуль, и в положении "вниз", когда двоичная единица. Предположим, что ключ *неисправен*. С вероятностью p он переключается в неверном направлении на T-секундный интервал. Наличие ошибки коммутации в течение каждого интервала не зависит от ошибки коммутации в любое другое время. Считайте, что $E\{z(t)\} = \pm\sqrt{E}$.

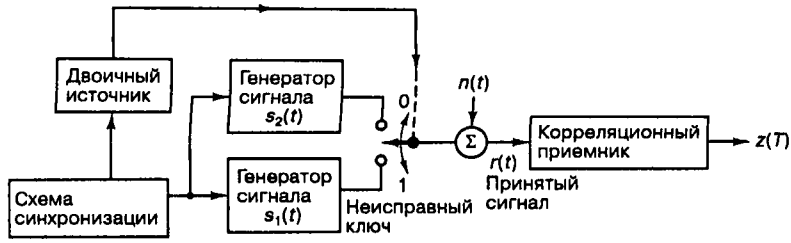


Рис. 34.2

а) Запишите условные вероятности $p(z|s_1)$ и $p(z|s_2)$.

б) Корреляционный приемник наблюдает сигнал $r(t)$ в течение интервала $(0, T)$. Нарисуйте блочную диаграмму оптимального приемника для минимизации вероятности битовой ошибки, если известно, что коммутатор сбоит с вероятностью p.

в) Какая система предпочтительнее

$$p = 0{,}1 \text{ и } E_b/N_0 = \infty$$

или

$$p = 0 \text{ и } E_b/N_0 = 7 \text{ дБ?}$$

4.12. а) Рассмотрим систему, использующую 16-ричную модуляцию PSK с вероятностью символьной ошибки $P_E = 10^{-5}$. При присвоении символам битового значения используется код Грея. Чему приблизительно равна вероятность битовой ошибки?

б) Повторите п. а для 16-ричной ортогональной модуляции FSK.

4.13. Рассмотрим систему ортогональной модуляции MFSK с $M = 8$; при равновероятных сигналах $s_i(t) = A\cos 2\pi f_i t$, $i = 1, \ldots, M$, $0 \le t \le T$, где $T = 0{,}2$ мс. Амплитуда несущей, A, равна 1 мВ, а двусторонняя спектральная плотность шума AWGN $N_0/2$ равна 10^{-11} Вт/Гц. Вычислите вероятность битовой ошибки, P_B.

4.14. Система со скоростью передачи данных 100 Кбит/с для передачи по каналу с шумом AWGN с использованием модуляции MPSK с когерентным обнаружением требует вероятности битовой ошибки $P_B = 10^{-3}$. Ширина полосы системы равна 50 кГц. Пусть частотная передаточная функция системы имеет вид приподнятого косинуса с коэффициентом сглаживания $r = 1$ и для присвоения символам битового значения используется код Грея.

а) Чему при заданной P_B равно отношение E_s/N_0?

б) Какое требуется отношение E_b/N_0?

4.15. Система, использующая дифференциальную модуляцию MPSK и когерентное обнаружение, работает в канале с шумом AWGN при $E_b/N_0 = 10$ дБ. Чему равна вероятность символьной ошибки при $M = 8$ и равновероятных символах?

4.16. Если основным критерием производительности системы является вероятность битовой ошибки, какую из следующих схем модуляции стоит выбрать для передачи по каналу с шумом AWGN?

Когерентная 8-ричная ортогональная FSK с $E_b/N_0 = 8$ дБ

или

Когерентная 8-ричная PSK с $E_b/N_0 = 13$ дБ

Приведите вычисления. (При присвоении символам битового значения предполагается использование кода Грея.)

4.17. Пусть демодулятор/детектор схемы с модуляцией BPSK содержит ошибку синхронизации, состоящую в смещении времени pT, где $0 \le p \le 1$. Другими словами, обнаружение символов начинается и завершается раньше (позже) на время pT. Предполагается равновероятная передача сигналов и идеальная частотная и фазовая синхронизация. Отметим, что эффект ухудшения достоверности вследствие известного фиксированного смещения можно вычислить, используя аналитические выражения, данные в тексте главы. В то же время, если ошибка фазы будет включать случайное смещение, вычисление его воздействия потребует стохастического рассмотрения (см. главу 10).

а) Выведите выражение для вероятности битовой ошибки P_b в зависимости от p.

б) Пусть в приемнике $E_b/N_0 = 9{,}6$ дБ и $p = 0{,}2$; вычислите ухудшение P_b в зависимости от смещения времени.

в) Если ошибку, описанную в данном примере, компенсировать не удается, насколько большее отношение E_b/N_0 понадобится для восстановления P_b, соответствующей $p = 0$?

4.18. Используя все приведенные условия, повторите задачу (4.17) для когерентного обнаружения потока битов в модуляции BFSK.

4.19. Пусть демодулятор/детектор схемы с модуляцией BPSK содержит ошибку синхронизации, состоящую в смещении времени pT, где $0 \le p \le 1$. Допустим также, что существует постоянная ошибка оценки фазы ϕ. Предполагается равновероятная передача сигналов и идеальная частотная синхронизация.

а) Выведите выражение для вероятности битовой ошибки P_b в зависимости от p и ϕ.

б) Пусть в приемнике $E_b/N_0 = 9{,}6$ дБ, $p = 0{,}2$ и $\phi = 25°$; вычислите ухудшение P_b в зависимости от смещения времени и фазы.

в) Если ошибки, описанные в данном примере, компенсировать не удается, насколько большее отношение E_b/N_0 понадобится для восстановления P_b, соответствующей $p = 0$ и $\phi = 0°$?

4.20. Чаще всего используемым методом синхронизации является корреляция с известной последовательностью Баркера, которая при надлежащей синхронизации дает яркий корреляционный пик, а при ее отсутствии — малый корреляционный выход. С помощью короткой последовательности Баркера 1 0 1 1 1 (первым является левый крайний бит) спроектируйте дискретный согласованный фильтр, подобный приведенному на рис. 4.10, который согласовывается с данной последовательностью. Докажите его пригодность, изобразив как функцию времени выход в зависимости от входа, на который подана последовательность 1 0 1 1 1.

Вопросы для самопроверки

4.1. В какой точке системы определяется отношение E_b/N_0 (см. раздел 4.3.2)?

4.2. Амплитудная или фазовая манипуляция представляется как совокупность точек или векторов на плоскости. Почему подобное представление нельзя использовать для ортогональной передачи сигналов, например сигналов FSK (см. раздел 4.4.4)?

4.3. Чему при передаче сигналов MFSK равно минимальное расстояние между тонами, обеспечивающее *ортогональность* сигналов (см. раздел 4.5.4)?

4.4. Какие преимущества при представлении синусоид дает *комплексная запись* (см. разделы 4.2.1 и 4.6)?

4.5. Схемы цифровой модуляции относятся к одному из двух классов с противоположными поведенческими характеристиками: схемы с *ортогональной* передачей сигналов и схемы с *модуляцией фазы/амплитуды*. Опишите поведение каждого класса (см. раздел 4.8.2).

4.6. Почему двоичная фазовая манипуляция (binary phase shift keying — BPSK) и четверичная фазовая манипуляция (quaternary phase shift keying — QPSK) имеют одинаковую вероятность битовой ошибки (см. раздел 4.8.4)?

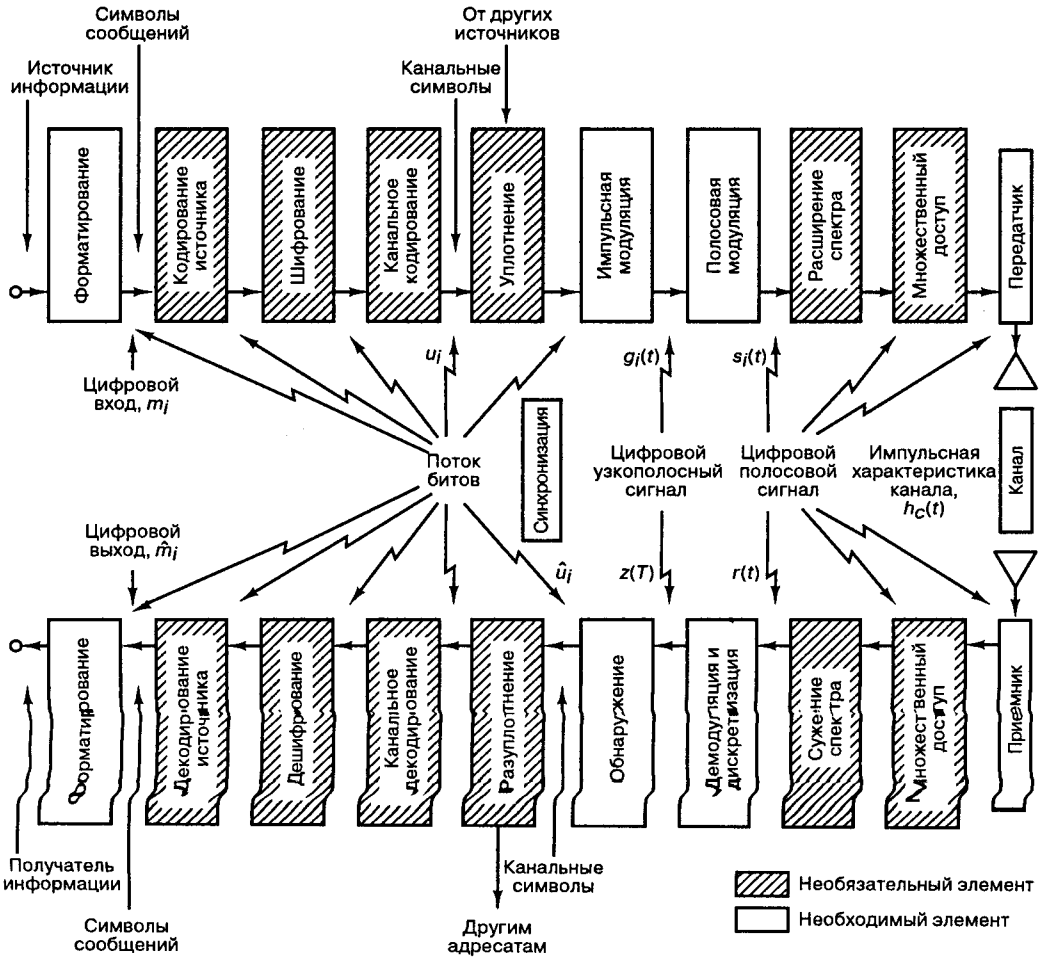
4.7. Почему при многофазной манипуляции (multiple-phase shift keying — MPSK) *эффективность использования полосы* повышается с увеличением размерности сигнального пространства (см. разделы 4.8.2 и 4.8.3)?

4.8. Почему при ортогональной передаче, например передаче сигналов MFSK, достоверность передачи повышается с увеличением размерности сигнального пространства (см. раздел 4.8.5)?

4.9. Применение кода Грея является одним из редких случаев в цифровой связи, где определенное преимущество может быть получено *безвозмездно*. Объясните, почему (см. раздел 4.9.4).

ГЛАВА 5

Анализ канала связи



5.1. Что такое бюджет канала связи

Когда говорим о *канале связи* (communication link), какую часть системы мы подразумеваем? Это физический канал или область между передатчиком и приемником? Нет, это нечто большее. Канал представляет собой тракт связи, который начинается с информационного источника, проходит через все этапы кодирования и модуляции, передатчик, физический канал, приемник (со всеми его этапами обработки) и завершается на получателе информации.

Что такое анализ канала связи? Какова его роль при разработке системы связи? Анализ канала связи и его результат, бюджет канала, состоят из вычисления и табулирования полезной мощности сигнала и паразитной мощности шума в приемнике. Бюджет канала — это расчет баланса потерь и прибыли; он определяет подробное соотношение между ресурсами передачи и приема, источниками шума, поглотителями сигнала и результатами процессов, выполняемых в канале. Некоторые параметры бюджета являются статистическими (например, скидка на замирание сигнала, которое описывается в главе 15). Бюджет — это метод *оценки*, позволяющий определить достоверность передачи системы связи. В главах 3 и 4 мы рассматривали графики зависимости вероятности ошибки от отношения E_b/N_0, имеющие "водопадоподобную" форму, подобную показанной на рис. 3.6. В этих главах для различных типов модуляции мы связали вероятность ошибки с отношением E_b/N_0 при гауссовом шуме. После того как выбрана схема модуляции, требования к определенной вероятности ошибки диктуют выбор операционной точки на кривой зависимости; другими словами, требуемая достоверность передачи определяет значение E_b/N_0, которое должно быть доступным в приемнике для получения этой достоверности. Основная задача анализа канала связи — это определить *действительную* рабочую точку системы на графике, изображенном на рис. 3.6, и установить, что вероятность ошибки, связанная с этой точкой, меньше (или равна) требуемой. Из множества спецификаций, анализов и табличных представлений, используемых для разработки системы связи, бюджет канала занимает особое место, поскольку обеспечивает обзор системы в целом.

Изучая бюджет канала, можно многое узнать об общей структуре и производительности системы. Например, из энергетического резерва канала связи можно узнать, как система удовлетворяет многочисленным требованиям — идеально, с натяжкой или вообще не удовлетворяет. Бюджет канала связи может показывать, существуют ли какие-либо аппаратные ограничения и можно ли их компенсировать за счет других частей канала. Вообще, бюджет канала часто используется для расчета компромиссов системы и изменения конфигурации; кроме того, он способствует пониманию различных аспектов и взаимозависимостей на уровне подсистем. Краткое изучение бюджета канала и сопровождающей его документации позволяет судить о том, был ли анализ выполнен точно или представляет грубую оценку. Вместе с другими методами моделирования бюджет канала помогает предсказать вес и размер оборудования, первоначальные энергетические требования, технические риски и стоимость системы. Бюджет канала — это один из самых важных документов управляющего системой; он представляет "итоговый отчет" по поиску оптимальной производительности системы.

5.2. Канал

Среда распространения, или электромагнитный тракт связи, соединяющий передатчик и приемник, называется *каналом* (channel). Вообще, каналы связи могут состоять из провод-

ников, коаксиальных и оптоволоконных кабелей, а также (в случае передачи в радиодиапазоне частот) волноводов, атмосферы или открытого пространства. Для большинства наземных каналов связи пространство канала проходит через атмосферу. Для спутниковых каналов связи канал, в основном, проходит через открытое пространство. Следует напомнить, что хотя некоторые атмосферные явления происходят на высоте до 100 км, основная часть атмосферы лежит все же ниже 20 км. Следовательно, на атмосферу приходится только небольшая часть (0,05%) общей длины (35 800 км) тракта связи. Большая часть предлагаемой главы представляет анализ канала связи в контексте подобной спутниковой связи. Вопросы наземных беспроводных каналов связи будут рассмотрены в главе 15.

5.2.1. Понятие открытого пространства

Понятие *открытого пространства* (free space) подразумевает канал, свободный от любых помех распространению в диапазоне радиочастот, таких как поглощение, отражение, преломление или дифракции. Если часть канала приходится на атмосферу, эта часть должна быть однородной и удовлетворять всем указанным условиям. Предполагается, что земля находится бесконечно далеко (или что ее коэффициент отражения пренебрежимо мал). Предполагается также, что энергия, передаваемая на радиочастотах, является функцией только расстояния от передатчика (и, как в оптике, подчиняется закону обратных квадратов). Каналы открытого пространства описывают идеальный тракт распространения радиочастот; на практике распространение через атмосферу и возле поверхности земли подвержено поглощению, отражению, дифракции и рассеиванию, что корректирует передачу в открытом пространстве. Атмосферное поглощение рассмотрено в последующих разделах. Отражение, дифракция и рассеивание, которые имеют важную роль в определении производительности наземной связи, рассмотрены в главе 15. Кроме того, всестороннее обсуждение этих вопросов представлено в работе [1].

5.2.2. Снижение достоверности передачи

В главе 3 было установлено, что существует две основные причины снижения достоверности передачи. Первая — это уменьшение отношения сигнал/шум. Вторая — это искажение сигнала, которое может быть вызвано межсимвольной интерференцией (intersymbol interference — ISI). В главах 3 и 15 рассматриваются определенные методы выравнивания, уменьшающие последствия ISI. В данной главе мы обсудим "бухгалтерию" усиления и рассеивания мощности сигнала. В бюджет канала мы не будем включать межсимвольную интерференцию, поскольку ее особенностью является то, что повышение мощности сигнала не всегда устраняет искажение, вызванное ISI (см. раздел 3.3.2.)

Для цифровой связи вероятность ошибки зависит от отношения E_b/N_0 в приемнике, определенного в формуле (3.30) следующим образом.

$$\frac{E_b}{N_0} = \frac{S}{N}\left(\frac{W}{R}\right)$$

Другими словами, E_b/N_0 — это мера нормированного отношения сигнал/шум (S/N или SNR). Если не оговорено противное, под SNR подразумевается отношение *средней* мощности сигнала к *средней* мощности шума. Сигналом может быть информационный сигнал, узкополосная волна или модулированная несущая. Уменьшение SNR может происходить двумя способами: (1) путем снижения желаемой мощности сигнала и (2) посредством повышения мощности шума или мощности сигналов, интерферирующих с полезным сигна-

лом. Эти механизмы будем называть, соответственно, *ослаблением* (или *потерями*) и *шумом* (или *интерференцией*). Ослабление происходит при поглощении, отклонении или отражении части сигнала при его прохождении к заданному приемнику; таким образом, часть переданной энергии не доходит до пункта назначения. Существует несколько источников электрических шумов и интерференции, возникающих вследствие различных механизмов, — тепловой шум, галактический шум, атмосферные помехи, помехи от коммутирующих элементов, перекрестные помехи и интерферирующие сигналы от других источников. При промышленном использовании термины *потеря* и *шум* часто не различаются, поскольку их эффект на систему одинаков.

5.2.3. Источники возникновения шумов и ослабления сигнала

На рис. 5.1 представлена блочная диаграмма спутникового канала связи с источниками возникновения шумов и ослабления сигнала. На данном рисунке механизмы ослабления (или потерь) сигнала показаны затененными, а источники шума — штрихованными прямоугольниками. Источники, ослабляющие сигнал и вносящие шум, представлены сетчатыми прямоугольниками. Ниже приводится перечень источников (21 наименование) ухудшения качества передачи, в котором описаны важнейшие "вкладчики" в ухудшение отношения SNR. Нумерация списка соответствует нумерации, приведенной на рис. 5.1

1. *Потери, связанные с ограничением полосы.* Все системы используют в передатчике фильтры для передачи энергии в ограниченной или выделенной полосе. Это позволяет исключить интерференцию с сигналами других каналов или пользователей, а также удовлетворить требования органов государственного регулирования. Подобная фильтрация уменьшает общее количество передаваемой энергии; результат — *ослабление* сигнала.
2. *Межсимвольная интерференция (intersymbol interference — ISI).* Как показывалось в главе 3, фильтрация в системе — передатчике, канале и приемнике — может привести к межсимвольной интерференции. Принятые импульсы перекрываются; хвост одного импульса "размывается" на соседние символьные интервалы, что мешает процессу обнаружения. Даже при отсутствии теплового шума, неидеальная фильтрация, ограничение полосы системы и замирание в каналах приводят к возникновению межсимвольной интерференции.
3. *Фазовый шум гетеродина.* При использовании в процессе смешения сигналов гетеродина, случайное смещение фазы добавляет к сигналу *фазовый шум.* При использовании в корреляционном приемнике опорного сигнала случайное смещение фазы может привести к уменьшению возможностей детектора, а следовательно, к *ослаблению* сигнала. В передатчике случайное смещение фазы может привести к размыванию полосы выходного сигнала, которая затем будет ограничена выходным фильтром, что приведет к ослаблению сигнала.
4. *Преобразование амплитудной модуляции в фазовую* (AM/PM conversion). Данное преобразование — это явление *фазового шума*, проявляющееся в нелинейных устройствах, таких как лампа бегущей волны (traveling-wave tube — TWT, ЛБВ). Флуктуации амплитуды сигнала (амплитудная модуляция) порождают колебания фазы, вносящие *фазовый шум* в сигналы, которые выделяются с помощью когерентного детектирования. Преобразование амплитудной модуляции в фазовую также может приводить к возникновению дополнительных боковых полос, что вызывает *ослабление* сигнала.

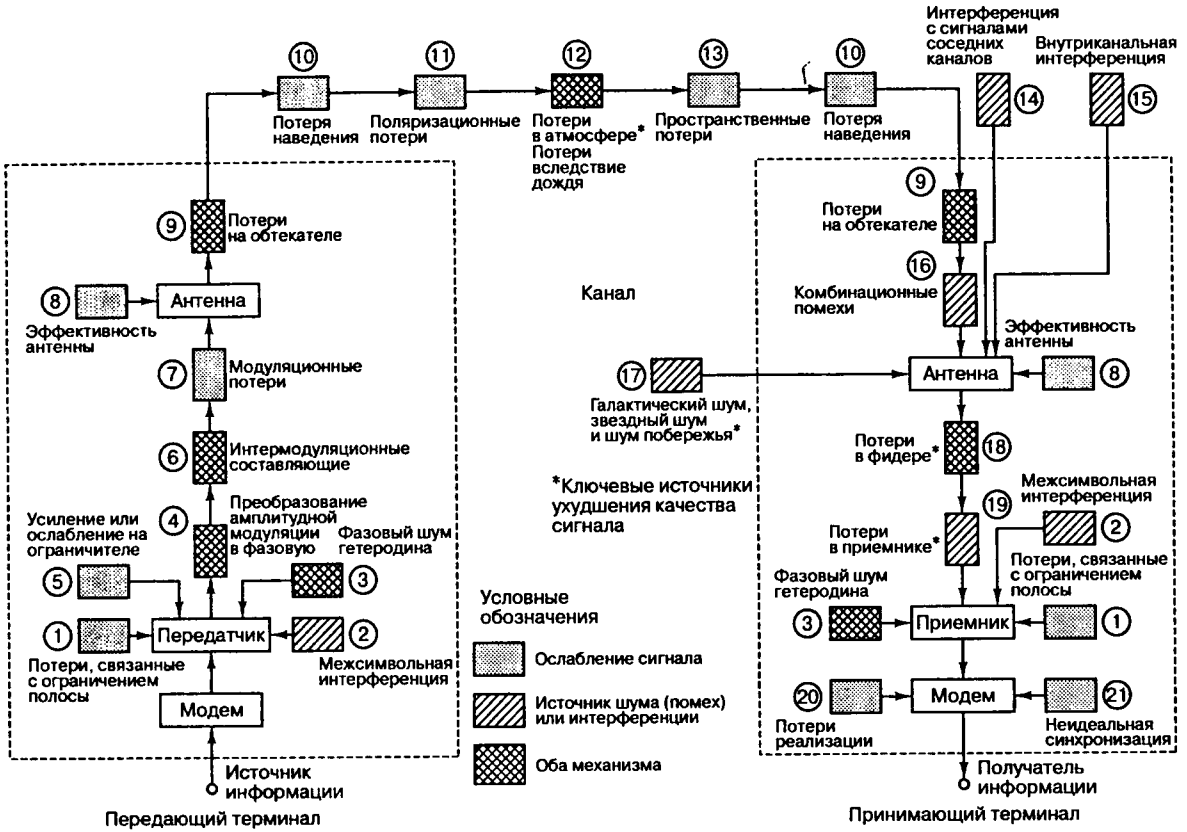


Рис. 5.1. Спутниковый канал связи "приемник-передатчик" с типичными источниками ослабления сигнала и помех

5. *Усиление или ослабление на ограничителе.* Ограничитель с резким порогом может усилить более мощный из двух сигналов и подавить более слабый; это может привести как к *усилению*, так и к *ослаблению* сигнала [2].
6. *Интермодуляционные (IM) составляющие, возникающие в результате взаимодействия нескольких несущих.* Когда несколько сигналов, которые передаются на разных несущих частотах, одновременно присутствуют в нелинейном устройстве, таком, например, как ЛБВ, может возникнуть мультипликативное взаимодействие между частотами несущих, что может привести к возникновению комбинационных сигналов суммарных и разностных частот. Перераспределение энергии между этими паразитными сигналами (интермодуляционные, или IM-составляющие) представляет потерю энергии сигнала. Кроме того, если эти IM-составляющие появляются в частотной области того или другого полезного сигнала, это приводит к увеличению уровня *шума* для соответствующего сигнала.
7. *Модуляционные потери.* Бюджет канала связи — это расчет принятой полезной мощности (или энергии). Полезной считается только та мощность, которая связана с сигналами, переносящими информацию. Достоверность передачи является функцией удельной энергии, приходящейся на один символ. Любая мощность, используемая для передачи несущей, отличной от той, что модулирует сигнал (символы), представляет *потери* модуляции. (Стоит, правда, отметить, что энергия несущей может использоваться для обеспечения синхронизации.)
8. *Эффективность антенны.* Антенны — это преобразователи, превращающие электронные сигналы в электромагнитные поля и наоборот. Кроме того, они используются для фокусировки электромагнитной энергии в заданном направлении. Чем больше апертура (поверхность) антенны, тем выше результирующая плотность мощности сигнала в заданном направлении. Эффективность антенны описывается отношением ее эффективной апертуры к физической. Механизмы, приводящие к снижению эффективности (*уменьшению интенсивности* сигнала), называются убыванием амплитуды, затенением апертуры, рассеиванием, переизлучением, приемом паразитных сигналов, дифракцией по краям и потерями вследствие диссипации [3]. Типичная эффективность, получаемая при суммарном воздействии всех названных механизмов, равна порядка 50–80%.
9. *Ослабление и шум на обтекателе.* Обтекатель — это специальная оболочка, применяемая для некоторых антенн в целях защиты от погодных воздействий. Обтекатель, находящийся на пути сигнала, будет рассеивать и поглощать некоторую энергию сигнала, что приведет к ослаблению *сигнала*. Основной закон физики утверждает, что тело, способное поглощать энергию, также излучает энергию (при температуре свыше 0 K). Часть этой энергии приходится на полосу приемника и вносит посторонний *шум*.
10. *Потеря наведения.* Если передающая либо принимающая антенна направлена неидеально, существует возможность *потери* сигнала.
11. *Поляризационные потери.* Поляризация электромагнитного поля определяется как направление в пространстве, вдоль которого лежат силовые линии поля, а поляризация антенны описывается поляризацией ее поля излучения. При неверном согласовании передающей и принимающей антенн сигнал может *ослабляться*.
12. *Атмосферные помехи и шум атмосферы.* Атмосфера отвечает за ослабление сигнала, а также вносит нежелательные помехи. Основная часть атмосферы лежит ниже высоты

20 км; но даже в пределах этого относительно короткого пути работают важные механизмы потерь и шумов. На рис. 5.2 приведены теоретические графики одностороннего поглощения по направлению к зениту. Зависимости приведены для нескольких высот (начиная с уровня моря — 0 км) для составляющих водяного пара с плотностью 7,5 г/м^3 возле земной поверхности. Величина *ослабления* сигнала вследствие поглощения кислородом (O_2) и водяными парами показана как функция несущей частоты. Локальные максимумы поглощения расположены в окрестности 22 ГГц (водяной пар), 60 и 120 ГГц (O_2). Также стоит отметить, что атмосфера вносит в канал энергию *шумов*. Как и в случае обтекателя, молекулы, поглощающие энергию, также излучают энергию. Молекулы кислорода и водяного пара излучают шум по всему спектру радиочастот. Часть этого шума, приходящаяся на полосу данной системы связи, ухудшает ее отношение сигнал/шум. Ливень является основной атмосферной причиной *ослабления* сигнала и основным фактором, вносящим *шум*. Чем он интенсивнее, тем большую энергию сигнала он поглотит. Кроме того, в дождливый день через луч антенны, направленный на приемник, проходит больше атмосферных шумов, чем в ясный день. Вообще, атмосферные помехи — это относительно обширная тема, и мы еще вернемся к ней в следующих разделах.

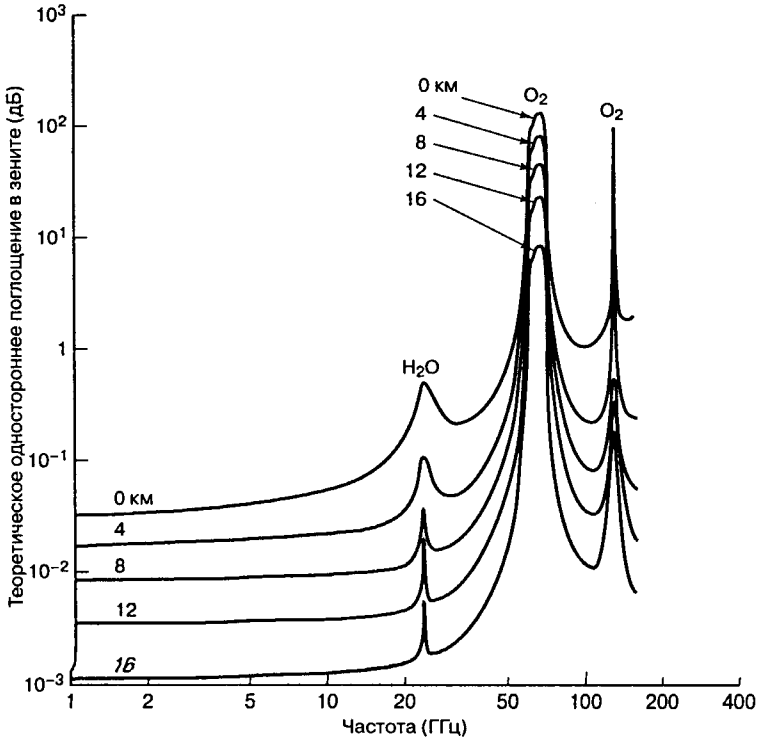


Рис. 5.2. Теоретическое вертикальное одностороннее поглощение от заданной высоты до верхней границы атмосферы для водяного пара плотностью 7,5 г/м^3 на поверхности. (Поглощение дождем или облаками не учитывается.) (Перепечатано с разрешения Национального комитета по аэронавтике и исследованию космического пространства из NASA Reference Publication 1082(03), "Propagation Effects Handbook for Satellite Systems Design", June, 1983, Fig. 6.2-1, p. 218.)

13. *Пространственные потери.* Интенсивность электрического поля, а следовательно, и интенсивность сигнала (плотности мощности или плотности потока мощности) уменьшаются с расстоянием. Для канала спутниковой связи пространственные потери — это наибольшие потери, вызванные одним фактором, приводящим к *ослаблению* в системе (данный фактор отнесен к *ослаблению* сигнала, потому что не вся излучаемая энергия фокусируется на целевой принимающей антенне).
14. *Помехи соседнего канала* (adjacent channel interference — ACI). Эта *интерференция* характеризуется нежелательными сигналами, которые поступают из других частотных каналов, или энергией, привносимой в интересующий нас канал. Возможность такого "заползания" соседнего сигнала определяется модуляционным спектральным сглаживанием, а также шириной и формой основного спектрального лепестка сигналов.
15. *Внутриканальная интерференция.* Данной *интерференцией* называется ухудшение качества, вызванное интерферирующими сигналами, которые появляются в пределах полосы частот сигнала. Она может вноситься по-разному, например, посредством случайных передач, недостаточного разграничения вертикальной и горизонтальной поляризации или приема паразитных сигналов боковым лепестком антенны (низкоэнергетическим лучом, окружающим основной луч антенны). Кроме того, внутриканальная интерференция может вноситься другими пользователями данного спектра.
16. *Комбинационные помехи.* Интермодуляционные составляющие, описанные в п. 6, происходят от сигналов с множественными несущими, взаимодействующими в нелинейном устройстве. Подобные составляющие иногда называются *активной взаимной модуляцией*; как говорилось в п. 6, они могут либо приводить к потере энергии сигнала, либо быть причиной внесения в канал шума. В данном пункте мы имеем дело с *пассивной взаимной модуляцией*; это явление вызывается взаимодействием сигналов с множественными несущими, имеющими нелинейные компоненты на выходе передатчика. Эти нелинейности обычно появляются на пересечении соединительных звеньев волноводов, корродированных поверхностях и поверхностях с плохим электрическим контактом. Электромагнитные поля значительной мощности, имеющие диодоподобную характеристику (рабочий потенциал), порождают мультипликативные составляющие, а следовательно, — помехи. Если подобные помехи будут излучаться на близлежащую принимающую антенну, они могут серьезно ухудшить качество функционирования приемника.
17. *Галактический или космический шум, звездный шум и шум побережья.* Все небесные тела, такие как звезды и планеты, излучают энергию. Подобная энергия шума, поступающая в зону обзора антенны, отрицательно сказывается на отношении сигнал/шум.
18. *Потери в фидере.* Уровень принятого сигнала может быть крайне мал (например, 10^{-12} В); следовательно, он может быть особенно чувствителен к воздействию шума. По этой причине в начале приемника находится область, где прилагаются значительные усилия, чтобы максимально снизить уровень шума, пока сигнал не будет в достаточной степени усилен. Волновод или кабель (фидер) между принимающей антенной и собственно приемником не только приводит к поглощению сигнала, но и вносит тепловой шум; подробно об этом рассказывается в разделе 5.5.3.
19. *(Собственный) шум приемника.* Это тепловой *шум*, порождаемый приемником; подробно этот вопрос рассмотрен в разделах 5.5.1–5.5.4.

20. *Потери аппаратной реализации.* Эти потери представляют собой разность между теоретической эффективностью обнаружения и реальной, определяемой несовершенством системы: ошибками синхронизации, уходом частоты, конечными временами нарастания и спада сигналов и конечнозначной арифметикой.
21. *Неидеальная синхронизация.* Если фаза несущей, фаза поднесущей и синхронизация символов организованы идеально, вероятность ошибки представляет собой однозначную функцию отношения E_b/N_0, рассмотренную в главах 3 и 4. К сожалению, названные величины реализуются не идеально, что приводит к *потерям.*

5.3. Мощность принятого сигнала и шума

5.3.1. Дистанционное уравнение

Основная задача бюджета канала — доказать, что система связи будет работать согласно плану; т.е. качество сообщений (достоверность передачи) будет удовлетворять заданным требованиям. Бюджет канала отслеживает "потери" и "прибыли" (усиление и ослабление) передаваемого сигнала от начала его формирования в передатчике до полного получения в приемнике. Вычисления показывают, чему равно отношение E_b/N_0 в приемнике и какой запас прочности существует. Процесс вычисления бюджета канала начинается с *дистанционного уравнения*, связывающего принятую мощность с расстоянием между передатчиком и приемником. Вывод этого уравнения дан ниже.

В системах радиосвязи несущая распространяется от передатчика посредством передающей антенны. Передающая антенна — это устройство, преобразовывающее электрические сигналы в электромагнитные поля. В приемнике принимающая антенна выполняет обратное преобразование; она превращает электромагнитные поля в электрические сигналы. Вывод уравнения, связывающего приемник и передатчик, обычно начинается с рассмотрения ненаправленного источника радиоизлучения, равномерно передающего в 4π стерадиан. На рис. 5.3 показан идеальный источник, называемый *изотропным излучателем* (isotropic radiator). Поскольку площадь поверхности сферы радиуса d равна $4\pi d^2$, плотность мощности $p(d)$ данной сферы с центром в источнике излучения связана с переданной мощностью P_t.

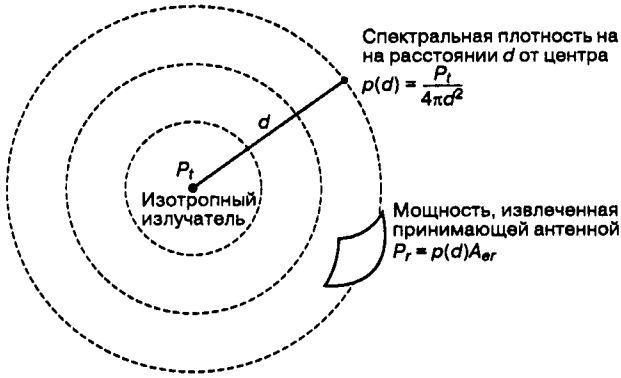


Рис. 5.3. Дистанционное уравнение. Выражение принятой мощности через расстояние

$$p(d) = \frac{P_t}{4\pi d^2} \text{ Вт/м}^2 \tag{5.1}$$

Мощность, извлеченную принимающей антенной, можно записать следующим образом.

$$P_r = p(d)A_{er} = \frac{P_t A_{er}}{4\pi d^2} \tag{5.2}$$

Здесь параметр A_{er} — это сечение захвата (эффективная площадь) принимающей антенны, определяемое следующим образом.

$$A_{er} = \frac{\text{общая извлеченная мощность}}{\text{плотность потока мощности падающего луча}} \tag{5.3}$$

Если рассматриваемая антенна является передающей, ее эффективная площадь обозначается как A_{et}. Если не указано, выполняет ли антенна принимающую или передающую функцию, эффективная площадь обозначается через A_e.

Эффективная площадь антенны A_e и ее физическая площадь поверхности A_p связаны коэффициентом эффективности η.

$$A_e = \eta A_p \tag{5.4}$$

Это говорит о том, что не вся мощность падающего луча была извлечена; вследствие различных механизмов [3] происходят потери. Номинальное значение η для параболической антенны составляет 0,55, а для рупорной — 0,75.

Определим параметр антенны, который связывает выходную (или входную) мощность с мощностью изотропного излучателя и именуется *коэффициентом направленного действия*.

$$G = \frac{\text{максимальная интенсивность мощности}}{\text{средняя интенсивность мощности в } 4\pi \text{ стерадиан}} \tag{5.5}$$

При отсутствии любых диссипативных потерь или потерь вследствие несогласованности импедансов коэффициент направленного действия антенны (в направлении максимальной интенсивности излучения) определяется из формулы (5.5). В то же время, если существует некоторая диссипация или несогласованность, коэффициент направленного действия антенны уменьшается на множитель, соответствующий объему потерь [4]. В данной главе мы будем предполагать, что диссипативные потери равны нулю, а импедансы согласованы идеально. Таким образом, формула (5.5) описывает *максимальный коэффициент направленного действия антенны*; как показано на рис. 5.4, его можно рассматривать как результат концентрации изотропного излучения в некоторой ограниченной области, меньшей 4π стерадиан. Теперь мы можем определить *эффективную излученную мощность* относительно изотропного излучателя (эффективная изотропно-излучаемая мощность — effective isotropic radiated power, EIRP) как произведение переданной мощности P_t и коэффициента усиления передающей антенны G_t.

$$\text{EIRP} = P_t G_t \tag{5.6}$$

Пример 5.1. Эффективная изотропно-излучаемая мощность

Покажите, что при надлежащем выборе антенн можно получить одинаковое значение EIRP как при использовании передатчика с $P_t = 100$ Вт, так и при использовании передатчика с $P_t = 0{,}1$ Вт.

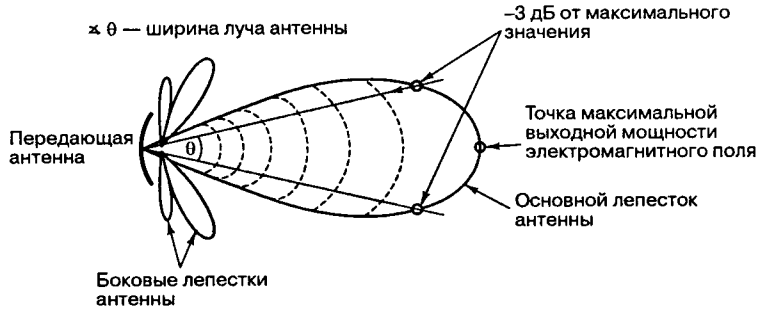


Рис. 5.4. Коэффициент направленного действия антенны — результат концентрации изотропного излучения

Решение

На рис. 5.5, *а* показан передатчик с $P_t = 100$ Вт, соединенный с изотропной антенной; значение $\text{EIRP} = P_t G_t = 100 \times 1 = 100$ Вт. На рис. 5.5, *б* показан передатчик с $P_t = 0{,}1$ Вт, соединенный с антенной, имеющей $G_t = 1000$; $\text{EIRP} = P_t G_t = 0{,}1 \times 1000 = 100$ Вт. Если измерители напряженности поля расположены так, как показано на рисунке, то измеряемая с их помощью эффективная мощность не будет отличаться.

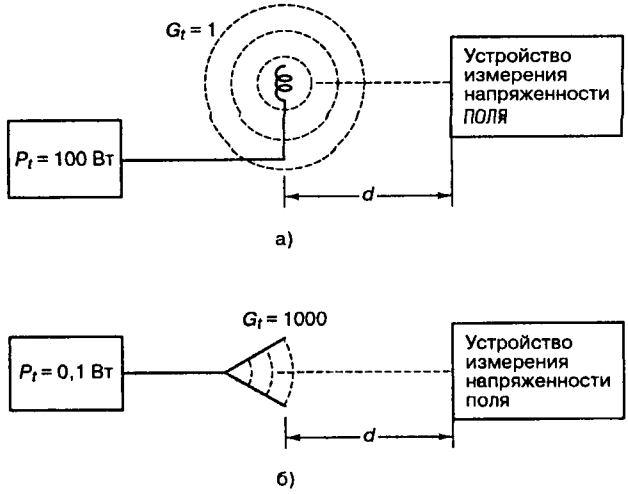


Рис. 5.5. Два различных способа получения одинакового значения EIRP

5.3.1.1. Возвращаясь к дистанционному уравнению

Если антенна передатчика имеет некоторый коэффициент направленного действия, отличающийся от предоставляемого изотропной антенной, в уравнении (5.2) мы меняем P_t на EIRP, что дает следующее.

$$P_r = \text{EIRP}\frac{A_{er}}{4\pi d^2} \tag{5.7}$$

Связь между коэффициентом усиления антенны G и эффективной площадью A_e дается выражением [4].

$$G = \frac{4\pi A_e}{\lambda^2} \quad (\text{для } A_e >> \lambda^2) \tag{5.8}$$

Здесь λ — длина волны несущей. Длина волны λ и частота f связаны соотношением $\lambda = c/f$, где c — скорость света ($\approx 3 \times 10^8$ м/с). *Теорема взаимности* утверждает, что для данной антенны при данной длине волны коэффициенты направленного действия приема и передачи идентичны [4].

Зона обзора антенны является мерой телесного угла, в котором сконцентрирована большая часть мощности поля. Зона обзора — это мера анизотропных свойств антенны; она обратно пропорциональна усилению антенны, т.е. антеннам с большим коэффициентом усиления соответствует более узкая зона обзора. Часто зону обзора выражают не через телесный угол, а через плоский *угол раствора антенны* (beamwidth), измеряемый в радианах или градусах. На рис. 5.4 показана диаграмма направленности антенны и дана иллюстрация общего определения угла раствора антенны. Угол раствора — это угол, образованный точками, в которых максимальная мощность поля ослаблена на 3 дБ. Как угол раствора зависит от частоты сигнала и размера антенны? Из уравнения (5.8) можно видеть, что усиление антенны увеличивается с уменьшением длины волны (увеличением частоты); также усиление антенны увеличивается с увеличением эффективной площади. Увеличение усиления антенны равносильно фокусировке плотности потока энергии в меньшем угле раствора; следовательно, увеличение частоты сигнала или размера антенны приводит к *сужению угла раствора*.

Эффективную площадь изотропной антенны можно вычислить, положив в уравнении (5.8) $G = 1$, что позволяет получить следующее выражение для A_e.

$$A_e = \frac{\lambda^2}{4\pi} \tag{5.9}$$

Затем для нахождения принятой мощности P_r, при изотропной принимающей антенне, подставляем уравнение (5.9) в уравнение (5.7), что дает следующее.

$$P_r = \frac{\text{EIRP}}{(4\pi d/\lambda)^2} = \frac{\text{EIRP}}{L_s} \tag{5.10}$$

Здесь совокупность коэффициентов $(4\pi d/\lambda)^2$ называется *потерями в тракте* (path loss) или *потерями в свободном пространстве* (free-space loss) и обозначается через L_s. Формула (5.10) показывает, что мощность, принятая изотропной антенной, равна эффективной переданной мощности, сниженной только за счет потерь в тракте связи. Если принимающая антенна не является изотропной, то после замены в уравнении (5.7) A_{er} выражением $G_r\lambda^2/4\pi$ из уравнения (5.8) получаем более общую формулу.

$$P_r = \frac{\text{EIRP}\, G_r \lambda^2}{(4\pi d)^2} = \frac{\text{EIRP}\, G_r}{L_s} \tag{5.11}$$

Здесь G_r — коэффициент усиления принимающей антенны. Полученное уравнение (5.11) называется *дистанционным*.

5.3.2. Мощность принятого сигнала как функция частоты

Поскольку и передающую, и принимающую антенны можно выразить через усиление или площадь, P_r можно выразить четырьмя различными способами.

$$P_r = \frac{P_t G_t A_{er}}{4\pi d^2} \tag{5.12}$$

$$P_r = \frac{P_t A_{et} A_{er}}{\lambda^2 d^2} \tag{5.13}$$

$$P_r = \frac{P_t A_{et} G_r}{4\pi d^2} \tag{5.14}$$

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi d)^2} \tag{5.15}$$

В этих выражениях A_{er} и A_{et} — эффективные площади принимающей и передающей антенн.

В уравнениях (5.12)–(5.15) зависимая переменная — это мощность принятого сигнала P_r, а независимые переменные — это такие параметры, как переданная мощность, коэффициент усиления антенны, площадь антенны, длина волны и расстояние между антеннами. Допустим, возник вопрос: как меняется принятая мощность при увеличении длины волны (или уменьшении частоты), при фиксированных остальных параметрах? Если рассматривать уравнения (5.12) и (5.14), то кажется, что P_r и длина волны вообще не связаны. Из уравнения (5.13) величина P_r вроде бы обратно пропорциональна квадрату длины волны, а из уравнения (5.15) она прямо пропорциональна квадрату длины волны. Нет ли здесь противоречия? Разумеется, нет; кажущаяся противоречивость уравнений (5.12)–(5.15) исчезает, если вернуться к формуле (5.8) и вспомнить, что коэффициент усиления антенны и ее площадь связаны через длину волны. Когда следует употреблять каждое из уравнений (5.12)–(5.15) для определения зависимости P_r от длины волны? Представим уже *сконструированную систему, т.е. антенны уже построены (зафиксированы A_{et} и A_{er}). В* этом случае подходящим выбором для вычисления P_r является уравнение (5.13), сформулированное для антенн фиксированного размера. Из этого уравнения видим, что принятая мощность увеличивается при уменьшении длины волны.

Рассмотрим уравнение (5.12), где независимыми переменными являются G_t и A_{er}. Итак, желательно, чтобы G_t и A_{er} были фиксированными при вычислении зависимости P_r от длины волны. Как изменится усиление при передаче на фиксированное расстояние, если уменьшить независимую переменную λ? G_t увеличится (см. уравнение (5.8)). Но мы не хотим увеличения G_t — оно нужно нам фиксированным. Другими словами, чтобы обеспечить неизменность G_t, нам необходимо уменьшать размер передающей антенны при уменьшении длины волны. Рассуждая подобным образом, приходим к выводу, что уравнение (5.12) удобно использовать при *фиксированном усилении передающей антенны* (или растворе антенны) и при переменном параметре A_{et}. Подобным образом уравнение (5.14) используется при фиксированных A_{et} и G_r, а уравнение (5.15) — при фиксированных коэффициентах усиления передающей и принимающей антенн (или растворах антенн).

На рис. 5.6 показано спутниковое приложение, где для обзора земной поверхности требуется луч со спутниковой антенны (раствор антенны равен порядка 17°). Поскольку коэффициент усиления спутниковой антенны G_t должен быть фиксированным, результирующая мощность P_r (см. уравнение (5.12)) не зависит от длины волны. Если передача ведется на определенной частоте f_1 ($= c/\lambda_1$), то изменение ее на f_2, где

$f_2 > f_1$, приведет к уменьшению обзора (поскольку при данной антенне увеличится G_t); таким образом, для поддержания требуемого обзора или раствора антенны размер этой антенны должен быть уменьшен. Итак, при увеличении несущей частоты антенны обзор земной поверхности уменьшается.

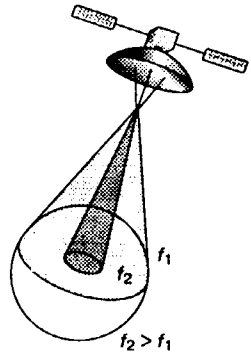


Рис. 5.6. Принятая мощность как функция частоты

5.3.3. Потери в тракте зависят от частоты

Из уравнения (5.10) можно видеть, что потери в тракте L_s зависят от длины волны (частоты). Довольно часто возникает вопрос: почему потери в тракте, подчиняющиеся простому геометрическому закону ослабления (ослабление обратно пропорционально квадрату расстояния), зависят от частоты? Ответ заключается в том, что потери в тракте, выраженные в уравнении (5.10), *определены* для изотропной принимающей антенны ($G_r = 1$). Вообще, потери в тракте — это весьма удобный параметр; он представляет гипотетическую потерю мощности, *которая произойдет, если принимающая антенна будет изотропной.* Из рис. 5.3 и уравнения (5.1) видно (из чисто геометрических соображений), что плотность мощности $p(d)$ — это функция расстояния, $p(d)$ *не* является функцией частоты. В то же время, поскольку потери в тракте заданы для $G_r = 1$, когда мы находим некоторую мощность P_r с помощью *изотропной антенны*, результат описывается выражением (5.10). Снова акцентируем внимание на том, что L_s можно рассматривать как совокупность параметров, которой было присвоено неудачное имя *потери в тракте.* Название представляет чисто геометрический эффект и не акцентирует внимания на том, что $G_r = 1$. Пожалуй, лучшим названием было бы *потери распространения при единичном усилении.* В системах радиосвязи потери в тракте — это наибольший единичный источник ослабления мощности сигнала. В спутниковых системах потери в тракте для канала связи со спутником в полосе С (6 ГГц) обычно составляют порядка 200 дБ.

Пример 5.2. Проект антенны для измерения потерь в тракте

Предложите эксперимент для измерения потерь в тракте L_s при частотах f_1 = 30 МГц и f_2 = 60 МГц, если расстояние между передатчиком и приемником равно 100 км. В обоих случаях найдите эффективную площадь принимающей антенны и вычислите потери в тракте в децибелах.

Решение

Два канала измерения L_s для частот f_1 и f_2 показаны на рис. 5.7. Для обоих приемников удельная мощность $p(d)$ одинакова и равна следующему.

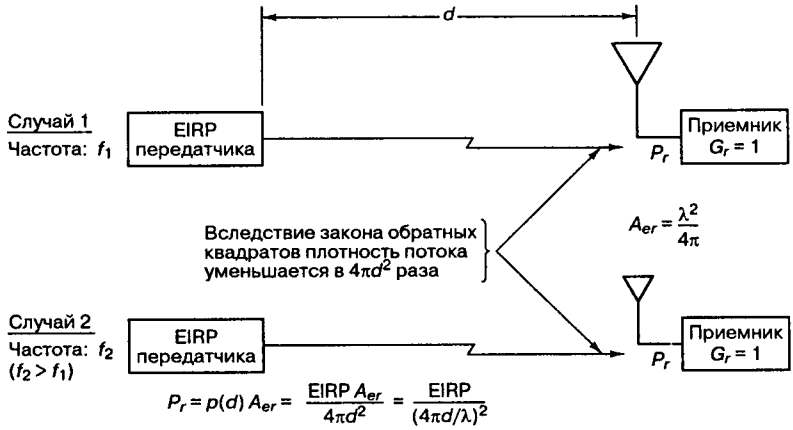


Рис. 5.7. Зависимость потерь в тракте от частоты. Предполагаемый эксперимент измерения потерь для двух различных частот

$$p(d) = \frac{\text{EIRP}}{4\pi d^2}$$

Это снижение удельной мощности происходит *исключительно* вследствие закона обратных квадратов. Действительная мощность, полученная каждым приемником, находится, как показано на рис. 5.7, посредством умножения плотности мощности $p(d)$ в приемнике на эффективную площадь собирающей антенны A_{er}. Поскольку потери в тракте определены для $G_r = 1$, эффективные площади A_{er1} и A_{er2} для частот f_1 и f_2 находятся с использованием уравнения (5.9).

$$A_{er} = \frac{\lambda^2}{4\pi} = \frac{(c/f)^2}{4\pi}$$

$$A_{er1} = \frac{(3\times 10^8/30\times 10^6)^2}{4\pi} \approx 8 \text{ м}^2$$

$$A_{er2} = \frac{(3\times 10^8/60\times 10^6)^2}{4\pi} \approx 2 \text{ м}^2$$

Далее для обоих случаев находим потери в тракте (в децибелах).

$$L_{s1} = 10\times \lg\left(\frac{4\pi d}{\lambda_1}\right)^2 = 10\times \lg\left(\frac{4\pi\times 10^5}{3\times 10^8/30\times 10^6}\right)^2 = 102 \text{ дБ}$$

$$L_{s2} = 10\times \lg\left(\frac{4\pi d}{\lambda_2}\right)^2 = 10\times \lg\left(\frac{4\pi\times 10^5}{3\times 10^8/60\times 10^6}\right)^2 = 108 \text{ дБ}$$

5.3.4. Мощность теплового шума

Тепловой шум вызывается тепловым движением электронов во всех проводящих элементах. Он создается в местах соединения антенны и приемника и в первых каскадах приемника. Спектральная плотность мощности шума постоянна для всех частот,

вплоть до 10^{12} Гц, что определило название *белый шум*. Как показывалось в разделе 1.5.5, процесс теплового шума в приемниках системы связи моделируется как процесс аддитивного белого гауссового шума (additive white Gaussian noise — AWGN). Физическая модель [5, 6] теплового шума — это генератор шума со среднеквадратическим напряжением холостого хода, равным $4\kappa T°W\Re$, где

κ (константа Больцмана) = $1{,}38 \times 10^{-23}$ Дж/К или Вт/КГц
$= -228{,}6$ дБВт/КГц,
$T°$ — температура, Кельвин
W — ширина полосы, Герц

и

$\Re$ — сопротивление, Ом

Максимальная мощность теплового шума N, которую можно подать с выхода генератора шума на вход усилителя, равна следующему.

$$N = \kappa T°W \text{ Ватт} \tag{5.16}$$

Следовательно, максимальная номинальная односторонняя спектральная плотность мощности шума N_0 (мощность шума на 1 Гц полосы) на выходе усилителя равна следующему.

$$N_0 = \frac{N}{W} = \kappa T° \text{ Ватт/Герц} \tag{5.17}$$

Может показаться, что мощность шума должна зависеть от значения сопротивления — но это не так. Рассмотрим такой аргумент. Соединим электрически большое и малое сопротивление так, чтобы они формировали замкнутую пару и их физические температуры были одинаковы. Если бы мощность шума зависела от сопротивления, то наблюдался бы поток полезной мощности от большего сопротивления к меньшему; большее сопротивление охлаждалось бы, а меньшее — нагревалось. Но это противоречит нашему жизненному опыту, не говоря уже о втором начале термодинамики. Следовательно, мощность, поступающая от большего сопротивления к меньшему, должна равняться мощности, получаемой этим большим сопротивлением.

Как видно из уравнения (5.16), мощность, подаваемая источником теплового шума, зависит от температуры окружающей среды источника (*шумовой температуры*). Это позволяет ввести для источников шума полезное понятие *эффективной шумовой температуры* (причем источники не обязательно должны быть тепловыми по природе — галактика, атмосфера, интерферирующие сигналы), влияющей на работу принимающей антенны. Эффективная шумовая температура подобного источника шума определяется как температура гипотетического источника теплового шума, дающего эквивалентную паразитную мощность. Подробнее шумовая температура рассматривается в разделе 5.5.

Пример 5.3. Максимальная номинальная мощность шума

Используя генератор со среднеквадратическим напряжением, равным $4\kappa T°W\Re$, покажите, что максимальная мощность шума, которую можно подать из такого источника на усилитель, равна $N_i = \kappa T°W$.

Решение

Теорема из области теории электрических цепей утверждает, что максимальная мощность подается на нагрузку, если полное сопротивление (импеданс) нагрузки равно комплексно сопряженному импедансу генератора [7]. В нашем случае импеданс генератора — это чистое сопро-

тивление, $\Re$; следовательно, условие передачи максимальной мощности удовлетворяется, если сопротивление усилителя равно $\Re$. Пример подобной схемы приведен на рис. 5.8. Источник теплового шума представлен электрически эквивалентной моделью, состоящей из бесшумного сопротивления, последовательно соединенного с идеальным генератором напряжения со среднеквадратическим напряжением $\sqrt{4\kappa T^\circ W\Re}$. Теперь входное сопротивление усилителя равно $\Re$. Напряжение шума, поступающего на вход усилителя, равно всего половине напряжения генератора, что следует из основных законов электрических схем. Таким образом, мощность шума, поданную на вход усилителя, можно выразить следующим образом.

$$N_i = \frac{\left(\sqrt{4\kappa T^\circ W\Re}/2\right)^2}{\Re} = \frac{4\kappa T^\circ W\Re}{4\Re} =$$

$$= \kappa T^\circ W$$

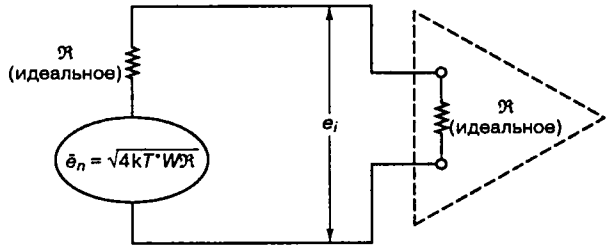


Рис. 5.8. Электрическая модель максимального теплового шума на входе усилителя

5.4. Анализ бюджета канала связи

При оценке производительности системы наибольший интерес представляет такой параметр, как отношение сигнал/шум (signal-to-noise ratio — SNR), или E_b/N_0. Причина — это основной фактор, определяющий возможность обнаружения сигналов при шуме с приемлемой вероятностью ошибки. Поскольку в спутниковых системах связи наиболее распространенной структурой сигнала является модулированная несущая с постоянной огибающей, в качестве интересующего нас отношения SNR мы можем использовать среднее *отношение мощности несущей к шуму* (carrier power-to-noise power) *C/N*. Фактически для передачи сигналов с постоянной огибающей данное додетекторное отношение SNR часто используется в форме одного из эквивалентных выражений.

$$\frac{P_r}{N} \equiv \frac{S}{N} \equiv \frac{C}{N} \equiv \frac{C}{\kappa T^\circ W}$$

Здесь P_r, S, C и N — принятая мощность, мощность сигнала, мощность несущей и мощность шума, а κ, T° и W — это константа Больцмана, температура в Кельвинах и ширина полосы. Действительно ли P_r/N или S/N — это всегда одно и то же, что и отношение несущей к шуму (C/N)? Нет, мощность сигнала и мощность несущей совпадают только при передаче сигналов с постоянной огибающей (угловой модуляции). Рассмотрим, например, частотно-модулированную (frequency modulated — FM) несущую, выраженную через модулирующий сигнал $m(t)$.

$$s(t) = A\cos\left(\omega_0 t + K\int m(t)\,dt\right)$$

Здесь K — константа системы. Средняя мощность в модулирующем сигнале равна $\overline{m^2(t)}$. Повышение этой модулирующей мощности приводит только к увеличению частотного отклонения $s(t)$; это означает, что несущая расширяется на больший спектр, но ее средняя мощность $\overline{s^2(t)}$ остается равной $A^2/2$, независимо от мощности модулирующего сигнала. Таким образом, частотная модуляция (FM), являющаяся примером передачи сигналов с постоянной огибающей, характеризуется тем, что мощность принятого сигнала равна мощности несущей.

Для линейной модуляции, такой как амплитудная модуляция (amplitude modulation — AM), мощность несущей несколько отличается от мощности модулирующего сигнала. Рассмотрим, например, выражение несущей через модулирующий сигнал $m(t)$.

$$s(t) = [1 + m(t)]\,A\cos\omega_0 t$$

$$\overline{s^2(t)} = [1+m(t)]^2\,\frac{A^2}{2} =$$

$$= \frac{A^2}{2}[1 + \overline{m^2(t)} + \overline{2m(t)}]$$

Если предположить, что среднее $m(t)$ равно нулю, то среднюю мощность несущей можно записать следующим образом.

$$\overline{s^2(t)} = \frac{A^2}{2} + \frac{A^2}{2}\overline{m^2(t)}$$

Из приведенного выше выражения видно, что при амплитудной модуляции мощность несущей отличается от мощности сигнала. Итак, параметры C/N и P_r/N совпадают при передаче сигналов с постоянной огибающей (например, при модуляциях PSK или FSK) и отличаются в остальных случаях (например, при модуляциях ASK, QAM).

Выражение для P_r/N можно получить, разделив обе части уравнения (5.11) на мощность шума N.

$$\frac{P_r}{N} = \frac{\text{EIRP}\,G_r/N}{L_s} \tag{5.18}$$

Формула (5.18) применима к любому одностороннему радиочастотному каналу. При использовании *аналоговых приемников* ширина полосы шума (обычно называемая эффективной или эквивалентной полосой шума), видимая демодулятором, обычно превышает ширину полосы сигнала, и отношение P_r/N — это основной параметр при определении возможности обнаружения сигнала и качества работы системы связи. При *цифровых приемниках* обычно реализуются корреляторы или согласованные фильтры, и ширина полосы сигнала обычно принимается равной ширине полосы шума. Как правило, мощность шума на входе не рассматривают, а обычной формулировкой отношения SNR для цифровых каналов связи является замещение мощности шума *спектральной плотностью мощности шума*. С помощью формулы (5.17) выражение (5.18) можно переписать следующим образом.

$$\frac{P_r}{N_0} = \frac{\text{EIRP}\,G_r/T^\circ}{kL_s L_0} \tag{5.19}$$

Здесь эффективная шумовая температура системы $T°$ — это функция шума, излучаемого на антенну, и теплового шума, генерируемого на первых каскадах приемника. Отметим, что коэффициент усиления принимающей антенны G_r и системную температуру $T°$ можно объединить в один параметр $G_r/T°$, иногда именуемый *добротностью приемника* (receiver figure-of-merit). Причина такой трактовки этих членов раскрывается в разделе 5.6.2.

Следует обратить внимание на то, что эффективная температура $T°$ — это параметр, *моделирующий* результат воздействия различных источников шума; подробнее этот вопрос рассмотрен в разделе 5.5. В формуле (5.19) был введен множитель L_o, описывающий все факторы ослабления и ухудшения, которые не учтены остальными членами уравнения (5.18). Множитель L_o включает большой набор различных источников ослабления и ухудшения, перечисленных ранее. Итак, в уравнении (5.19) связываются ключевые параметры любого анализа канала связи: отношение спектральной плотности мощности принятого сигнала к шуму (P_r/N_0), эффективная переданная мощность (EIRP), добротность приемника ($G_r/T°$) и потери (L_s, L_o). В настоящее время мы пытаемся развить методологический подход к отслеживанию потерь и прибылей в канале связи. Имея вначале некоторый ресурс мощности, мы с помощью формулы (5.19) можем вычислить суммарное отношение сигнал/шум, имеющее место на "лицевой стороне" детектора (додетекторной точке). Нашей целью является система "бухучета" (весьма подобная используемой в коммерции), бронирующая активы и пассивы и подводящая итог в виде чистого дохода (или потери). Формула (5.19) имеет как раз подобный, нужный нам предпринимательско-коммерческий вид. Все параметры (эффективная излученная мощность, добротность приемника), входящие в числитель, подобны коммерческим активам, а все параметры, фигурирующие в знаменателе, — пассивам.

Итак, предполагая, что вся принятая мощность P_r находится в модулирующем (переносящем информацию) сигнале, мы можем связать E_b/N_0 и SNR из уравнения (3.30) и записать следующее.

$$\frac{E_b}{N_0} = \frac{P_r}{N}\left(\frac{W}{R}\right) \tag{5.20,а}$$

$$\frac{E_b}{N_0} = \frac{P_r}{N_0}\left(\frac{1}{R}\right) \tag{5.20,б}$$

и

$$\frac{P_r}{N_0} = \frac{E_b}{N_0}R \tag{5.20,в}$$

Здесь R — скорость передачи битов. Если часть принятой мощности — это мощность несущей (т.е. имеем потерю мощности сигнала), мы по-прежнему можем использовать уравнение (5.20), за исключением того, что мощность несущей дает вклад в множитель потерь L_o в формуле (5.19). Полученная в уравнении (5.20) фундаментальная связь между E_b/N_0 и P_r/N_0 весьма пригодится нам в дальнейшем при проектировании и оценке систем (см. главу 9).

5.4.1. Два важных значения E_b/N_0

E_b/N_0 — это (согласно принятым обозначениям) отношение энергии бита к спектральной плотности мощности шума, необходимое для получения заданной вероятности ошибки. Для облегчения вычисления пределов рабочего диапазона или запаса прочности M необ-

ходимо различать *требуемое* отношение E_b/N_0 и реальное (или *принятое*) отношение E_b/N_0. С этого момента первое мы будем обозначать как $(E_b/N_0)_{\text{треб}}$, а последнее — $(E_b/N_0)_{\text{прин}}$. Иллюстрация приведена на рис. 5.9, где на графике обозначены две рабочие точки. Первая связана с $P_B = 10^{-3}$; далее будем называть эту рабочую точку требуемой системной достоверностью передачи. Предположим, что заданная достоверность получается при $(E_b/N_0)_{\text{треб}}$, равном 10 дБ. Вы думаете, что наша задача — создать систему, демодулятор которой получит *точно* эти 10 дБ? Разумеется, нет; мы определим и спроектируем систему с запасом прочности, так что реально принятое $(E_b/N_0)_{\text{прин}}$ будет несколько больше $(E_b/N_0)_{\text{треб}}$.

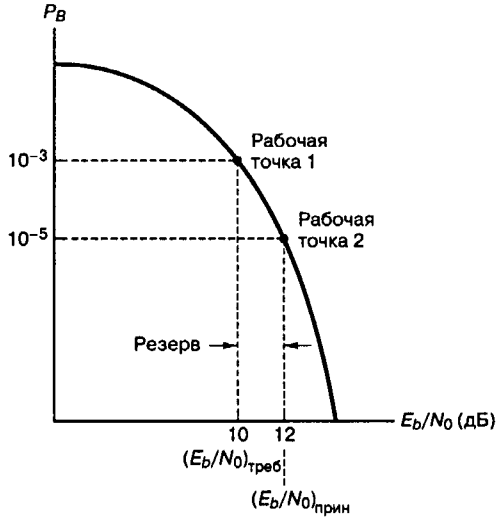


Рис. 5.9. Два важных значения E_b/N_0

Таким образом, мы должны разработать систему, которая бы работала на второй рабочей точке, показанной на рис. 5.9; в нашем случае $(E_b/N_0)_{\text{прин}} = 12$ дБ и $P_B = 10^{-5}$. Для данного примера мы можем описать запас прочности, или *энергетический резерв линии связи* (link margin), как дающий улучшение P_B на два порядка или (более привычная формулировка) энергетический запас линии связи можно описать как обеспечивающий на 2 дБ большее отношение E_b/N_0, чем требуется. Перепишем выражение (5.20), введя параметр энергетического резерва линии связи M.

$$\frac{P_r}{N_0} = \left(\frac{E_b}{N_0}\right)_{\text{прин}} R = N\left(\frac{E_b}{N_0}\right)_{\text{треб}} R \tag{5.21}$$

Разность в децибелах между $(E_b/N_0)_{\text{прин}}$ и $(E_b/N_0)_{\text{треб}}$ дает энергетический резерв линии связи.

$$M(\text{дБ}) = \left(\frac{E_b}{N_0}\right)_{\text{прин}} (\text{дБ}) - \left(\frac{E_b}{N_0}\right)_{\text{треб}} (\text{дБ}) \tag{5.22}$$

Параметр $(E_b/N_0)_{\text{треб}}$ отражает различия в структурах систем; эти различия могут быть вызваны отличиями схем модуляции или кодирования. Большее, чем ожидалось, отношение $(E_b/N_0)_{\text{треб}}$ может объясняться субоптимальной системой передачи в радиочас-

тотном диапазоне, дающей значительные ошибки синхронизации или допускающей больший шум в процессе обнаружения, чем идеальный согласованный фильтр.

Объединяя уравнения (5.19) и (5.21) и выражая энергетический резерв линии связи M, получаем следующее.

$$M = \frac{\text{EIRP}\, G_r / T^\circ}{(E_b / N_0)_{\text{треб}}\, R \kappa L_s L_0} \tag{5.23}$$

Уравнение (5.23), *выражение энергетического резерва линии связи*, содержит все параметры, влияющие на достоверность передачи по каналу связи. Некоторые из этих параметров определяются относительно конкретных точек системы. Например, отношение E_b/N_0 определяется на входе приемника. Если говорить более точно, то на входе детектора (додетекторной точке), где амплитуда напряжения демодулируемого сигнала пропорциональна принятой энергии, составляющей основу процесса принятия решения относительно значения принятого символа. Подобным образом любой параметр, описывающий принятую энергию или мощность, полезную или паразитную, также определяется относительно этой додетекторной точки. Добротность приемника G_r/T° определяется на входе принимающей антенны, где G_r — усиление принимающей антенны, а T° — эффективная температура системы (см. раздел 5.5.5). Эффективная мощность излучения EIRP — это мощность, связанная с электромагнитной волной на выходе передающей антенны. Итак, всегда нужно помнить, что каждый из параметров E_b/N_0, G_r/T° и EIRP вычисляется в определенной точке системы и никак иначе.

5.4.2. Бюджет канала обычно вычисляется в децибелах

Поскольку бюджет канала обычно вычисляется в децибелах, уравнение (5.23) можно переписать следующим образом.

$$M(\text{дБ}) = \text{EIRP}(\text{дБВт}) + G_r(\text{дБ[i]}) - \left(\frac{E_b}{N_0}\right)_{\text{треб}}(\text{дБ}) - R(\text{дББит/с}) - \tag{5.24}$$
$$- \kappa T^\circ(\text{дБВт/Гц}) - L_s(\text{дБ}) - L_o(\text{дБ})$$

Мощность переданного сигнала EIRP выражается в децибел-ваттах (дБВт); спектральная плотность мощности шума N_0 — в децибел-ваттах на герц (дБВт/Гц); усиление антенны G_r — в децибелах относительно изотропного усиления (дБ[i]); скорость передачи данных R — в децибелах относительно величины 1 бит/с (дББит/с); все остальные члены выражаются в децибелах (дБ). Численные значения параметров, фигурирующих в уравнении (5.24), составляют бюджет канала связи, полезное средство распределения ресурсов связи. Для поддержания положительного баланса мы должны найти приемлемое соотношение между всеми параметрами; мы можем снизить мощность передатчика путем предоставления избыточного резерва или увеличить скорость передачи данных путем снижения $(E_b/N_0)_{\text{треб}}$ (посредством выбора лучших схем модуляции и кодирования). Любой децибел в уравнении (5.24), независимо от параметра, не лучше и не хуже любого другого децибела — децибел есть децибел. Система передачи "не знает и знать не хочет", откуда приходят децибелы. Пока в приемнике обеспечивается надлежащее отношение E_b/N_0, система имеет необходимую достоверность передачи. Впрочем, введем еще два условия, которые необходимо будет удовлетворить при получении заданной вероятности ошибки, — должна поддерживаться синхрони-

зация и должно минимизироваться или компенсироваться искажение, вызванное межсимвольной интерференцией. Может возникнуть вопрос: если система не отдает предпочтения источнику поступления децибелов в отношение E_b/N_0, то как мы должны распределять приоритеты поиска достаточного числа децибелов. Ответ таков: мы должны искать наиболее рентабельные децибелы. Это и будет путеводной нитью нескольких следующих глав, посвященных кодам коррекции ошибок, поскольку именно для этой области характерно историческое развитие в направлении снижения стоимости оборудования, позволяющего получить более достоверную передачу.

5.4.3. Какой нужен резерв

Вопрос о величине энергетического запаса, встроенного в систему, возникает довольно часто. Ответ на него заключается в следующем. Если строго описать (учесть наиболее неблагоприятные варианты) все источники усилений и ослаблений сигнала и шума и считать дисперсию параметров канала (например, вследствие погодных условий) максимальной из возможных, то потребуется незначительная дополнительная надбавка энергетического запаса. Требуемый запас прочности зависит от степени достоверности каждой позиции бюджета канала. Для системы, в которой задействованы новые технологии или новые рабочие частоты, потребуется больший запас, чем для системы, которая создавалась и тестировалась уже неоднократно. Иногда в бюджете канала связи как отдельная позиция фигурирует затухание вследствие погодных условий. В других случаях требуемое значение энергетического запаса отражает требования канала при данном ухудшении параметров вследствие дождя. Для спутниковой связи на полосе частот С (линия связи "земля-спутник" использует частоту 6 ГГц, линия связи "спутник-земля" — частоту 4 ГГц), где все параметры хорошо известны и ведут себя довольно хорошо, систему можно проектировать всего лишь с 1 дБ энергетического запаса. Настроенные только на прием телевизионные станции, которые используют параболические антенны диаметром 16 футов и работают в полосе частот С, часто проектируются с энергетическим запасом, составляющим всего доли децибела. В то же время телефонная связь через спутник, которая использует стандарт 99,9% доступности канала, требует значительно большего энергетического запаса; в некоторых системах INTELSAT резерв составляет порядка 4–5 дБ. Если вычисления выполняются не для самого неблагоприятного варианта, а для фактически имеющегося, расчет обычно производится для совместимых дисперсий оборудования в рабочем диапазоне температур, перепадов напряжения в линии и длительностей передач. Кроме того, для спутниковой связи могут приниматься предположения о возможных ошибках отслеживания местонахождения спутника.

Проекты с использованием высоких частот (например, 14/12 ГГц) обычно требуют значительных (погодных) энергетических запасов, поскольку атмосферные потери крайне разнообразны и их влияние увеличивается с частотой. Следует отметить, что побочные продукты поглощения вследствие атмосферных потерь больше шума антенны. При использовании малошумящих усилителей даже небольшие погодные изменения могут привести к увеличению температуры антенны на 40–50 К. В табл. 5.1 показан анализ канала связи для спутника непосредственного вещания, предложенный Федеральной комиссии по средствам связи (Federal Communications Commission — FCC) США корпорацией Satellite Television. Отметим, что бюджет для линии связи "спутник-земля" рассчитан для двух альтернативных погодных условий: ясной погоды и ослабления на 5 дБ вследствие дождя. Ослабление сигнала из-за атмосферного поглощения составляет толь-

ко малую долю децибела при ясной погоде и 5 дБ — при дожде. Следующий пункт в таблице для линии связи "спутник-земля", $G/T°$ домашнего приемника, показывает дополнительное ухудшение качества, вызванное дождем; принимающая антенна излучает дополнительный тепловой шум, что приводит к увеличению эффективной шумовой температуры системы $T°$ и уменьшению $G/T°$ домашнего приемника (от 9,4 дБ/К до 8,1 дБ/К). Следовательно, при выделении дополнительного энергетического запаса на потери вследствие погодных условий, одновременно следует выделять дополнительный резерв для компенсации увеличения шумовой температуры системы.

Таблица 5.1. Спутник непосредственного вещания (Direct Broadcast Satellite — DBS), предложенный Satellite Television Corp.

Линия связи "земля-спутник"		
EIRP наземной станции	86,6 дБВт	
Потери в свободном пространстве (17,6 ГГц, угол возвышения 48°)	208,9 дБВт	
Предполагаемое поглощение вследствие дождя	12,0 дБВт	
$G/T°$ спутника	7,7 дБ/К	
$C/\kappa T°$ линии связи "земля-спутник"	102,0 дБГц	
	Атмосферные условия	
Линия связи "спутник-земля "	Ясно	Поглощение 5 дБ вследствие дождя
EIRP спутника	57,0 дБВт	57,0 дБВт
Потери в свободном пространстве (12,5 ГГц, угол возвышения 30°)	206,1 дБ	206,1 дБ
Поглощение в атмосфере	0,14 дБ	5,0 дБ
$G/T°$ домашнего приемника (параболическая антенна 0,75 м)	9,4 дБ/К	8,1 дБ/К
Потеря наведения приемника (ошибка 0,5°)	0,6 дБ	0,6 дБ
Рассогласование по поляризации (среднее)	0,04 дБ	0,04 дБ
$C/\kappa T°$ линии связи "спутник-земля"	88,1 дБГц	82,0 дБГц
Общее $C/\kappa T°$	87,9 дБГц	82,0 дБГц
Общее C/N (на 16 МГц)	15,9 дБ	10,0 дБ
Эталонное пороговое C/N	10,0 дБ	10,0 дБ
Резерв относительно порога	5,9 дБ	0,0 дБ

Небольшое замечание относительно спутниковых каналов связи: в промышленности часто встречаются выражения типа "канал *может* быть закрыт", т.е. значение энергетического запаса в децибелах положительно и удовлетворяются существующие требования к достоверности передачи, или "канал *не может* быть закрыт" — значение энергетического запаса отрицательно и существующие требования к достоверности передачи *не* будут удовлетворяться. Хотя при использовании выражений "канал закрывается" или "канал не закрыт" создается впечатление работы по принципу "включено/выключено", на самом деле

незакрытый канал (или отрицательный энергетический запас) означает, что достоверность передачи не удовлетворяет системным требованиям; это не обязательно означает прекращение связи. Рассмотрим, например, систему, показанную на рис. 5.9, с $(E_b/N_0)_{\text{треб}} = 10$ дБ и $(E_b/N_0)_{\text{прин}} = 8$ дБ. Пусть 8 дБ соответствует $P_B = 10^{-2}$. Следовательно, энергетический запас равен –1 дБ, а фактическая вероятность появления ошибочного бита в 10 раз превышает заданную. В то же время, несмотря на сниженную достоверность передачи, канал по-прежнему может использоваться.

5.4.4. Доступность канала

Доступность канала обычно является мерой долговременного использования канала, сформулированной на среднегодовой основе; для данного географического местоположения доступность канала показывает процентное отношение времени, в течение которого канал может быть закрыт. Например, для конкретного канала связи между Вашингтоном и спутниковым ретранслятором долговременная синоптическая ситуация может быть такой, что погодного запаса 10 дБ достаточно для закрытия канала связи 98% времени; для 2% времени проливные дожди приводят к большему, чем на 10 дБ, ухудшению параметра SNR, так что канал не закрывается. Поскольку воздействие шума на SNR зависит от частоты сигнала, доступность канала и требуемый энергетический запас должны изучаться в контексте конкретной частоты передачи.

На рис. 5.10 обобщаются значения доступности каналов глобальных спутников на частоте 44 ГГц. Данный график иллюстрирует процентное отношение видимости земной поверхности (каналы закрыты и заданная вероятность ошибки достигается) как функцию энергетического запаса для трех равномерно размещенных геостационарных спутников. *Геостационарный спутник* расположен на круговой орбите в той же плоскости, что и земная экваториальная плоскость, и его синхронная высота над уровнем моря равна 35 800 км. Период обращения спутника равен периоду обращения Земли; таким образом, спутник стационарно висит над определенной точкой земной поверхности. На рис. 5.10 показано семейство кривых видимости, отличающихся требуемыми значениями параметра доступности канала, от качественного (доступность 95%) до достаточно точного (99%). Вообще, при фиксированном энергетическом запасе видимость обратно пропорциональна требуемой доступности, а при фиксированной доступности она монотонно растет с увеличением запаса [8]. На рис. 5.11–5.13 для трех различных значений энергетического запаса канала затененными и чистыми областями показаны части земной поверхности, в которых канал 44 ГГц не может быть закрыт 99% времени. На рис. 5.11 показан охват каналом различных мест при энергетическом запасе 14 дБ. Отметим, что с помощью рисунка можно вычислить области наибольших ливней, такие как Бразилия и Индонезия. На рисунке представлены результаты расчета канала, выполненного с использованием синоптической модели Земли.

На рис. 5.11 выделяются заштрихованные полоски на восточных и западных границах поля зрения каждого спутника. Как вы думаете, почему канал недоступен в данных областях? На краях земной поверхности, видимой со спутника, расстояние между спутником и наземной станцией больше расстояния между точкой, находящейся непосредственно под спутником, и спутником. Ухудшение качества происходит вследствие сочетания трех элементов: (1) большее расстояние распространения приводит к уменьшению спектральной плотности мощности на при-

нимающей антенне; (2) в местах, расположенных на границе охвата, усиление, получаемое с помощью спутниковой антенны, снижается, если антенна специально не спроектирована для равномерного охвата всего поля зрения (обычная схема — это –3 дБ на крайних лучах по сравнению с пиковой амплитудой в центре луча); и (3) при распространении к точкам на границе охвата сигналу приходится пройти больший путь через атмосферы (это объясняется наклонным путем и кривизной земной поверхности). Последнее является самым важным для сигналов на частотах, наиболее поглощаемых атмосферой. Почему подобные заштрихованные области отсутствуют около северного и южного полюсов на рис. 5.11? Снегопад не имеет (на распространение сигнала) такого же отрицательного эффекта, как ливень; данный феномен называется *эффект замораживания*.

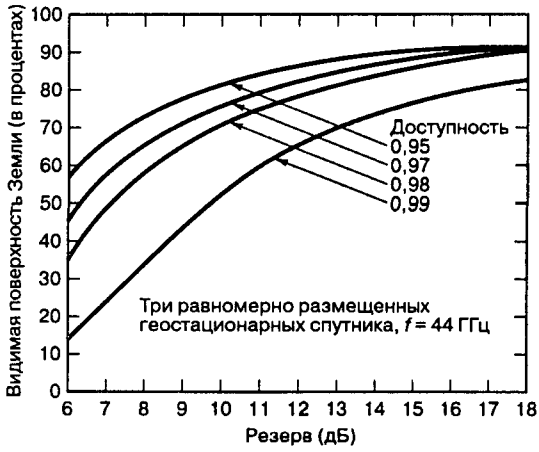


Рис. 5.10. Зависимость охвата земной поверхности от энергетического запаса линии связи при различных значениях доступности канала. (Перепечатано с разрешения Lincoln Laboratory из L. M. Schwab. "World-Wide Link Availability for Geostationary and Critically Inclined Orbits Including Rain Effects", Lincoln Laboratory, Rep. DCA-9, Jan., 27, 1981, Fig. 14, p. 38)

На рис. 5.12 показаны части земной поверхности, которые 99% времени могут (и не могут) закрывать канал 44 ГГц с запасом 10 дБ. Отметим, что, по сравнению с запасом 14 дБ, затененные области стали значительно больше; теперь восточный берег Соединенных Штатов, Средиземноморье и большая часть Японии 99% времени не могут закрывать канал. На рис. 5.13 подобные рабочие характеристики канала показаны для энергетического запаса 6 дБ. Если на рис. 5.11 можно определить регионы наибольшей дождливости, то на рис. 5.13 видны наиболее засушливые регионы Земли. Видим, что подобными областями являются юго-западные части Соединенных штатов, большая часть Австралии, побережья Перу и Чили, а также пустыня Сахара в Африке.

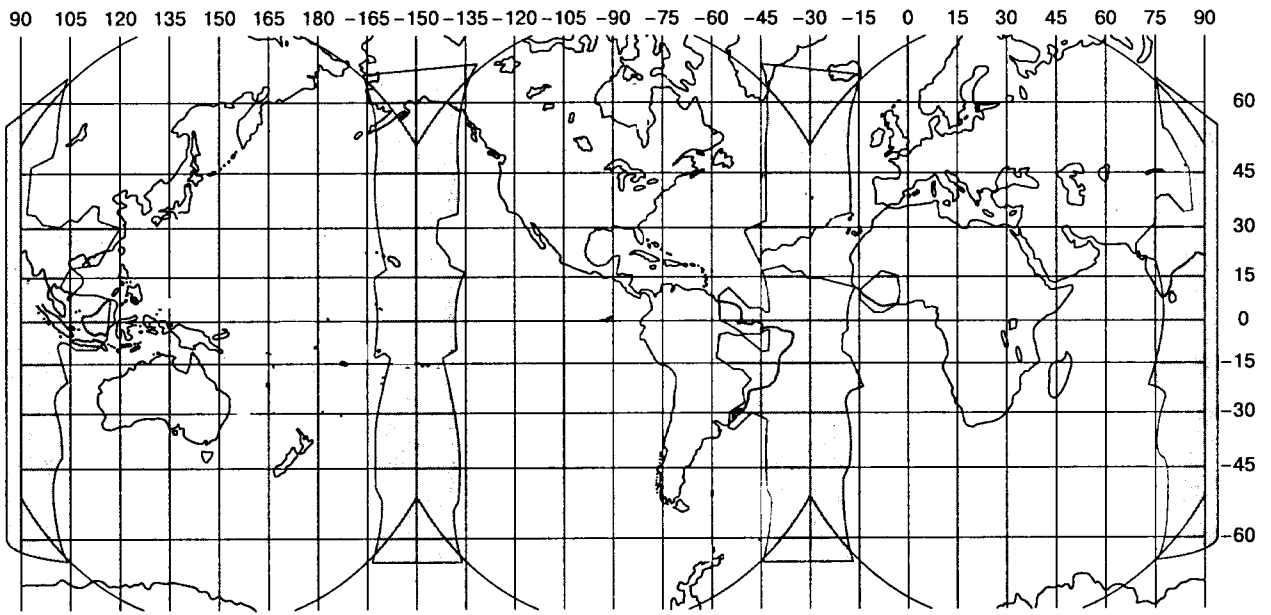


Рис. 5.11. Зависимость охвата земной поверхности (незатененные области) от энергетического запаса линии связи при 0,99 доступности канала для трех равномерно размещенных геостационарных спутников; f = 44 ГГц, энергетический запас канала равен 14 дБ. (Перепечатано с разрешения Lincoln Laboratory из L. M. Schwab. "World-Wide Link Availability for Geostationary and Critically Inclined Orbits Including Rain Effects", Lincoln Laboratory, Rep. DCA-9, *Jan., 27, 1981, Fig. 17, p. 42.)*

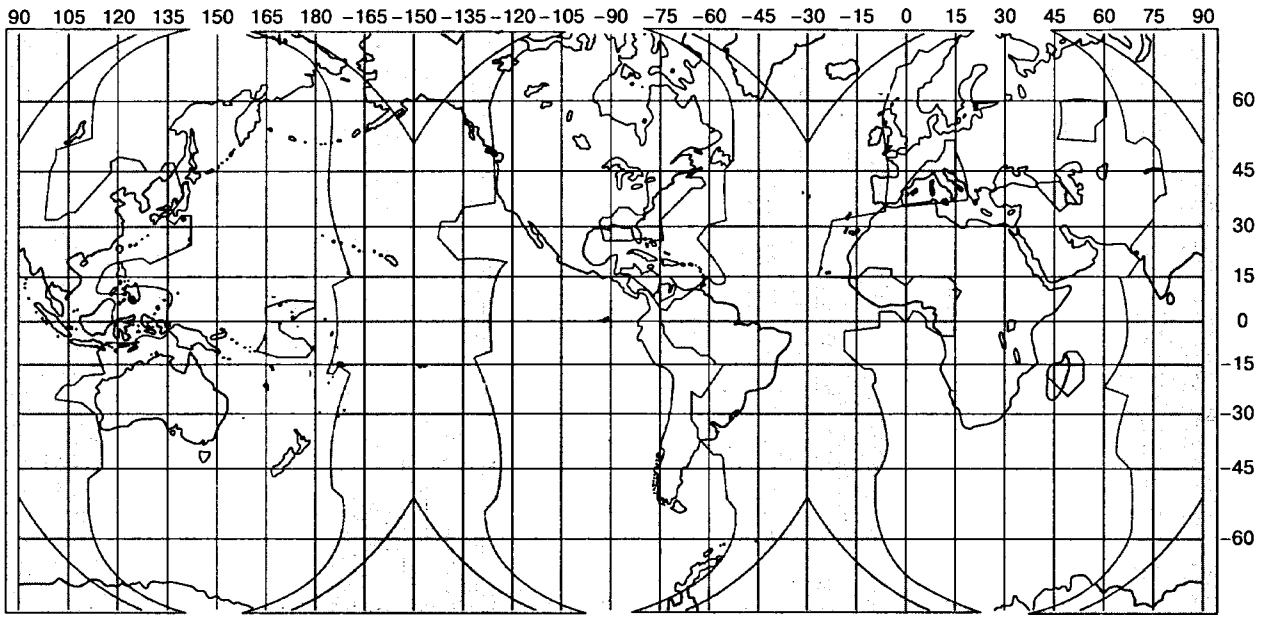


Рис. 5.12. Зависимость охвата земной поверхности (незатененные области) от энергетического запаса линии связи при 0,99 доступности канала для трех равномерно размещенных геостационарных спутников; $f = 44$ ГГц, энергетический запас канала равен 10 дБ. (Перепечатано с разрешения Lincoln Laboratory из L. M. Schwab. "World-Wide Link Availability for Geostationary and Critically Inclined Orbits Including Rain Effects", Lincoln Laboratory, Rep. DCA-9, *Jan., 27, 1981, Fig. 18, p. 43.)*

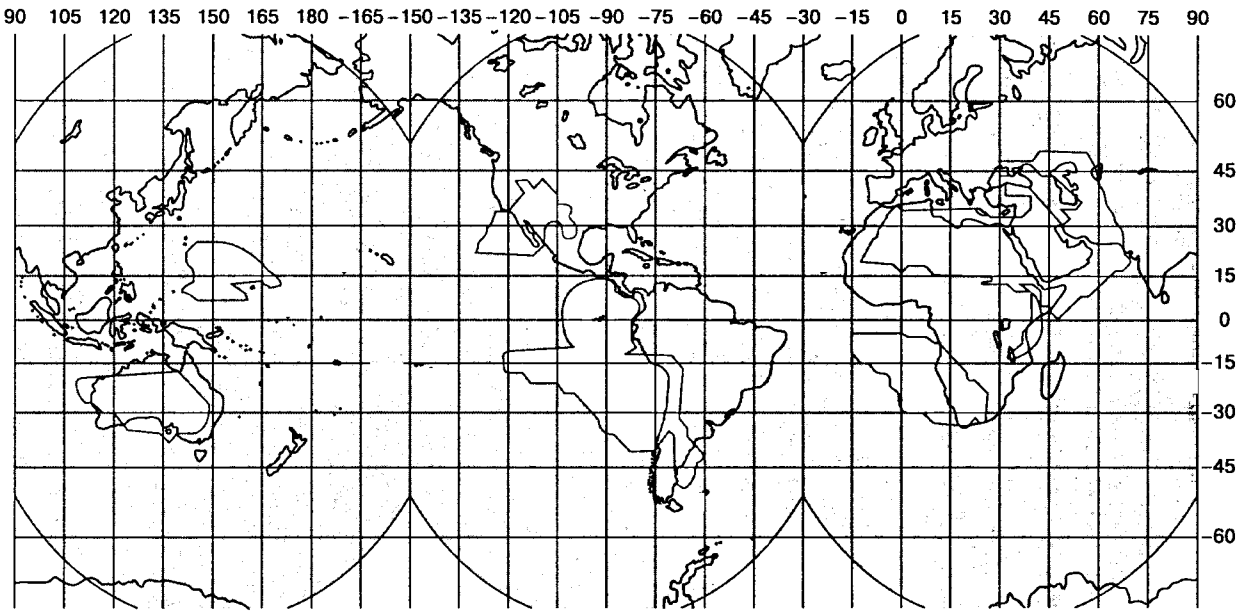


Рис. 5.13. Зависимость охвата земной поверхности (незатененные области) от энергетического запаса линии связи при 0,99 доступности канала для трех равномерно размещенных геостационарных спутников; f = 44 ГГц, энергетический запас канала равен 6 дБ. (Перепечатано с разрешения Lincoln Laboratory из L. M. Schwab. "World-Wide Link Availability for Geostationary and Critically Inclined Orbits Including Rain Effects", Lincoln Laboratory, Rep. DCA-9, *Jan., 27, 1981, Fig. 19, p. 44.)*

5.5. Коэффициент шума, шумовая температура системы

5.5.1. Коэффициент шума

Коэффициент шума F (или шум-фактор) (noise figure) связывает значение параметра SNR на входе сети со значением на выходе. Таким образом шум-фактор измеряет ухудшение SNR, вызванное прохождением через сеть. Пример сказанного приведен на рис. 5.14. На рис. 5.14, *а* показано значение параметра SNR на *входе усилителя* (обозначено как $(SNR)_{in}$) в зависимости от частоты. Максимальное значение на 40 дБ превышает минимальный уровень шума. На рис. 5.14, *б* значение параметра SNR показано на выходе усилителя (обозначено как $(SNR)_{out}$). За счет усиления на усилителе мощность сигнала возросла на 20 дБ, но при этом усилитель добавил к сигналу собственный шум. Максимальное значение сигнала на выходе всего на 30 дБ превышает минимальный уровень шума. Получаем, что ухудшение SNR на пути от входа до выхода составляет 10 дБ; это равносильно утверждению, что коэффициент шума усилителя равен 10 дБ. Коэффициент шума — это параметр, выражающий шумовые свойства двухпортовой сети или некоторого устройства, такого как усилитель, относительно эталонного источника шума в входном порту. Записать шум-фактор можно следующим образом.

$$F = \frac{(SNR)_{in}}{(SNR)_{out}} = \frac{S_i / N_i}{GS_i / G(N_i + N_{ai})}, \tag{5.25}$$

где

S_i — мощность сигнала во входном порту усилителя
N_i — мощность шума во входном порту усилителя
N_{ai} — шум усилителя относительно входного порта
G — коэффициент усиления усилителя

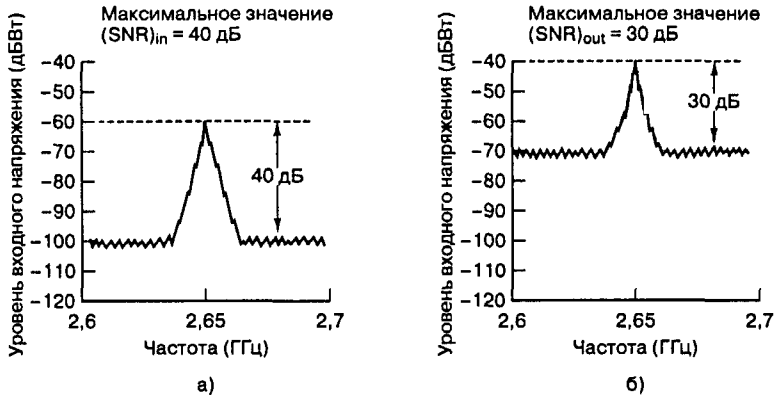


Рис. 5.14. Уровни шума и сигнала усилителя как функция частоты: а) вход усилителя; б) выход усилителя

Иллюстрация уравнения (5.25) приведена на рис. 5.15. На рис. 5.15, *а* представлен *реализуемый усилитель* с коэффициентом усиления $G = 100$ и мощностью внутреннего шума $N_a = 10$ мкВт. Мощность источника шума, внешнего по отношению к усилителю, равна $N_i = 1$ мкВт. На рис. 5.15, *б* усилитель предполагается *идеальным*, и мы при-

писали шумовые свойства реального усилителя, изображенного на рис. 5.15, *а*, внешнему источнику N_{ai}, последовательно соединенному с исходным источником N_i. Значение N_{ai} получается путем уменьшения N_a на величину, равную коэффициенту усиления усилителя. Как показано на рис. 5.15, *б*, уравнение (5.25) соотносит все шумы с входом усилителя, независимо от того, где в действительности присутствует шум — на входе устройства или вне его. Как видно из рис. 5.15, мощность шума на выходе реального усилителя идентична тому, что дает эквивалентная модель.

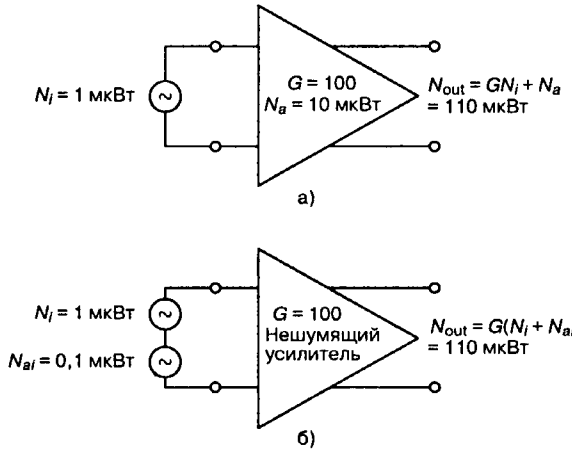


Рис. 5.15. Пример трактовки шума в усилителях

После упрощения уравнения (5.25) получаем следующее.

$$F = \frac{N_i + N_{ai}}{N_i} = 1 + \frac{N_{ai}}{N_i} \tag{5.26}$$

Из полученного уравнения видим, что коэффициент шума выражает шумовые свойства сети относительно входного источника шума; коэффициент шума — это *не* абсолютная мера шума. Идеальный усилитель или идеальная сеть, не вносящие шума ($N_{ai} = 0$), имеют шум-фактор, равный единице (0 дБ).

Для практического использования понятия *шум-фактор* мы должны научиться делать объективные сравнения устройства на основе уравнения (5.26). Следовательно, в качестве *эталонного* мы должны выбрать значение N_i. Шум-фактор любого устройства будет представлять меру того, насколько более шумным (по сравнению с эталонным) является рассматриваемое устройство. В 1944 году Фриис (Friis) [9] предложил, чтобы шум-фактор определялся для источника шума при эталонной температуре $T_0° = 290$ К. Впоследствии это предложение было принято IEEE как часть стандартного определения шум-фактора [10]. Из уравнения (5.17) видим, что для задания максимальной доступной спектральной плотности мощности шума из любого источника достаточно задать температуру этого источника. Значение 290 К было выбрано в качестве эталонного, поскольку именно оно является разумной приближенной оценкой температуры источника большинства каналов связи. Кроме того, если выбрать $T_0° = 290$ К, то вычисление спектральной плотности шума N_0 при этой температуре дает эстетически красивое значение.

$$N_0 = \kappa T_0° = 1{,}38 \times 10^{-23} \times 290 = 4{,}00 \times 10^{-21} \text{ Вт/Гц}$$

или (в децибелах)

$$N_0 = -204 \text{ дБВт/Гц}$$

Теперь, когда мы определили шум-фактор F относительно источника шума с температурой 290 К, важно отметить, что соотношения (5.25) и (5.26) справедливы строго, только если N_i — это источник шума с температурой 290 К. При других N_i нужно переименовать коэффициент F в уравнениях (5.25) и (5.26) и использовать термин *эксплуатационный коэффициент шума* F_{op}. Связь F_{op} и F показана ниже, в уравнении (5.48).

5.5.2. Шумовая температура

Преобразовав уравнение (5.26), можем записать следующее.

$$N_{ai} = (F - 1)N_i \tag{5.27}$$

Из уравнения (5.16) можем подставить $N_i = \kappa T_0° W$ и $N_{ai} = \kappa T_R° W$, где $T_0°$ — эталонная температура источника, а $T_R°$ — *эффективная шумовая температура* приемника (или сети). Затем можем записать следующее.

$$\kappa T_R° W = (F - 1)\kappa T_0° W$$

или

$$T_R° = (F - 1)\, T_0°$$

Температура $T_0°$ выбрана равной 290 К, поэтому получаем следующее.

$$T_R° = (F - 1)\, 290 \text{ К} \tag{5.28}$$

В уравнении (5.26) понятие коэффициента шума использовано для описания шумовых характеристик усилителя. Уравнение (5.28) — это альтернативная (и при этом эквивалентная) характеристика, именуемая *эффективной шумовой температурой*. Напомним, что шум-фактор — это измерение относительно эталона. Шумовая температура такого ограничения не имеет.

Характеристики источников шума (в контексте уравнения (5.17)) можно описывать как через доступную спектральную мощность шума, так и эффективную шумовую температуру. Уравнение (5.28) показывает, что шумовые свойства усилителя можно смоделировать с помощью введения дополнительного источника шума, подобного изображенному на рис. 5.15, *б*, работающего при некоторой эффективной температуре, обозначенной $T_R°$. Для чисто резистивного оконечного устройства $T_R°$ всегда превышает температуру окружающей среды (разумеется, если устройство не охлаждается специально). Важно заметить, что в реактивных оконечных устройствах, таких как неохлаждаемые параметрические усилители или другие малошумящие устройства, $T_R°$ может быть значительно меньше 290 К, даже если температура окружающей среды выше этой величины [11]. Чтобы записать выход усилителя как функцию его эффективной температуры, мы можем использовать уравнения (5.16), (5.25) и (5.28):

$$N_{\text{out}} = GN_i + Gn_{ai} = \tag{5.29,а}$$

$$= G\kappa T_g° W + G\kappa T_R° W = G\kappa(T_g° + T_R°)W = \tag{5.29,б}$$

$$= G\kappa T_g° W + (F - 1)G\kappa T_0° W, \tag{5.29,в}$$

где $T_g°$ — температура источника, а $T_0°$ равна 290 К.

5.5.3. Потери в линии связи

Отличия между сетями усилителей и сетями с потерями в линии можно рассматривать в контексте механизмов *потерь* и *шумов*, описанных ранее. Сети с шумами рассматривались в разделах 5.5.1 и 5.5.2 и подразумевали использование усилителей. Говорилось, что ухудшение параметра SNR происходит вследствие введения в линию связи дополнительного шума (усилителя), как показано на рис. 5.15. В то же время в случае линии с потерями мы должны показать, что ухудшение параметра SNR происходит вследствие поглощения сигнала при фиксированном уровне шума (когда температура линии меньше (или равна) температуры источника). Впрочем, и в этом случае ухудшение будет выражено через увеличение коэффициента шума или эффективной шумовой температуры.

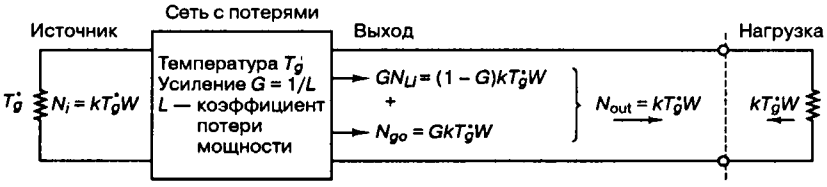


Рис. 5.16. Линия с потерями: импеданс и температура согласованы на обоих концах

Рассмотрим линию (или сеть) с потерями, показанную на рис. 5.16. Предположим, что линия согласована с источником и нагрузкой по импедансу. Определим потерю мощности следующим образом.

$$L = \frac{\text{мощность на входе}}{\text{мощность на выходе}}$$

Коэффициент усиления сети G равен $1/L$ (меньше единицы для линии с потерями). Пусть все компоненты работают с температурой $T_g°$. Общий шум, поступающий с выхода сети в нагрузку, равен

$$N_{out} = \kappa T_g° W,$$

поскольку при температуре $T_g°$ выход сети выглядит как чистое сопротивление. Для обеспечения теплового равновесия общая мощность, поступающая с нагрузки обратно в сеть, также должна равняться N_0. Напомним, что доступная мощность шума $\kappa T°W$ зависит исключительно от температуры, ширины полосы и согласования импедансов; она не зависит от значения сопротивления. N_{out} можно разбить на два компонента, N_{go} и GN_{Li}.

$$N_{out} = \kappa T_g° W = N_{go} + GNL,i \tag{5.30}$$

где

$$N_{go} = G\kappa T_g° W \tag{5.31}$$

является компонентом выходной мощности шума, связанным с источником, GK_{Li} — компонентом выходной мощности шума, отвечающим за сеть с потерями, а N_{Li} — шумом сети, измеряемым относительно ее входа. Объединяя уравнения (5.30) и (5.31), можем записать следующее.

$$\kappa T_g° W = G\kappa T_g° W + GN_{Li} \tag{5.32}$$

Выразим N_{Li}

$$N_{Li} = \frac{1-G}{G} \kappa T_g^\circ W = \kappa T_L^\circ W \tag{5.33}$$

Следовательно, эффективная шумовая температура линии равна следующему.

$$T_L^\circ = \frac{1-G}{G} T_g^\circ \tag{5.34}$$

Поскольку $G = 1/L$, то

$$T_L^\circ = (L-1) T_g^\circ \tag{5.35}$$

В качестве эталонной температуры выберем $T_g^\circ = 290$ К. Тогда можем записать следующее.

$$T_L^\circ = (L-1)290 \text{ К} \tag{5.36}$$

С помощью уравнений (5.28) и (5.36) можем выразить *шум-фактор для линии с потерями.*

$$F = 1 + \frac{T_L^\circ}{290} = L \tag{5.37}$$

Если сеть является линией с потерями, такой что $F = L$ и $G = 1/L$, то N_{out} в уравнении (5.29,в) приобретает следующий вид.

$$N_{out} = \frac{\kappa T_g^\circ W}{L} + \left(1 - \frac{1}{L}\right) \kappa T_0^\circ W \tag{5.38}$$

Отметим, что некоторые авторы используют параметр L для обозначения величины, обратной к введенному нами коэффициенту потерь. В таких случаях шум-фактор $F = 1/L$.

Пример 5.4. Линия с потерями

Линия с температурой $T_0^\circ = 290$ К проложена от источника с шумовой температурой $T_g^\circ = 1450$ К. Мощность входящего сигнала S_i равна 100 пиковатт (пВт), а ширина полосы сигнала W — 1 ГГц. Коэффициент потерь линии $L = 2$. Определите $(SNR)_{in}$, эффективную температуру линии T_L°, мощность выходного сигнала S_{out} и $(SNR)_{out}$.

Решение

$$N_i = \kappa T_g^\circ W = 1{,}38 \times 10^{-23} \text{ Вт/КГц} \times 1450 \text{ К} \times 10^9 \text{ Гц} =$$

$$= 2 \times 10^{-11} \text{ Вт} = 20 \text{ пВт}$$

$$(SNR)_{in} = \frac{100 \text{ пВт}}{20 \text{ пВт}} = 5 \text{ (7 дБ)}$$

$$T_L^\circ = (L-1)\ 290 \text{ К} = 290 \text{ К}$$

$$S_{out} = \frac{S_i}{L} = \frac{100 \text{ пВт}}{2} = 50 \text{ пВт}$$

Используя уравнение (5.29), получаем следующее.

$$N_{\text{out}} = \frac{\kappa T_g \circ W}{L} + \left(1 - \frac{1}{L}\right) \kappa T_0 \circ W =$$

$$= \frac{2 \times 10^{-11}}{2} \text{Вт} + \frac{1}{2}(4 \times 10^{-12}) \text{Вт} = 12 \text{ пВт}$$

и

$$(\text{SNR})_{\text{out}} = \frac{50 \text{ пВт}}{12 \text{ пВт}} = 4{,}17 \ (6{,}2 \text{ дБ})$$

5.5.4. Суммарный шум-фактор и общая шумовая температура

Если две сети соединены последовательно, как показано на рис. 5.17, *а*, суммарный шум-фактор можно записать следующим образом.

$$F_{\text{общ}} = F_1 + \frac{F_2 - 1}{G_1} \tag{5.39}$$

Здесь G_1 — коэффициент усиления, связанный с сетью 1. Если последовательно соединены n сетей, выражение (5.39) приобретает следующий вид.

$$F_{\text{общ}} = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \ldots + \frac{F_n - 1}{G_1 G_2 \cdots G_{n-1}} \tag{5.40}$$

Можете ли вы, изучив уравнение (5.40), предположить, чем следует руководствоваться при проектировании входного каскада приемника (особенно первого каскада или первой пары каскадов)? На входе приемника сигнал более уязвим к дополнительному шуму; следовательно, первый каскад должен иметь максимально низкий шум-фактор F_1. Кроме того, поскольку шум-фактор каждого последующего каскада ослабляется на коэффициенты усиления предыдущих каскадов, это приводит к тому, что мы стремимся получить максимально возможный коэффициент G_1. Одновременное получение максимально низкого F_1 и максимально высокого G_1 — задачи противоречивые; следовательно, всегда необходим некоторый компромисс.

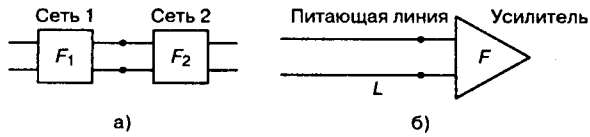


Рис. 5.17. Сети, соединенные последовательно

Уравнения (5.40) и (5.28) можно объединить и выразить эффективную шумовую температуру последовательности n каскадов.

$$T_{\text{общ}} \circ = T_1 \circ + \frac{T_2 \circ}{G_1} + \frac{T_3 \circ}{G_1 G_2} + \ldots + \frac{T_n \circ}{G_1 G_2 \cdots G_{n-1}} \tag{5.41}$$

На рис. 5.17, *б* показана питающая линия, последовательно соединенная с усилителем; после этого обычно следует принимающая антенна. Используя уравнение (5.39) для нахождения $F_{\text{общ}}$ подобной линии с потерями, можем записать следующее:

$$F_{\text{общ}} = L + L(F - 1) = LF, \tag{5.42}$$

поскольку шум-фактор линии с потерями равен L, а коэффициент усиления линии — $1/L$. По аналогии с уравнением (5.36) общую температуру можно записать следующим образом.

$$T_{\text{общ}} = (LF - 1)290 \text{ K} \tag{5.43}$$

Общую температуру канала и усилителя можно также записать иначе.

$$\begin{aligned} T_{\text{общ}}^{\circ} &= (LF - 1 + L - L)290 \text{ K} = \\ &= [(L - 1) + L(F - 1)]290 \text{ K} = \\ &= T_L^{\circ} + LT_R^{\circ} \end{aligned} \tag{5.44}$$

5.5.4.1. Сравнение шум-фактора и шумовой температуры

Поскольку и шум-фактор F и эффективная шумовая температура T° характеризуют шумовые характеристики устройств, некоторые инженеры вынуждены выбирать одну из этих мер. В то же время оба параметра имеет четко определенную "сферу деятельности". Для наземных приложений практически универсальным является шум-фактор F; здесь понятие ухудшения параметра SNR для источника с температурой 290 K имеет смысл, поскольку температура наземных источников обычно близка к 290 K. Значения наземных шум-факторов обычно принадлежат диапазону 1–10 дБ.

Для космических приложений более удобным критерием качества является параметр T°. Диапазон температур для коммерческих систем обычно находится между 30 и 150 K. Недостатком использования шум-факторов для подобных малошумящих сетей является то, что все получаемые значения близки к единице (0,5–1,5 дБ), что создает определенные затруднения при сравнении устройств. Для малошумящих приложений F (в децибелах) необходимо выражать с точностью до двух знаков после запятой, чтобы оно давало разрешение или точность, сравнимую с точностью, которую даст T°. Для приложений космической связи эталонная температура в 290 K не является настолько подходящей, как для наземных приложений. Если же использовать эффективную температуру, то для описания ухудшения никакой эталонной температуры не требуется (разве что абсолютный нуль K). Эффективная входная шумовая температура просто сравнивается с эффективной шумовой температурой источника. Вообще, приложения, в которых фигурируют малошумящие устройства, лучше описывать с помощью эффективной температуры, а не шум-фактора.

5.5.5. Эффективная температура системы

На рис. 5.18 представлена упрощенная схема принимающей системы, причем указаны те области (антенна, линия связи и предварительный усилитель), которые играют основную роль в ухудшении параметра SNR. Влияние предварительного усилителя уже обсуждалось ранее — оно заключается во введении в линию дополнительного шума. Также рассматривались потери в линии — сигнал поглощается при фиксированном уровне шума (если температура линии меньше (или равна) температуры источника). Оставшиеся источники ухудшения качества сигнала могут быть как естественными, так и искусственными. Естественные источники — это молнии, небесные источники радиоизлучения, атмосферные источники и тепловое излучение от земли и других физических структур. Искусственные — это излучение от автомобильных систем зажигания и других электрических приборов, а также радиопередача от других пользователей, использующих ту же полосу, что и приемник. Общий объем шума, вносимого

перечисленными внешними источниками, можно описать как $kT_{ant}W$, где T_{ant} является *температурой антенны*. Антенна подобна линзе: вносимый ею шум определяется тем, "на что смотрит антенна". Если антенна нацелена на прохладную область неба, вводится крайне малый объем теплового шума. Температура антенны — это мера эффективной температуры, проинтегрированной по всей поверхности антенны.

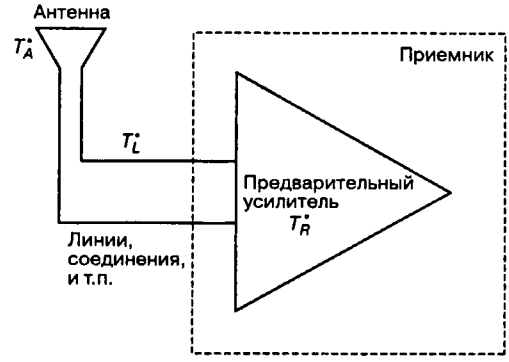


Рис. 5.18. Основные источники шума принимающей системы

Теперь мы можем определить *температуру системы* $T_S°$, сложив все вклады в шум системы (выраженные через эффективную температуру). Суммарное выражение выглядит следующим образом.

$$T_S° = T_A° + T_{общ}° \tag{5.45}$$

Здесь $T_A°$ — температура антенны, а $T_{общ}°$ — общая температура линии и предварительного усилителя. В уравнении (5.45) указаны два основных источника шума и интерференции, вызывающие ухудшение качества работы приемника. Один источник, описываемый членом $T_A°$, представляет ухудшение работоспособности, навязываемое "внешним миром", проходящим через антенну. Второй источник, описываемый членом $T_{общ}°$, — это тепловой шум, вызванный движением электронов во всех проводниках. Поскольку температура системы $T_S°$ — это новая суммарная температура, включающая $T_A°$ и суммарную эффективную температуру линии и предварительного усилителя, может возникнуть вопрос: почему уравнение (5.45) не содержит тех же множителей последовательного уменьшения, что и в уравнении (5.41)? Мы предполагаем, что антенна *не имеет диссипативных частей*; ее коэффициент усиления, в отличие от усилителя или аттенюатора, может рассматриваться как коэффициент расширения спектра сигнала. Какая бы эффективная температура не вводилась при проходе через антенну, это не зависит от самой антенны; антенна представляет шум источника (или температуру источника) на входе линии.

Используя уравнение (5.44), мы можем модифицировать уравнение (5.45) следующим образом.

$$\begin{aligned} T_S° &= T_A° + T_L° + LT_R° = \\ &= T_A° + (L-1)290\text{ K} + L(F-1)290\text{ K} = \\ &= T_A° + (LF-1)290\text{ K} \end{aligned} \tag{5.46}$$

(5.47)

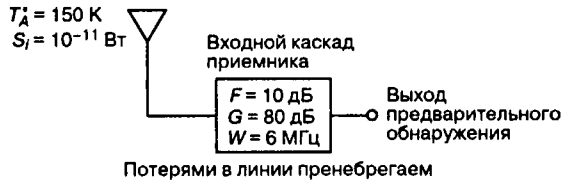
Если LF выражено в децибелах, мы должны вначале изменить его размерность, и $T_S°$ приобретет следующий вид.

$$T_S° = T_A° + (10^{LF/10} - 1)290 \text{ K}$$

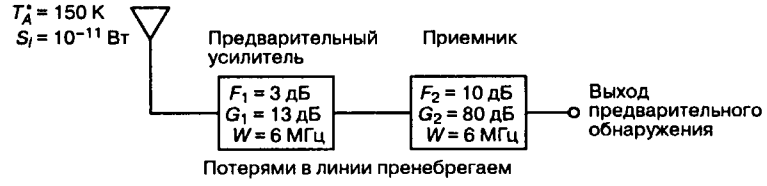
Уравнения (5.45)–(5.47) описывают температуру системы T_s на оконечных устройствах принимающей антенны, а уравнения (5.10) и (5.11) — мощность P_r, полученную принимающей антенной. Данные определения используются в этой главе; кроме того, их предпочитают разработчики систем, антенн, а также люди, работающие на передающей стороне линии. Важно отметить, что существует альтернативный набор определений, используемых разработчиками систем, которые предпочитают описывать температуру системы T_S' и принятую мощность P_r' на входе приемника. Если предположить, что антенна и приемник связаны устройством, которое не сложнее линии с потерями, то параметры T_S и T_S' (как и P_r' и P_r) отличаются в L раз (напомним, что L — коэффициент потерь в линии). Иными словами, $T_S = LT_S'$ и $P_r = LP_r'$. При вычислении принятого SNR (определяемого в следующем разделе) с помощью определений принятой мощности и температуры системы, соотнесенных с приемником, результат не будет отличаться от того, который был получен при использовании определений, связанных со входом приемника. Причина в том, что множитель L входит и в числитель, и в знаменатель отношения SNR, поэтому он просто сокращается.

Пример 5.5. Шум-фактор и температура шума

На входе приемника, показанном на рис. 5.19, *а*, шум-фактор равен 10 дБ, усиление равно 80 дБ, а ширина полосы — 6 МГц. Мощность сигнала на входе S_i равна 10^{-11} Вт. Допустим, что потери в линии отсутствуют и температура антенны равна 150 K. Найдите $T_R°$, $T_S°$, N_{out}, $(SNR)_{in}$ и $(SNR)_{out}$.



а)



б)

Рис. 5.19. Улучшение входного каскада приемника за счет малошумящего предварительного усилителя

Решение

Вначале преобразуем все значения в децибелах в размерные величины.

$$T_R° = (F - 1)290 \text{ K} = 2610 \text{ K}$$

Использование уравнения (5.46) при $L = 1$ для малошумящей линии дает следующее.

$$T_S^\circ = T_A^\circ + T_R^\circ = 150\ \text{K} + 2610\ \text{K} = 2760\ \text{K}$$

$$N_{\text{out}} = G\kappa T_A^\circ + G\kappa T_R^\circ W = G\kappa T_S^\circ W =$$

$$= 10^8 \times 1{,}38 \times 10^{-23} \times 6 \times 10^6 (150\ \text{K} + 2610\ \text{K}) =$$

= 1,2 мкВт (вклад от источника) + 21,6 мкВт (вклад от входного каскада) = 22,8 мкВт

$$(\text{SNR})_{\text{in}} = \frac{S_i}{\kappa T_A^\circ W} = \frac{10^{-11}}{1{,}24 \times 10^{-14}} = 806{,}5\ (29{,}1\ \text{дБ})$$

$$(\text{SNR})_{\text{out}} = \frac{S_{\text{out}}}{N_{\text{out}}} = \frac{10^8 \times 10^{-11}}{22{,}8 \times 10^{-6}} = 43{,}9\ (16{,}4\ \text{дБ})$$

Заметим, что в приведенном примере шум усилителя значительно больше шума источника и является основной причиной ухудшения параметра SNR.

Пример 5.6. Улучшение параметра SNR с помощью малошумящего предварительного усилителя

Используйте предварительный усилитель, как показано на рис. 5.19, *б*, с шум-фактором 3 дБ, усилением 13 дБ и шириной полосы 6 МГц для улучшения SNR приемника, описанного в примере 5.5. Определите $T_{\text{общ}}^\circ$ объединения предварительного усилителя и приемника. Найдите T_s°, $F_{\text{общ}}$, N_{out} и $(\text{SNR})_{\text{out}}$. Потери в линии будем считать нулевыми.

Решение

Как и ранее, вначале все значения, выраженные в децибелах, приводятся к размерному виду.

$$T_{R1}^\circ = (F_1 - 1)290\ \text{K} = 290\ \text{K}$$

$$T_{R2}^\circ = (F_2 - 1)290\ \text{K} = 2610\ \text{K}$$

$$T_{\text{общ}}^\circ = T_{R1}^\circ + \frac{T_{R2}^\circ}{G_1} = 290\ \text{K} + \frac{2610\ \text{K}}{20} = 420{,}5\ \text{K}$$

$$T_S^\circ = T_A^\circ + T_{\text{общ}}^\circ = 150\ \text{K} + 420{,}5\ \text{K} = 570{,}5\ \text{K}$$

$$F_{\text{общ}} = F_1 + \frac{F_2 - 1}{G_1} = 2 + \frac{9}{20} = 2{,}5\ (4\ \text{дБ})$$

$$N_{\text{out}} = G\kappa T_A^\circ W + G\kappa T_{\text{общ}}^\circ W = G\kappa T_S^\circ W =$$

$$= 20 \times 10^8 \times 1{,}38 \times 10^{-23} \times 6 \times 10^6 (150\ \text{K} + 420{,}5\ \text{K}) =$$

= 24,8 мкВт (вклад источника) + 69,6 мкВт (вклад входного каскада) = 94,4 мкВт

$$(SNR)_{\text{out}} = \frac{S_{\text{out}}}{N_{out}} = \frac{10^{-11} \times 20 \times 10^8}{94{,}4 \times 10^{-6}} = 212{,}0\ (23{,}3\ \text{дБ})$$

Итак, при добавлении предварительного усилителя выходной шум увеличивается (с 22,8 мкВт в примере 5.5) до 94,4 мкВт. И все же, несмотря на увеличение мощности шума, более низкая температура системы приводит к улучшению параметра SNR на 6,9 дБ (с 16,4 дБ в примере 5.5 до 23,3 дБ в данном примере). Цена, которую мы платим за это улучшение, — необходимость улучшения $F_{\text{общ}}$ на 6 дБ (с 10 дБ в примере 5.5 до 4 дБ в данном примере).

Нежелательный шум частично *вносится посредством антенны* ($\kappa T_A^\circ W$) и частично *генерируется внутренне* в входном каскаде приемника ($\kappa T_{\text{общ}}^\circ W$). Объем улучшения системы, который может дать проектирование входного каскада, зависит от того, какая часть общего шума вносится входным каскадом. Из примера 5.5 мы видели, что входной каскад вносит большую часть шума. Следовательно, как было сделано в примере 5.6, обеспечение малошумящего предварительного усилителя значительно улучшает системное отношение сигнал/шум (SNR). В следующем примере рассматривается, когда большая часть шума вносится посредством антенны; мы увидим, что в этом случае введение малошумящего предварительного усилителя не дает ощутимого улучшения параметра SNR.

Пример 5.7. Попытка улучшения SNR при больших значениях T_A°

Повторите примеры 5.6 и 5.5 с единственным изменением: пусть T_A° =8000 К. Другими словами, большая часть шума теперь вносится антенной; допустим, все поле зрения антенны заполняет очень горячее тело (солнце). Вычислите улучшение параметра SNR, которое дается предварительным усилителем, использованным в примере 5.6 (рис. 5.19, *б*), после чего сравните результат с ответом примера 5.6.

Решение

Без предварительного усилителя

$$N_{\text{out}} = G\kappa W(T_A^\circ + T_R^\circ) =$$

$$= 10^8 \times 1{,}38 \times 10^{-23} \times 6 \times 10^6 (8000\ \text{K} + 2610\ \text{K}) =$$

= 66,2 мкВт (вклад источника) + 21,6 мкВт (вклад входного каскада) = 87,8 мкВт

$$(\text{SNR})_{\text{out}} = \frac{S_{\text{out}}}{N_{\text{out}}} = \frac{10^8 \times 10^{-11}}{87{,}8 \times 10^{-6}} = 11{,}4\ (10{,}6\ \text{дБ})$$

С предварительным усилителем

$$N_{\text{out}} = 20 \times 10^8 \times 1{,}38 \times 10^{-23} \times 6 \times 10^6 (8000\ \text{K} + 420{,}5\ \text{K}) =$$

= 1324,8 мкВт (вклад источника) + 69,6 мкВт (вклад входного каскада) = 1394,4 мкВт

$$(\text{SNR})_{\text{out}} = \frac{20 \times 10^8 \times 10^{-11}}{1{,}39 \times 10^{-3}} = 14{,}4\ (11{,}6\ \text{дБ})$$

Таким образом, в данном случае улучшение параметра SNR равно всего 1 дБ, что значительно меньше полученных ранее 6,9 дБ. Если основные источники шума находятся внутри приемника, улучшить SNR можно за счет введения малошумящих устройств. В то же время, если основные источники шума являются внешними, то улучшение входного каскада приемника не имеет существенного значения.

Шум-фактор — это определение, основанное на использовании эталонного значения 290 К. Если температура источника отличается от 290 К, как в примерах 5.5–5.7, то необходимо определить *рабочий* или *эффективный шум-фактор*, описывающий реальную зависимость между $(\text{SNR})_{\text{in}}$ и $(\text{SNR})_{\text{out}}$. Если в качестве отправной точки использовать уравнения (5.25) и (5.27), рабочий шум-фактор можно выразить следующим образом.

$$F_{\text{раб}} = \frac{S_i / kT_A W}{GS_i / G(kT_A W + N_{ai})} =$$
$$= \frac{kT_A W + N_{ai}}{kT_A W} = 1 + \frac{(F-1)kT_0 W}{kT_A W} = \tag{5.48}$$
$$= 1 + \frac{T_0}{T_A}(F-1)$$

5.5.6. Шумовая температура неба

Принимающая антенна собирает случайные шумы, излученные галактикой, солнцем и наземными источниками, что вместе составляет фоновый шум неба. Фон неба появляется как комбинация галактического воздействия, уменьшающегося с частотой, и атмосферного воздействия, которое становится существенным при частоте порядка 10 ГГц (и увеличивается с частотой). Пример температуры неба, измеренной с земли, приведен на рис. 5.20 (учтены оба названных механизма). Заметим, что существует область между 1 и 10 ГГц, где температура достигает наименьшего значения; галактический шум становится достаточно малым при 1 ГГц и для спутниковой связи шум излучения абсолютно черного тела (вследствие поглощения атмосферой) не является существенным, если он ниже 10 ГГц. (Для других приложений, например пассивной радиометрии, это по-прежнему является проблемой.) Эта область, известная как *микроволновое* (или *космическое*) *окно*, представляет особый интерес для спутниковой связи или космической дальней связи. Низкий шум неба — это основная причина того, что системы в основном используют несущие частоты, принадлежащие этой части спектра. Кривые на рис. 5.20, показывающие галактический и атмосферный шумы, показаны в виде семейства кривых с разными углами возвышения θ. При $\theta = 0$ принимающая антенна направлена на линию горизонта, и в процессе распространения сигнал проходит наибольший возможный путь через атмосферу. При $\theta = 90°$ антенна направлена на зенит, и минимальная часть пути сигнала приходится на атмосферу. Таким образом, верхняя кривая семейства демонстрирует почти наихудшую (*почти* — потому что погода считается ясной) зависимость температуры шума от частоты, а нижняя представляет наиболее благоприятный случай. На рис. 5.20 также показан график зависимости температуры шума от частоты *при дожде*. Поскольку интенсивность любого ливня можно выразить только статистически, показанные температуры шума — это значения, когда дожди идут 25% времени (в зените). Какая спектральная область является наиболее благоприятной для космической связи, если принимать во внимание дожди? Это нижняя часть космического окна. По этой причине системы, подобные SGLS (Space Ground Link Subsystem) (военные) и Unified S-Band Telemetry, Tracking, and Control System (NASA), расположены в полосе частот 1,8–2,4 ГГц.

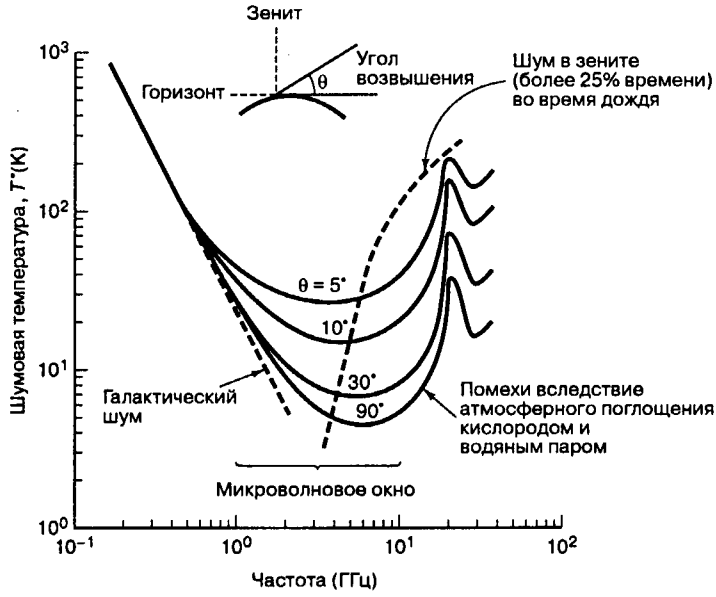


Рис. 5.20. Шумовая температура неба

5.5.6.1. Радиокарта неба

Различные исследователи изображали излучение галактического шума как функцию частоты. На рис. 5.21 представлена подобная карта радиотемператур, взятая из работы [12]. На ней изображены температурные контуры неба в районе 250 МГц при рассмотрении с земли. Вообще, небо состоит из локализированных галактических источников (Солнце, Луна, планеты и т.д.), каждый из которых имеет собственную температуру. Карта — это эффективная взвешенная сумма температур отдельных галактических источников плюс постоянный фон неба. Координаты карты, *склонение* и *прямое восхождение*, можно рассматривать как небесную широту и долготу относительно земной поверхности (прямое восхождение измеряется в часовых углах, причем 24 часа соответствуют полному обороту Земли). На рис. 5.21 температурные контуры показаны для температур от 90 до 1000 К. Измерения проводились так, чтобы воздействие Солнца было исключено (ночное небо). Луч антенны в центре карты указывает размер области неба, в пределах которой производились измерения (каждое измерение — это усреднение по площади луча). Чем уже луч, тем лучше разрешение температурных контуров; чем шире луч, тем разрешение хуже.

На рис. 5.22 представлена другая радиокарта для частоты 600 МГц, взятая из работы [13]. При этой частоте, как было показано на рис. 5.20, галактический шум снижается, по сравнению с рис. 5.21; наиболее низкой из показанных температур является 8 К, наиболее высокой — 280 К. Если внимательно изучить рис. 5.21 и 5.22, то можно обнаружить область наибольшего излучения шума. Она расположена в овальной области в середине правой части каждой карты; продольная ось овала определяет положение на *нашей галактической плоскости*, где подобное излучение космического шума является наиболее интенсивным.

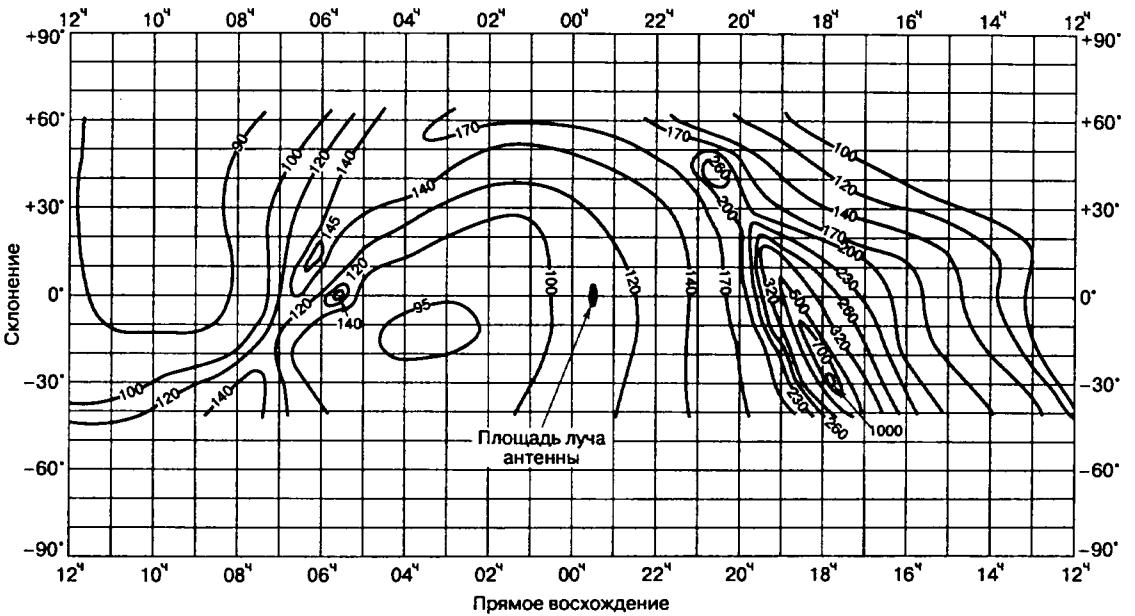


Рис. 5.21. Радиокарта небесного фона при 250 МГц (Перепечатано с разрешения журнала Sky and Telescope, *Cambridge, Mass., из работы H. C. Ko and J. D. Kraus. "A Radio Map of the Sky at 1,2 Meters,"* Sky Telesc., *vol. 16, Feb., 1957, p. 160.)*

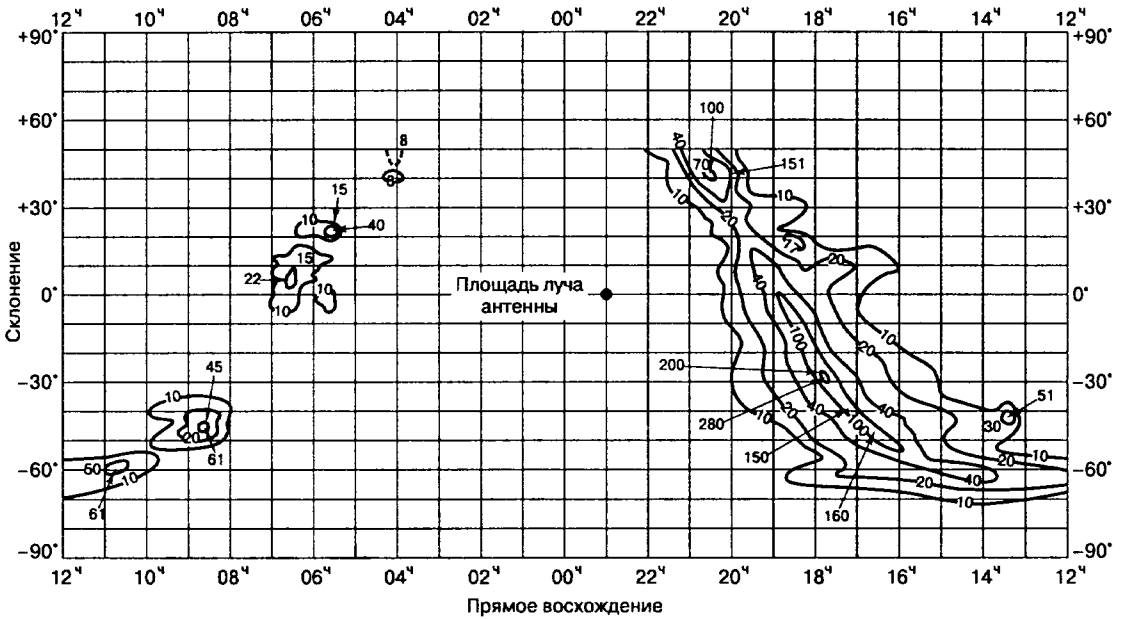


Рис. 5.22. Радиокарта небесного фона при 600 МГц. (Перепечатано с разрешения авторов из работы J. H. Piddington and G. H. Trent. "A survey of Cosmic Radion Emission at 600 Mc/s", Aust. J. Phys., *vol. 9, Dec., 1956, Fig. 1, pp. 483–486.)*

5.6. Пример анализа канала связи

В разделе 5.4 мы вывели соотношения между основными параметрами канала связи. В данном разделе мы используем эти соотношения для расчета простого бюджета канала, показанного в табл. 5.2. Данная таблица может показаться "страшным" перечнем терминов; может создаться впечатление, что бюджет канала представляет сложный процесс обработки имеющейся информации. На самом деле это не так, и для подтверждения этого мы приведем рис. 5.23. На этом рисунке набор пунктов из таблицы сведен к нескольким ключевым параметрам. Вообще, цель анализа канала связи — определить, достигается ли требуемая достоверность передачи. Для этого отношение E_b/N_0 в реально принятом сигнале сравнивается с тем, которое необходимо для удовлетворения спецификации системы. При этом необходимыми являются следующие параметры: EIRP (какая эффективная мощность была передана), добротность $G/T°$ (насколько приемник способен вобрать эту мощность), L_s (наибольшие отдельные потери, потери в свободном пространстве) и L_o (другие вклады в потери и ослабления сигнала). И это *ВСЕ*!

Таблица 5.2. Пример бюджета канала "наземный терминал — спутник": частота — 8 ГГц, расстояние — 21 915 морских миль (40 626 км)

1. Переданная мощность (дБВт)	(100 Вт)	20,0	P_t
2. Потери в передатчике (дБ)		<2,0>	L_o
3. Усиление передающей антенны (максимум дБ[i])		51,6	G_t
Диаметр параболической антенны (футы)	20,00		
Ширина луча половинной мощности (градусы)	0,45		
4. EIRP терминала (дБВт)		69,6	EIRP
5. Потери в тракте (дБ)	(угол возвышения 10°)	<202,7>	L_s
6. Скидка на замирание (дБ)		<4,0>	L_o
7. Другие потери (дБ)		<6,0>	L_o
8. Принятая изотропная мощность (дБВт)		−143,1	
9. Усиление принятой антенны (максимум дБ[i])			
Диаметр параболической антенны (футы)	3,00		
Ширина луча половинной мощности (градусы)	2,99		
10. Потери на границе охвата (дБ)		<2,0>	L_o
11. Мощность принятого сигнала (дБВт)		−110,0	P_r
Шум-фактор приемника в порту антенны (дБ)			11,5
Температура приемника (дБК)			35,8 (3806 К)
Температура принимающей антенны (дБ-К)			24,8 (300 К)
12. Температура системы (дБК)			36,1 (4106 К)
13. $G/T°$ системы (дБК)	−1,0		$G/T°$
14. Константа Больцмана (дБВт/КГц)			−228,60

Окончание табл. 5.2

15. Спектральная плотность шума (дБВт/Гц)		<–192, 5>	$N_0 = kT°$
16. Принятое P_r/N_0 (дБГц)		82,5	$(P_r/N_0)_r$
17. Скорость передачи данных (дББит/с)	(2 Мбит/с)	<63,0>	R
18. Принятое E_b/N_0 (дБ)		19,5	$(E_b/N_0)_r$
19. Потери реализации (дБ)		<1,5>	L_o
20. Требуемое E_b/N_0 (дБ)		<10,0>	$(E_b/N_0)_{треб}$
21. Резерв (дБ)		8,0	M

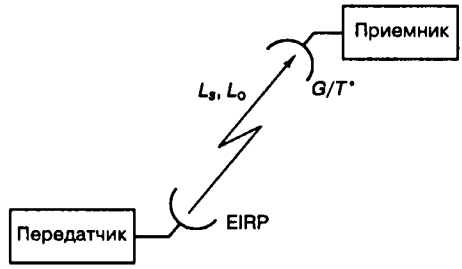


Рис. 5.23. Ключевые параметры анализа канала связи

5.6.1. Элементы бюджета канала

Пример бюджета канала, приведенный в табл. 5.2, состоит из трех столбцов чисел. Собственно бюджетом канала является средний из них. Другие состоят из вспомогательной информации, например информации о ширине луча антенны, или включают вычисления, дополняющие основную таблицу. Потери берутся в скобки (стандартная форма записи при учете использования системных ресурсов). Если значение не заключено в скобки — оно представляет усиление. Промежуточные суммы показаны в прямоугольниках. Начиная с вершины среднего столбца, мы алгебраически суммируем все ослабления и усиления. Окончательный энергетический резерв линии связи заключен в двойной прямоугольник и приведен под номером 21. Вычисления проводятся согласно уравнению (5.24) (ниже оно приводится повторно, только в этот раз параметры G_r и $T°$ собраны вместе в $G_r/T°$).

$$M(\text{дБ}) = \text{EIRP}(\text{дБВт}) + \frac{G_r}{T°}(\text{дБ/К}) - \left(\frac{E_b}{N_0}\right)_{\text{треб}}(\text{дБ}) - R(\text{дББит/с}) -$$
$$- \text{к}(\text{дБВт/Гц}) - L_s(\text{дБ}) - L_o(\text{дБ})$$

Рассмотрим пункты из табл. 5.2 подробнее.

1. Мощность передатчика равна 100 Вт (20 дБВт).
2. Потери в канале между передатчиком и антенной равны 2 дБ.
3. Усиление передающей антенны равно 51,6 дБ[i].
4. Суммарный вклад пп. 1–3 дает EIRP = 69,6 дБВт.

5. Потери в тракте вычисляются для указанного в заголовке таблицы диапазона, соответствующего углу возвышения 10° над наземной оконечной станцией.

6, 7. Скидка на погодное поглощение сигнала и некоторые другие, не указанные, потери.

8. Принятая изотропная мощность — это мощность, которую бы приняла антенна (–143,1 дБВт), если бы была изотропной.

9. Максимальный коэффициент усиления принимающей антенны равен 35,1 дБ[i].

10. Потери на границе охвата, вызванные внеосевым усилением антенны (по сравнению с максимальным усилением) и увеличенным диапазоном для пользователей на краях зоны обслуживания (здесь указаны номинальные потери, равные 2 дБ).

11. Мощность, подаваемая на вход приемника (сумма пп. 8–10), равна –110 дБВт.

12. Температура системы находится с помощью уравнения (5.46). Впрочем, в данном примере мы предполагали, что линия от антенны приемника до входного каскада является линией без потерь, так что коэффициент потерь в линии L равен 1, а температура системы, вычисленная в столбце 3, равна $T_S° = T_A + T_R$.

13. Добротность приемника $G/T°$ определяется при объединении усиления принимающей антенны G_r (см. п. 9) с температурой системы T_S. Как интересующий нас параметр, данное отношение помещается не в центральный столбец, а в левый. Причина в том, что G_r учитывается в п. 9, а T_S — в п. 15. Если поместить $G/T°$ в центральный столбец, это приведет к двойному табулированию указанных величин.

14. Константа Больцмана равна –228,6 дБВт/КГц.

15. Сложение константы Больцмана (в децибелах, п. 14) и температуры системы (в децибелах, п. 12) дает спектральную плотность мощности шума.

16. Мы можем записать спектральную плотность отношения принятого сигнала к шуму 82,5 дБГц, вычтя спектральную плотность шума в децибелах (п. 15) из мощности принятого сигнала в децибелах (п. 11).

17. Скорость передачи данных указана в дБбит/с.

18. Поскольку $E_b/N_0 = (1/R)(P_r/N_0)$, мы должны вычесть R в децибелах (п. 17) из P_r/N_0 в децибелах (п. 16), что дает $(E_b/N_0)_r = 19{,}5$ дБ.

19. Потери реализации (здесь 1,5 дБ) учитывают отличия теоретически предсказанной достоверности обнаружения и работы реального детектора.

20. Это и есть требуемое E_b/N_0, результат выбора модуляции и кодирования и задания вероятности ошибки.

21. Разность принятого и требуемого E_b/N_0 в децибелах (здесь учтены потери реализации) дает окончательный энергетический резерв.

Пункты потери или усиления, показанные в бюджете канала, — первое приближение *идеального* или *упрощенного* результата, за которым следует параметр потерь или усилений, уточняющий этот результат. Другими словами, бюджет канала обычно придерживается *модульного* принципа разделения усилений и ослаблений, чтобы расчет можно было легко приспособить к нуждам любой системы. Рассмотрим следующие примеры этого формата. В табл. 5.2 п. 1 представляет мощность передатчика, которая подается с передатчика посредством изотропной передающей антенны (упрощение). В то же время только после применения модулей потерь в канале и усиления на пере-

дающей антенне (реальный результат) получается передаваемое EIRP, показанное в п. 4. Подобным образом п. 8 показывает мощность, принятую изотропной антенной (упрощение). В то же время только в п. 11 мы увидим (реальную) принятую мощность сигнала после применения модулей усиления принимающей антенны и потерь на границе охвата.

5.6.2. Добротность приемника

Ниже следует объяснение причины частого объединения усиления антенны и температуры системы в единый параметр $G/T°$. На заре развития спутниковой связи G_r и $T_S°$ задавались отдельно. Подрядчик, согласившийся с заданными требованиями, желал оставить себе некоторый резерв для удовлетворения каждого требования в отдельности. Даже если пользователя обычно интересовал лишь конечный результат (итоговая строка бюджета), а не явные значения G_r или $T_S°$, подрядчик не мог использовать потенциальные компромиссы. В результате получалась переопределенная система, более дорогая, чем необходимо. Распознание этой переопределенности привело к определению антенны, входного каскада приемника и единого параметра $G/T°$ (иногда еще называемого *чувствительностью приемника*), так что теперь могли использоваться рентабельные компромиссы между структурами антенны и приемника.

5.6.3. Принятая изотропная мощность

Еще одной областью переопределения структуры приемника является отдельное задание требуемого P_r/N_0 (или E_b/N_0) и $G/T°$ приемника. Если P_r/N_0 и $G/T°$ задаются раздельно, подрядчик обязан получать каждое заданное значение. Он должен планировать некоторый резерв по обоим пунктам. Как и при $G/T°$, рассмотренном в предыдущем разделе, существуют определенные преимущества задания P_r/N_0 и $G/T°$ в виде одного параметра; этот новый параметр, называемый *принятой изотропной мощностью* (received isotropic power — RIP), можно записать следующим образом.

$$\text{RIP(дБВт)} = \frac{P_r}{N_0}\text{(дБГц)} - \frac{G}{T°}\text{(дБ/К)} - \text{к(дБВт/КГц)} \tag{5.49}$$

При переводе в отношения

$$\text{RIP} = \frac{P_r}{\kappa T°}\left(\frac{\kappa T°}{G_r}\right) = \frac{P_r}{G_r} \tag{5.50}$$

Важно отметить, что P_r/N_0 — это отношение спектральных плотностей сигнала и шума (signal-to-noise ratio — SNR) до обнаружения, *требуемое* для получения определенной достоверности передачи при использовании указанной схемы модуляции (обычно в этот параметр включается резерв, учитывающий *потери при реализации детектора*). Обозначим теоретическое SNR, необходимое для получения определенной вероятности ошибки P_B, как $(P_r/N_0)_{\text{т-тр}}$. Затем можем записать следующее.

$$\frac{P_r}{N_0} = L_0'\left(\frac{P_r}{N_0}\right)_{\text{т-тр}} \tag{5.51}$$

Здесь L'_o является потерями реализации и учитывает аппаратные и операционные потери в процессе обнаружения. Объединяя уравнения (5.50) и (5.51), можем записать следующее.

$$\text{RIP} = L'_0\left(\frac{P_r}{\kappa T^\circ}\right)_{\text{т-тр}} \frac{\kappa T}{G_r} \tag{5.52}$$

Задание параметра RIP позволяет подрядчику, перед которым стоит задача получения определенной вероятности ошибки, оперировать значением одного параметра. Подрядчик может использовать связь P_r/N_0 и G/T° или L'_o. При увеличении G/T° производительность детектора может ухудшаться и наоборот.

5.7. Спутниковые ретрансляторы

Спутниковые ретрансляторы повторно передают все получаемые сообщения (с трансляцией на несущую частоту). *Регенеративные* (цифровые) ретрансляторы перед повторной передачей регенерируют, т.е. демодулируют и восстанавливают цифровую информацию, заложенную в принятый сигнал. *Нерегенеративные* ретрансляторы только усиливают и повторно передают сообщение. Следовательно, нерегенеративный ретранслятор может использоваться с различными форматами модуляции (одновременно или последовательно без какой-либо коммутации), а регенеративный обычно проектируется для работы только с одним форматом модуляции (или очень малым количеством). В процессе анализа канала связи для регенеративного спутникового ретранслятора каналы "земля-спутник" и "спутник-земля" рассматриваются раздельно. Для вычисления общей вероятности битовой ошибки в канале регенеративного ретранслятора необходимо отдельно определить вероятности появления ошибочного бита в каждом из двух каналов. Пусть P_u и P_d — вероятность появления ошибочного бита в каналах "земля-спутник" (*uplink*) и "спутник-земля" (*downlink*). Бит будет безошибочно передан между двумя оконечными наземными устройствами, если в обоих последовательных каналах бит будет передан либо точно, либо с ошибкой. Следовательно, общая вероятность точной передачи бита равна следующему.

$$P_c = (1 - P_u)(1 - P_d) + P_u P_d \tag{5.53}$$

Общая вероятность появления ошибочного бита равна

$$P_B = 1 - P_c = P_u + P_d - 2P_u P_d \tag{5.54}$$

При малых значениях P_u и P_d общая вероятность ошибки получается при простом суммировании вероятностей появления ошибки в отдельных каналах.

$$P_B \approx P_u + P_d \tag{5.55}$$

5.7.1. Нерегенеративные ретрансляторы

Анализ канала связи для нерегенеративного ретранслятора — это анализ полного "оборота" сигнала (передача на спутник и ретрансляция на наземное оконечное устройство). Нерегенеративный ретранслятор имеет несколько уникальных особенностей — это зависимость общего отношения SNR от SNR канала "земля-спутник" и совместное использование мощности канала "спутник-земля" каждым сигналом и шумом канала "земля-

спутник". С этого момента при обращении к ретранслятору или транспондеру будем подразумевать *нерегенеративный ретранслятор*, и для простоты будем предполагать, что транспондер работает в собственном линейном диапазоне.

Возможности спутникового транспондера ограничены мощностью канала "спутник-земля", мощностью наземного оконечного устройства, которая подается в канал "земля-спутник", шумом спутника и наземной оконечной станции, а также шириной полосы канала. Как правило, основные ограничения накладывает один из этих параметров; довольно часто — это мощность канала "спутник-земля" или ширина полосы канала. Важнейшие параметры линейного спутникового канала связи показаны на рис. 5.24. Ретранслятор передаст все сигналы канала "земля-спутник" (или шум, при отсутствии сигнала) без какой-либо обработки, за исключением усиления и трансляции по частоте. Предположим, что в пределах полосы приемника W существуют множественные каналы "земля-спутник" (используемые одновременно) и их разделение производится с помощью метода, известного как *множественный доступ с частотным разделением* (frequency-division multiple access — FDMA). Технология FDMA — это метод совместного использования ресурсов связи посредством распределения между пользователями раздельных участков полосы транспондера; подробно технология FDMA рассмотрена в главе 11. Эффективная мощность канала "спутник-земля" $EIRP_s$ является константой, и поскольку мы предполагаем использование линейного транспондера, $EIRP_s$ разделена между множественными сигналами (и шумами) канала "земля-спутник" пропорционально соответствующим уровням входного напряжения.

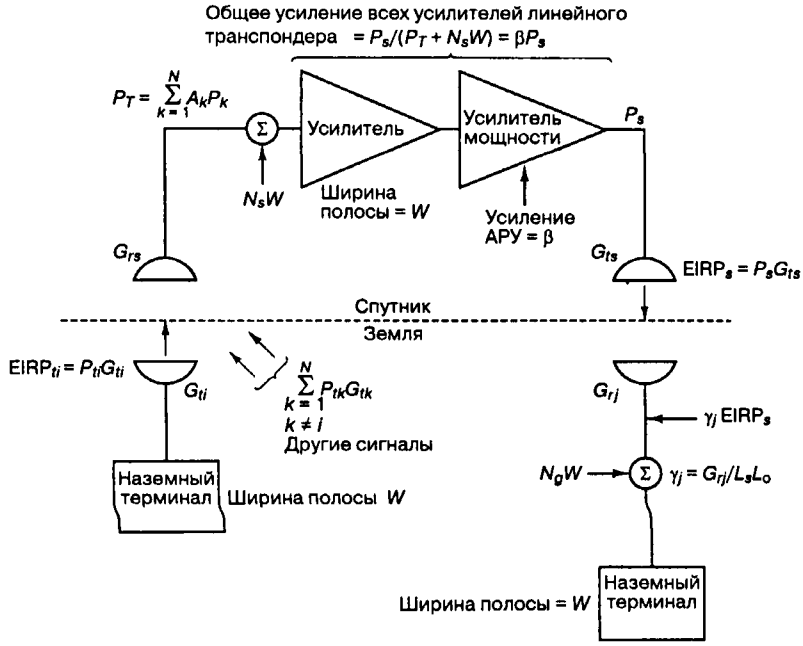


Рис. 5.24. Нерегенеративный спутниковый ретранслятор

Передача начинается с наземной станции (ширина полосы $\leq W$), скажем терминала i, причем EIRP терминала $EIRP_{ti} = P_{ti}G_{ti}$. Одновременно на спутник передаются другие сигналы (с других терминалов). Мощность EIRP с k-го терминала будем далее обозна-

чать просто P_k. На спутнике мощность общего принятого сигнала равна $P_T = \sum A_k P$, где A_k описывает потери распространения в канале "земля-спутник" и усиление спутниковой антенны для k канала. N_sW — это мощность шума в канале "земля-спутник", а N_s — общая спектральная плотность мощности шума, возникающего в спутниковом приемнике и излучающей спутниковой антенне. Общую мощность EIRP канала "спутник-земля" $\text{EIRP}_s = P_sG_{ts}$, где P_s — мощность на выходе спутникового транспондера, а G_{ts} — коэффициент усиления передающей антенны спутника, можно выразить следующим образом [14].

$$\text{EIRP}_s = \text{EIRP}_s\beta[A_iP_i + (P_T - A_iP_i) + N_sW] \tag{5.56}$$

Обе части формулы (5.56) выражают общую мощность EIRP спутника. Выражение $\beta[A_iP_i + (P_T - A_iP_i) + N_sW]$ в правой части является дробным пропорциональным распределением EIRP_s между различными пользователями и шумом канала, так что суммарное значение этого выражения равно 1. Полезность приведенного равенства вскоре станет очевидной. Общее усиление мощности в транспондере можно выразить как βP_s. Поскольку P_s фиксированы, а входящие сигналы могут быть различными, $\beta = 1/(P_T + N_sW)$ — это значение коэффициента автоматической регулировки усиления (automatic gain control — AGC). Общую мощность сигнала, принятого из канала "земля-спутник", P_T, можно записать как $A_iP_i + (P_T - A_iP_i)$, разделив, таким образом, мощность i-го сигнала и мощность остальных сигналов в транспондере. Общую мощность, принятую j-м наземным терминалом с шириной полосы W, можно записать следующим образом.

$$P_{rj} = \text{EIRP}_s\gamma_j\beta[A_iP_i + (P_T - A_iP_i) + N_sW] + N_gW \tag{5.57}$$

Здесь $\gamma_i = G_r/L_sL_o$ учитывает потери в канале "спутник-земля" и усиление принимающей антенны для j-го наземного терминала. $\text{EIRP}_s\gamma_j$ представляет часть мощности EIRP_s, принятой j-м наземным терминалом, а N_g — это спектральная плотность мощности шума, созданного и внесенного оборудованием приемной станции. Уравнение (5.57) описывает саму суть пропорционального разделения в ретрансляторе мощности канала "спутник-земля" между различными пользователями и шумом. Перепишем уравнение (5.57), заменив β его эквивалентом $1/(P_T + N_sW)$.

$$P_{rj} = EIRP_s\gamma_j\left(\frac{A_iP_i}{P_T + N_sW} + \frac{P_T - A_iP_i}{P_T + N_sW} + \frac{N_sW}{P_T + N_sW}\right) + N_gW \tag{5.58}$$

Для облегчения дальнейших рассуждений запишем уравнение (5.58) словами.

$$P_{rj} = EIRP_s\gamma_j\left(\frac{\text{мощность } S_i\ (UL)}{\text{общая мощность } (S+N)\ (UL)} + \frac{\text{равновесная мощность } (UL)}{\text{общая мощность } (S+N)\ (UL)} + \right.$$
$$\left. + \frac{\text{мощность шума } (UL)}{\text{общая мощность } (S+N)\ (UL)}\right) + N_gW$$

Здесь S — мощность сигнала, N — мощность шума, а (UL) — канал "земля-спутник" (uplink).

Можно ли из уравнения (5.58) определить важную связь, которая должна существовать между пользователями, совместно использующими нерегенеративный транс-

пондер? Пользователи должны *взаимодействовать*, не превышая договорные уровни мощности передачи. Из уравнения (5.58) видно, что часть мощности EIRP канала "спутник-земля", выделенной определенному пользователю (или относящейся к шуму канала), определяется отношением мощности этого пользователя к общей мощности суммарного сигнала плюс мощность шума. Следовательно, если один из пользователей, совместно использующих канал, решит "смошенничать" путем увеличения мощности своего сигнала, результатом будет улучшение уровня сигнала этого пользователя за счет сигналов других пользователей. Заметим также из уравнения (5.58), что шум канала "земля-спутник" использует ресурс канала "спутник-земля" наравне с другими пользователями. Такое включение шума канала "земля-спутник" в канал "спутник-земля" является отличительной особенностью нерегенеративных ретрансляторов.

Из уравнения (5.58) отношение P_r/N сигнала, переданного i-м передатчиком и принятого j-м терминалом, равно следующему.

$$\left(\frac{P_r}{N}\right)_{ij} \approx \frac{\text{EIRP}_s \gamma_j [A_i P_i/(P_T + N_s W)]}{\text{EIRP}_s \gamma_j [N_s W/(P_T + N_s W)] + N_g W} \tag{5.59}$$

Общее отношение P_r/N_0 сигнала, переданного i-м передатчиком и принятого j-м терминалом, равно следующему [14].

$$\left(\frac{P_r}{N_0}\right)_{ij} \approx \frac{\text{EIRP}_s \gamma_j \beta A_i P_i}{\text{EIRP}_s \gamma_j \beta N_s + N_g} \tag{5.60}$$

Уравнения (5.58)–(5.60) показывают, что шум ретранслятора уменьшает общее значение параметра SNR двумя способами — он "крадет" мощность EIRP канала "спутник-земля" и вносит вклад в общий шум системы. Если спутниковый шум канала "земля-спутник" доминирует, т.е. при $P_T << N_s W$, говорят, что *передача ограничена каналом "спутник-земля"*, и большая часть мощности EIRP_s канала "спутник-земля" бесполезно выделяется мощности шума канала "земля-спутник". В этом случае и если $\text{EIRP}_s \gamma_j >> N_g W$, уравнение (5.60) можно переписать следующим образом.

$$\left(\frac{P_r}{N_0}\right)_{ij} \approx \frac{\text{EIRP}_s \gamma_j A_i P_i / N_s W}{(\text{EIRP}_s \gamma_j / W) + N_g} \approx \frac{A_i P_i}{N_s} \tag{5.61}$$

Уравнение (5.61) показывает, что при передаче, ограниченной каналом "земля-спутник", общее отношение P_r/N_0 практически совпадает с SNR канала "земля-спутник". Более распространенной является *передача, ограниченная каналом "спутник-земля"*, когда $P_T >> N_s W$ и мощность EIRP спутника ограничена. В этом случае уравнение (5.60) можно переписать следующим образом.

$$\left(\frac{P_r}{N_0}\right)_{ij} \approx \frac{\text{EIRP}_s \gamma_j A_i P_i / P_T}{N_g} \tag{5.62}$$

Затем мощность транспондера распределяется между различными сигналами канала "земля-спутник"; небольшой шум канала "земля-спутник" передается по каналу "спутник-земля". Производительность ретранслятора в этом случае ограничена параметрами канала "спутник-земля".

Пример анализа канала связи для нерегенеративного ретранслятора ("полный оборот") приведен в табл. 5.3. Часть "земля-спутник" сама по себе не завершает бюджета канала, поскольку передача не демодулируется на спутнике. Без демодуляции *битов не существует*, а следовательно, не существует возможности измерения вероятности появления битовой ошибки. После полного оборота сигнал демодулируется на наземном терминале; и только после этого определяется окончательный резерв канала связи. Пример, приведенный в табл. 5.3, представляет одновременное обслуживание спутниковым транспондером 10 пользователей (частота канала "земля-спутник" — 375 МГц, частота канала "спутник-земля" — 275 МГц, расстояние — 22 000 морских миль или 40 779 км). В блоке "А" показано отношение $P_i/(P_T + N_sW)$, описывающее пропорциональное разделение мощности EIRP канала "спутник-земля" для интересующего нас сигнала. В данном примере, где все пользователи осуществляют передачу с равными уровнями мощности, каждому сигналу выделяется 9,8% EIRP канала "спутник-земля". В блоке "В" мы видим пропорциональное разделение EIRP канала "спутник-земля". Общая мощность равна 1514,7 Вт; интересующий нас пользователь получает 148,5 Вт; другие пользователи получают в сумме 1336,1; шум канала "земля-спутник" получает мощность 30,1 Вт.

Таблица 5.3. Бюджет канала связи для нерегенеративного спутникового ретранслятора с 10 пользователями

		Канал "земля-спутник"		Канал "спутник-земля"
Переданная мощность (дБВт)		27,0 (500,0 Вт)		13,0 (20,0 Вт)
Потери в передатчике (дБ)		1,0		1,0
Усиление антенны передатчика (максимум дБ[i])		19,0		19,8
Диаметр параболической антенны (футы)	10,00		15,00	
Ширина луча половинной мощности (градусы)	19,16		17,42	
EIRP (дБВт)		45,0		31,8 (1514,7 Вт)
Потери в тракте		176,1		173,4
Мощность переданного сигнала (дБВт)			**В**	21,7 (148,5 Вт)
Мощность других переданных сигналов (дБВт)				31,3 (1336,1 Вт)
Мощность шума, переданного по каналу "земля-спутник" (дБВт)				14,8 (30,1 Вт)
Другие потери (дБ)		2,0		2,0
Изотропная мощность принятого сигнала (дБВт)		−133,1		−153,7
Изотропная мощность принятого шума (дБВт)				−160,6
Усиление антенны приемника (максимум дБ[i])		22,5		16,3
Диаметр параболической антенны (футы)	15,00		10,00	

Окончание табл. 5.3

Ширина луча половинной мощности (градусы)	12,77			26,13		
Мощность принятого сигнала (дБВт)		−110,6			−137,4	
Мощность принятого шума (дБВт)					−144,3	
Температура антенны приемника (дБК)		24,6	(290 K)		20,0	(100 K)
Шум-фактор приемника в порте антенны (дБ)		10,8			2,0	
Температура приемника (дБК)		35,1	(3197 K)		22,3	(170 K)
Температура системы (дБК)		35,4	(3487 K)		24,3	(270 K)
$G/T°$ системы (дБ/К)		−12,9			−8,0	
Константа Больцмана (дБВт/КГц)		−228,6			−228,6	
Спектральная плотность шума (дБВт/Гц)		−193,2			−204,3	
Ширина полосы системы (дБГц)		75,6	(36,0 МГц)		75,6	(36,0 МГц)
Мощность шума (дБВт)		−117,6			−128,7	
Мощность шума канала "земля-спутник" + мощность шума канала "спутник-земля" (дБВт)					−128,6	
Одновременный доступ	10					
Мощность других принятых сигналов (дБВт)		−101,1				
Другие сигналы + шум (дБВт)		−101,0				
$P_r/(P_T + N_r/W)$ (дБ)	A	−10,1	(0,098)			
P_r/N (дБ)		7,0			−8,7	
Общее P_r/N (дБ)					−8,8	
P_r/N_0 (дБГц)		82,6			66,9	
Общее P_r/N_0 (дБГц)					66,8	
Скорость передачи данных (дБбит/с)					50,0	(100 000 бит/с)
Доступное E_b/N_0 (дБ)					16,8	
Требуемое E_b/N_0 (дБ)					10,0	
Резерв					6,8	

Оценить производительность, описанную в уравнении (5.60), можно, использовав значения E_b/N_0 (или P_r/N_0) каналов "земля-спутник" и "спутник-земля", объединенные следующим образом (*при отсутствии комбинационных помех*) [15].

$$\left(\frac{E_b}{N_0}\right)^{-1}_{\text{общ}} = \left(\frac{E_b}{N_0}\right)^{-1}_{u} + \left(\frac{E_b}{N_0}\right)^{-1}_{d} \tag{5.63}$$

Здесь индексы *общ*, *u* и *d* обозначают, соответственно, общее значение E_b/N_0, а также значения в канале "земля-спутник" (uplink) и "спутник-земля" (downlink).

Большинство коммерческих спутниковых транспондеров являются нерегенеративными. Однако очевидно, что в будущем коммерческие системы будут требовать встроенной обработки, коммутации или выборочной адресации сообщений и будут использовать регенеративную ретрансляцию для преобразования принятых сигналов в биты сообщений. Помимо возможности внедрения сложной обработки данных, одной из важных особенностей регенеративных ретрансляторов, по сравнению с нерегенеративными, является то, что каналы "земля-спутник" и "спутник-земля" разделяются, так что шум из первого не переходит во второй. Использование регенеративных спутниковых ретрансляторов позволяет значительно улучшить значения E_b/N_0, которые необходимы в обоих каналах, относительно значений, требуемых современными нерегенеративными ретрансляторами. В канале "земля-спутник" наблюдалось [16] увеличение E_b/N_0 порядка 5 дБ, а в канале "спутник-земля" — 6,8 дБ (использовалась когерентная модуляция QPSK с $P_B = 10^{-4}$).

5.7.2. Нелинейное усиление ретрансляторов

В большинстве спутниковых систем связи мощность существенно ограничена, и неэффективность, связанную с каскадами линейного усиления мощности, преодолевать обычно дорого. По этим причинам многие спутниковые ретрансляторы используют нелинейные усилители мощности. Эффективное усиление мощности получается за счет искажения сигнала, вызванного нелинейностью. Рассмотрим основные недостатки нелинейностей усилителей.

1. Комбинационные помехи (intermodulation (IM) noise), вызванные взаимодействием различных несущих. Вред является двояким; полезная мощность может теряться, переходя в энергию комбинационных помех (потери обычно составляют 1–2 дБ), и в виде интерференции в канал могут вноситься паразитные комбинационные произведения. Последняя проблема может быть достаточно серьезной.
2. Преобразования амплитудной модуляции в амплитудную модуляцию (AM-to-AM conversion) — это явление, обычное для нелинейных устройств, подобных лампам бегущей волны. На входе устройства любые флуктуации огибающей сигнала (амплитудная модуляция) подвергаются нелинейному преобразованию и приводят к искажению амплитуды на выходе устройства. Следовательно, работа лампы бегущей волны в нелинейной области не будет оптимальным выбором усиления мощности для схемы, основанной на модулировании амплитуды (такой, как QAM).
3. Переход амплитудной модуляции в фазовую (AM-to-PM conversion) — это еще одно явление, общее для нелинейных устройств. Флуктуации в огибающей сигнала производят колебания фазы, которые могут повлиять на достоверность передачи при использовании любой схемы, основанной на модулировании фазы (такой, как PSK или DPSK).
4. В ограничителях с резким порогом, ослабление слабых сигналов относительно сильных составляет порядка 6 дБ [2]. В лампах бегущей волны, работающих в режиме насыщения, подавление слабых сигналов происходит вследствие не только ограничения, но и того, что механизмы связывания сигнала в лампе оптимизированы в пользу сильных сигналов. В результате слабые сигналы могут ослабляться на 18 дБ [17].

Общепринятые нерегенеративные ретрансляторы обычно работают *с режекцией* из области высокого насыщения; это делается, чтобы избежать заметных комбинационных помех, и, следовательно, позволяет эффективно использовать всю полосу системы. Впрочем, режекция в линейную область — это компромисс; для получения полезного уровня выходной мощности некоторый уровень комбинационных помех должен быть приемлемым.

5.8. Системные компромиссы

Пример бюджета канала связи, приведенного в табл. 5.3, — это документ распределения ресурсов. Подобное табулирование канала связи позволяет исследовать потенциальные компромиссные проекты системы и оптимизировать производительность системы. Бюджет канала — это естественная начальная точка для рассмотрения всех потенциальных компромиссов: резерв или шум-фактор, размер антенны или мощность передатчика и т.д. В табл. 5.4 приведен пример расчетов для изучения возможных компромиссов между мощностью наземной передающей станции и шум-фактором в принимающем оконечном устройстве. Первая строка таблицы взята из бюджета канала, приведенного в табл. 5.3. Допустим, что вследствие некоторых физических ограничений на передающем наземном терминале системный инженер решил, что передатчик мощностью 500 Вт является непрактичным или что подобный передатчик дает системе излишне богатый канал "земля-спутник" (система плохо спроектирована). После этого инженер должен рассмотреть компромиссы между значением мощности передатчика и резервом мощности, учитывающим тепловой шум. Расчет потенциальных компромиссов является тривиальной задачей для компьютера. Табл. 5.4 была создана путем многократного повторения вычисления бюджета канала, причем при каждом следующем повторении P_t уменьшалось вдвое.

Таблица 5.4. Возможные компромиссы: P_t или энергетический резерв

$P_t(W)$	$(P_t/N_0)_u$ (дБГц)	$(P_t/N_0)_d$ (дБГц)	$(P_t/N_0)_{общ}$ (дБГц)	Резерв (дБ)
500,0	82,6	66,9	66,8	6,8
250,0	79,6	66,8	66,6	6,6
125,0	76,6	66,6	66,2	6,2
62,5	73,6	66,3	65,5	5,5
31,3	70,5	65,7	64,5	4,5
15,6	67,5	64,8	62,9	2,9
7,8	64,5	63,3	60,8	0,8
3,9	61,5	61,4	58,4	–1,6
2,0	58,4	59,0	55,7	–4,3
1,0	55,4	56,4	52,9	–7,2
0,5	52,4	53,6	49,9	–10,1

Каждое значение мощности передатчика (с шагом 3 дБ) — это выбор передатчиков, каналов "земля-спутник" и "спутник-земля" и энергетического резерва. Системный инженер должен всего лишь внимательно рассмотреть перечень, чтобы найти вероятного кандидата. Например, если инженера удовлетворяет резерв порядка 3–4 дБ,

он может снизить мощность передатчика с 500 Вт до 20 или 30 Вт. Он может также пожелать создать передатчик с мощностью, скажем, $P_t = 100$ Вт с дальнейшим использованием дополнительных компромиссов (возможно, руководствуясь опасениями относительно одной из подсистем, скажем антенны). Затем инженер создаст новую таблицу при фиксированном $P_t = 100$ Вт и снова выполнит последовательное вычисление бюджетов канала для создания подобного перечня других возможных компромиссов.

Заметим, что из табл. 5.4 можно определить обсуждавшиеся ранее области передач, ограниченных каналами "земля-спутник" и "спутник-земля". В первых строках, где отношение SNR в канале "земля-спутник" велико, уменьшение SNR этого канала на 3 дБ приводит к потере общего SNR всего на несколько десятых децибела. Системы с подобными характеристиками *ограничены каналом "спутник-земля"*; т.е. ограничения на производительность систем накладывают в основном параметры канала "спутник-земля" и система слабо реагирует на изменения параметров канала "земля-спутник". В нижних строках таблицы изменение отношения SNR в канале "земля-спутник" на 3 дБ меняет общее отношение SNR практически на те же 3 дБ. Здесь мы имеем дело с системами, *ограниченными каналом "земля-спутник"*, т.е. основные ограничения на производительность системы определяют параметры канала "земля-спутник".

5.9. Резюме

Среди множества анализов, поддерживающих разработку систем связи, бюджет канала связи занимает особое место, поскольку он позволяет охватить систему в целом. Изучая бюджет канала, можно узнать много полезного относительно структуры и производительности всей системы. Например, из энергетического резерва канала связи можно получить информацию о том, как система соответствует поставленным требованиям — с запасом, впритык или вообще не соответствует. Очевидными становятся все аппаратные ограничения и возможности их компенсации за счет других частей канала связи. Бюджет канала часто используется для рассмотрений компромиссных проектов системы и изменений конфигурации; кроме того, он способствует пониманию нюансов на уровне подсистем и взаимозависимости элементов системы. Объединенный с другими методами моделирования, бюджет канала может помочь предсказать вес, размер и стоимость системы. В данной главе показано, как формулируется этот бюджет и как можно его использовать для определения компромиссов. Стоит также отметить, что бюджет канала — это один из самых важных документов системного администратора; он представляет "итоговый подсчет", результат поиска системы с оптимальной достоверностью передачи.

Литература

1. Panter P. F. *Communication Systems Design: Line-of-Sight and Tropo-Scatter Systems.* R. E. Krieger Publishing Co, Inc., Melbourne, Fla., 1982.
2. Jones J. J. *Hard Limiting of Two Signals in Random Noise.* IEEE Trans. Inf. Theory, vol. IT9, January, 1963.
3. Silver S. *Microwave Antenna Theory and Design.* MIT Radiation Laboratory Series, Vol. 12, McGraw-Hill Book Company, New York, 1949.
4. Kraus J. D. *Antennas.* McGraw-Hill Book Company, New York, 1950.
5. Johnson J. B. *Thermal Agitation of Electricity in Conductors.* Phys. Rev., vol. 32, July, 1928, pp. 97–109.
6. Nyquist H. *Thermal Agitation of Electric Charge in Conductors.* Phys. Rev., vol. 32, July, 1928, pp. 110–113.

7. Desoer C. A. and Kuh E. S. *Basic Circuit Theory.* McGraw-Hill Book Company, New York, 1969.
8. Schwab L. M. *World-Wide Link Availability for Geostationary and Critically Inclined Orbits Including Rain Attenuation Effects.* Lincoln Laboratory, Rep. DCA-9, January, 27, 1981.
9. Friis H. T. *Noise Figure of Radio Receivers.* Proc. IRE, July, 1994, pp. 419–422.
10. IRE Subcommittee 7.9 on Noise. *Description of the Noise Performance of Amplifiers and Receiving Systems.* Proc. IEEE, March, 1963, pp. 436–442.
11. Blackwell L. A. and Kotzebue K. L. *Semiconductor Diode Parametric Amplifiers.* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1961.
12. Ko H. C. and Kraus J. D. *A Radio Map of the Sky at 1.2 Meters.* Sky Telesc., vol. 16, Februry, 1957, pp. 160, 161.
13. Piddington J. H. and Trent G. H. *A Survey of Cosmic Radio Emission at 600 Mc/s.* Aust. J. Phys., vol. 9, December, 1956, pp. 481–493.
14. Spilker J. J. *Digital Communications by Satellite.* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1977.
15. Pritchard W. L. and Sciulli J. A. *Satellite Communication Systems Engineering.* Prentice-Hall, Inc., Englewood Cliffs., N. J., 1986.
16. Campanella S. J., Assal F. and Berman A. *Onboard Regenerative Repeaters.* Int. Conf. Commun., Chicago, vol. 1., 1977, pp. 6.2-121–66.2-125.
17. Wolkstein H. J. *Suppression and Limiting of Undesired Signals in Travelling-Wave-Tube Amplifiers.* Publication ST-1583, RCA Rev., vol. 22, no. 2, June, 1961, pp. 280–291.

Задачи

5.1.

а) Чему (в децибелах) равно значение потерь в свободном пространстве для несущей частоты 100 МГц и расстояния 3 мили?

б) Выходная мощность передатчика равна 10 Вт. Пусть передающая и принимающая антенны являются изотропными, а другие потери отсутствуют. Вычислите принятую мощность в дБВт.

в) В п. б положим EIRP = 20 Вт. Чему равна принятая мощность в дБВт?

г) На сколько (в дБ) увеличится усиление параболической антенны при удвоении ее диаметра?

д) Чему должен быть равен диаметр параболической антенны, чтобы в системе, описанной в п. а, усиление антенны было равно 10 дБ? Эффективность антенны предполагать равной 0,55.

5.2. На выход передатчика подается 2 Вт на несущей частоте 2 ГГц. Пусть передающая и принимающая антенны являются параболическими с диаметром 3 фута каждая. Эффективность каждой антенны считать равной 0,55.

а) Вычислите усиление каждой антенны.

б) Вычислите EIRP переданного сигнала в дБВт.

в) Если антенны разделены расстоянием 25 миль, приходящимся на свободное пространство, чему (в дБВт) будет равна доступная мощность сигнала вне принимающей антенны?

5.3. В табл. 5.1 было приведено предложение от Satellite Television Corporation, предназначенное для спутника непосредственного вещания с EIRP = 57 дБВт и частотой передачи в канале "спутник-земля" 12,5 ГГц. Допустим, единственными потерями являются показанные потери в канале "спутник-земля". Предположим, информация, подаваемая в этот канал, состоит из цифрового сигнала (5×10^7 бит/с). Пусть требуемое отношение E_b/N_0 равно 10 дБ, температура системы в вашем домашнем приемнике — 600 К, а эффективность принимающей параболической антенны — 0,55. Чему равен минимальный диаметр антенны, с помощью которого можно закрыть канал? Как вы думаете, будут ли возражать соседи против такой "тарелки"?

5.4. Входное и выходное сопротивление усилителя равно 50 Ом, усиление — 60 дБ, а ширина полосы — 10 кГц. Если со входом соединяется сопротивление 50 Ом с температурой 290 K, среднеквадратическое значение мощности шума на выходе равно 10 мкВ. Определите эффективную шумовую температуру усилителя.

5.5. Шум-фактор усилителя равен 4 дБ, ширина полосы — 500 кГц, а входное сопротивление — 50 Ом. Вычислите напряжение входного сигнала, необходимое для получения на выходе SNR = 1 при присоединении усилителя к источнику сигнала с сопротивлением 50 Ом при температуре 290 K.

5.6. Рассмотрим систему связи, имеющую следующую спецификацию: частота передачи — 3 ГГц, схема модуляции — BPSK, вероятность появления ошибочного бита — 10^{-3}, скорость передачи данных — 100 бит/с, энергетический резерв линии — 3 дБ, EIRP — 100 Вт, усиление принимающей антенны — 10 дБ, расстояние между передатчиком и приемником — 40 000 км. Потерями в линии между принимающей антенной и приемником можно пренебречь.

а) Вычислите максимальную допустимую спектральную плотность мощности шума (в Вт/Гц) относительно входа приемника.

б) Чему равна максимально допустимая эффективная шумовая температура (в K) для приемника, если температура антенны равна 290 K?

в) Чему (в дБ) равен максимальный допустимый шум-фактор для приемника?

5.7. Шум-фактор предварительного усилителя приемника равен 13 дБ, усиление равно 60 дБ, а ширина полосы — 2 МГц. Температура антенны — 490 K, мощность входного сигнала — 10^{-12} Вт.

а) Найдите эффективную температуру (в K) предварительного усилителя.

б) Найдите температуру системы (в K).

в) Найдите выходное SNR (в дБ).

5.8. Дан приемник со следующими параметрами: усиление — 50 дБ, шум-фактор — 10 дБ, ширина полосы — 500 МГц, мощность входного сигнала — 50×10^{-12} Вт, температура источника $T_A°$ — 10 K, потери в линии — 0 дБ. Между антенной и приемником нужно ввести предварительный усилитель, который должен иметь усиление 20 дБ и ширину полосы 500 МГц. Найдите шум-фактор предварительного усилителя, получаемый при улучшении общесистемного SNR на 10 дБ.

5.9. Найдите максимально допустимую эффективную температуру системы $T_S°$, необходимую для *закрытия с минимальными требованиями* определенного канала с вероятностью битовой ошибки 10^{-5} при скорости передачи данных R = 10 Кбит/с. Канал имеет следующие параметры: частота передачи — 12 ГГц, EIRP — 10 дБВт, усиление принимающей антенны — 0 дБ, тип модуляции — кодировка BPSK с некогерентным обнаружением, другие потери — 0 дБ, расстояние между передатчиком и приемником — 100 км.

5.10. Рассмотрим приемник, сделанный из следующих трех каскадов: входной каскад — это предварительный усилитель с усилением 20 дБ и шум-фактором 6 дБ; второй каскад — сеть с потерями 3 дБ; выходной каскад — усилитель с усилением 60 дБ и шум-фактором 16 дБ.

а) Найдите общий шум-фактор всего приемника.

б) Повторите п. а для приемника без первого каскада.

5.11. а) Найдите эффективную шумовую температуру $T_R°$ приемника, состоящего из трех последовательно соединенных усиливающих каскадов с коэффициентами усиления 10, 16 и 20 дБ и эффективными шумовыми температурами 1800, 2700 и 4800 K.

б) Каким должно быть усиление первого каскада, чтобы вклад в $T_R°$ других каскадов снизился до 10% от вклада первого каскада?

5.12. Эффективная температура многокаскадного приемника должна быть равна 300 К. Пусть эффективная температура и коэффициенты усиления каскадов 2–4 равны, соответственно, $T_2° = 600$ К, $T_3° = T_4° = 2000$ К, $G_2 = 13$ дБ, $G_2 = G_4 = 20$ дБ.

а) Вычислите усиление G_1 первого каскада при $T_1° = 200$, 230, 265, 290, 295 и 300 К.

б) Изобразите компромиссные соотношения G_1 и $T_1°$.

в) Почему (относительно вклада в эффективную температуру приемника) можно пренебречь всеми каскадами по сравнению с четвертым?

г) Какая область компромиссов между $T_1°$ и G_1 (с практической инженерной точки зрения) заслуживает рассмотрения?

5.13. Приемник состоит из предварительного усилителя, за которым следуют множественные усиливающие каскады. Общая эффективная температура всех усиливающих каскадов равна 1000 К относительно выхода предварительного усилителя.

а) Вычислите эффективную шумовую температуру приемника относительно входа предварительного фильтра для однокаскадного предварительного усилителя с шумовой температурой 400 К и коэффициентами усиления 3, 6, 10, 16 и 20 дБ.

б) Повторите п. а для двухкаскадного предварительного усилителя с шумом 400 К на каскад и коэффициентами усиления 3, 6, 10 и 13 дБ на каскад.

в) Изобразите зависимость эффективной температуры приемника от коэффициента усиления первого каскада для пп. а и б.

5.14. а) В уравнении (5.42) показан шум-фактор сети, состоящей из линии с потерями, за которой следует усилитель. Выведите выражение для шум-фактора последовательного соединения трех таких сетей.

б) Рассмотрим сеть, составленную из усилителя, за которым следует линия с потерями. Выведите общее выражение для шум-фактора последовательного соединения трех таких сетей.

в) Приемник составлен из последовательного соединения принимающей антенны с температурой $T_A = 1160$ К, линии с потерями 1 с $L_1 = 6$ дБ, усилителя 1 с шум-фактором $F_1 = 3$ дБ и усилением 13 дБ, линии с потерями 2 с $L_2 = 10$ дБ и усилителя 2 с шум-фактором $F_2 = 6$ дБ и коэффициентом усиления $G_2 = 10$ дБ. Мощность входного сигнала равна 80 пиковатт (пВт), а ширина полосы сигнала — 0,25 ГГц. Определите мощность сигнала, шум-фактор и SNR во всех точках системы.

5.15. а) Усилитель с коэффициентом усиления 10 дБ и шум-фактором 3 дБ соединен непосредственно с выходом принимающей антенны (без линии с потерями между ними). За усилителем следует линия с коэффициентом потерь 10 дБ. Пусть мощность входного сигнала равна 10 пВт, температура антенны — 290 К, а ширина полосы сигнала — 0,25 ГГц. Найдите SNR в усилителе, на его выходе и вне линии с потерями.

б) Повторите п. а для антенны с температурой 1450 К.

5.16. Приемник с коэффициентом усиления 80 дБ и эффективной шумовой температурой 3000 К соединяется с антенной, шумовая температура которой равна 600 К.

а) Определите номинальную мощность шума, поступающего от источника в полосу 40 МГц.

б) Найдите мощность шума приемника относительно входа приемника.

в) Найдите мощность выходного шума приемника в полосе 40 МГц.

5.17. Антенна ориентирована так, что ее шумовая температура равна 50 К. Она соединена с предварительным усилителем, шум-фактором 2 дБ и номинальным усилением 30 дБ в эффективной полосе 20 МГц. Мощность сигнала на входе предварительного усилителя равна 10^{-12} Вт.

а) Определите эффективную шумовую температуру предварительного усилителя.

б) Найдите SNR вне предварительного усилителя.

5.18. Приемник с шум-фактором 13 дБ соединен с антенной с помощью 75 футов линии передачи, имеющей сопротивление 300 Ом и потери 3 дБ на 100 футов.

а) Вычислите общий шум-фактор линии и приемника.

б) Между линией и приемником внесен предварительный усилитель (усиление — 20 дБ, шум-фактор — 3 дБ). Определите общий шум-фактор линии, предварительного усилителя и приемника.

в) Вычислите общий шум-фактор, если предварительный усилитель вставлен между антенной и линией передачи.

5.19. Система спутниковой связи использует передатчик, дающий 20 Вт мощности на несущей частоте 8 ГГц, которая подается на параболическую антенну диаметром 2 фута. Расстояние к принимающей наземной станции равно 20 000 морских миль (37 072 км). Принимающая система использует 8-футовую параболическую антенну, а ее шумовая температура равна 100 К. Пусть эффективность антенны равна 0,55. Случайные потери равны 2 дБ.

а) Вычислите максимальную скорость передачи данных, если используется дифференциальная когерентная модуляция PSK (DPSK), а вероятность битовой ошибки не превышает 10^{-5}.

б) Повторите п. а, предполагая, что передача на наземную станцию ведется на несущей 2 ГГц.

5.20. Пусть автоматический космический аппарат с несущей 2 ГГц и транспондером 10 Вт работает в непосредственной близости от Сатурна (расстояние $7{,}9 \times 10^8$ миль от Земли). Размер антенны принимающей наземной станции равен 75 футов, шумовая температура системы — 20 К. Вычислите граничный размер антенны космического аппарата, необходимой для закрытия канала со скоростью передачи 100 бит/с. Пусть требуемое отношение E_b/N_0 — 10 дБ, а случайные потери не превышают 3 дБ. Эффективность каждой антенны считать равной 0,55.

5.21. а) Имеем входной каскад приемника со следующими параметрами: усиление — 60 дБ, ширина полосы — 500 МГц, шум-фактор — 6 дБ, мощность входного сигнала = $6{,}4 \times 10^{-11}$ Вт, температура источника, $T_A°$ — 290 К, потери в линии — 0 дБ. Между антенной и приемником введен предварительный усилитель со следующим характеристиками: усиление — 10 дБ, шум-фактор — 1 дБ. Определите общий шум-фактор (в дБ). Каково при данной реализации было получено улучшение шум-фактора (в дБ)?

в) Повторите п. б для $T_A° = 6000$ К. Чему (в дБ) равно улучшение SNR на выходе?

г) Повторите п. б для $T_A° = 15$ К. Чему (в дБ) равно улучшение SNR на выходе?

д) Какой вывод можно сделать из ответов на предыдущие вопросы относительно влияния улучшения шум-фактора на улучшение параметра SNR на выходе? Ответ аргументируйте.

5.22. а) Используя данные параметры канала, найдите максимальный допустимый шум-фактор приемника. Применяется когерентная схема BPSK с вероятностью битовой ошибки 10^{-5} при скорости передачи данных 10 Мбит/с. Частота передачи — 12 ГГц, мощность EIRP — 0 дБВт, диаметр принимающей антенны — 0,1 м (эффективность предполагается равной 0,55), а температура антенны — 800 К. Расстояние между передатчиком и приемником равно 10 км. Резерв равен 0 дБ; также предполагается отсутствие непредвиденных потерь.

б) Если в условии п. а удвоить скорость передачи данных, то как это скажется на шум-факторе?

в) Если в условии п. а удвоить диаметр антенны, то как это скажется на шум-факторе?

5.23. а) Десять пользователей одновременно (используя схему FDMA) получают доступ к нерегенеративному спутниковому ретранслятору с шириной полосы 50 МГц. Пусть мощность EIRP каждого пользователя равна 10 дБВт, $A_i = G_r/L_sL_o = -140$ дБ. Чему равна общая мощность P_T, полученная приемником спутника?

б) Пусть шумовая температура спутника равна 2000 К. Чему равно значение шума спутника (в Вт) относительно входа приемника?

в) Чему равно отношение SNR в канале "земля-спутник" для каждого пользовательского сигнала?

г) Пусть спутник получает одинаковые мощности от всех пользователей. Чему равна доля EIRP спутника, выделяемая каждому из 10 пользовательских сигналов? Если мощность EIRP канала "спутник-земля" равна 1000 Вт, какая мощность (в Вт) на одного пользователя приходится в этом канале?

д) Какую мощность выделяет спутник для ретрансляции шума канала "земля-спутник"?

е) Каким каналом ограничена система? Ответ аргументируйте.

ж) На наземной станции шумовая температура приемника равна 800 К. Чему равно общее среднее отношение спектральных плотностей сигнала к шуму (P_r/N_0) для отдельной пользовательской передачи в полосе 50 МГц? Коэффициент $\gamma = G_r/L_sL_o$ считать равным −140 дБ.

з) Пересчитайте P_r/N_0, используя приближенный результат, полученный при ответе на п. е.

и) При отсутствии комбинационных помех часто используется следующее соотношение.

$$\text{общее}\left(\frac{P_r}{N_0}\right)^{-1} = \left(\frac{P_r}{N_0}\right)^{-1}\text{(земля-спутник)} + \left(\frac{P_r}{N_0}\right)^{-1}\text{(спутник-земля)}$$

Вычислите P_r/N_0 с помощью приведенного выражения и сравните результат с ответами на пп. ж и з.

5.24. Сколько пользователей могут одновременно получать доступ к нерегенеративному спутниковому ретранслятору, ширина полосы которого равна 100 МГц, так чтобы каждому пользователю доставалось 50 Вт из общей мощности спутника 5000 Вт? Эффективная системная температура на спутнике $T_S^\circ = 3500$ К. Пусть мощность EIRP в каждом пользовательском канале равна 10 дБВт, а коэффициент $\gamma = G_r/L_sL_o$ равен −140 дБ.

5.25. Канал с шумом AWGN имеет следующие параметры и требования: скорость передачи данных — 2,5 Мбит/с; модуляция — когерентная BPSK с идеальной синхронизацией частоты, несущей и случайного смещения фазы; требуемая вероятность ошибки — 10^{-5}; несущая частота — 300 МГц; расстояние между передатчиком и приемником 100 км. Мощность передатчика 10^{-3} Вт; диаметры передающей и принимающей антенны равны 2 м, их эффективность — 0,55; температура принимающей антенны — 290 К; потери в канале от выхода принимающей антенны до входа приемника — 1 дБ, иные потери отсутствуют. Найдите максимальный граничный шум-фактор приемника (в дБ), который может закрыть канал.

5.26. Ручной радиоприемник принимает и передает данные со скоростью 1 Мбит/с и вероятностью битовой ошибки 10^{-7}. Он должен работать на расстоянии 10 км на несущей 3 ГГц. В качестве схемы модуляции используется DPSK, а $G/T^\circ = -30$ дБ/К. Данное радио может использоваться в машинах и подвергаться потерям вследствие замирания. Разработчик радио желает исследовать компромиссы между минимизацией требуемой мощности EIRP и максимизацией сопротивления замиранию. Создайте таблицу, в которой будут показаны несколько пар значений "EIRP-потери вследствие замирания". Интересующие нас значения EIRP должны принадлежать диапазону 300 мВт–

10 Вт. Можно ли удовлетворить системные требования, если потери вследствие замирания равны 20 дБ, а EIRP меньше 10 Вт?

5.27. Разработчик решил, что радио, описанное в задаче 5.26, не обязательно должно удовлетворять поставленным требованиям при использовании его в машинах, поэтому потери вследствие замирания можно положить равными 0 дБ. Пусть в передатчике выбрана минимальная номинальная мощность EIRP, соответствующая потерям в 0 дБ (из решения задачи 5.26). Чему равно минимальное значение переданной мощности, которую можно использовать, если эффективная площадь поверхности антенны равна 25 $см^2$?

Вопросы для самопроверки

5.1. Почему потери в свободном пространстве — это функция длины волны (см. раздел 5.3.3)?

5.2. Как связаны отношение принятого сигнала к шуму (*S/N*) и отношение мощности несущей к шуму (*C/N*) (см. раздел 5.4)?

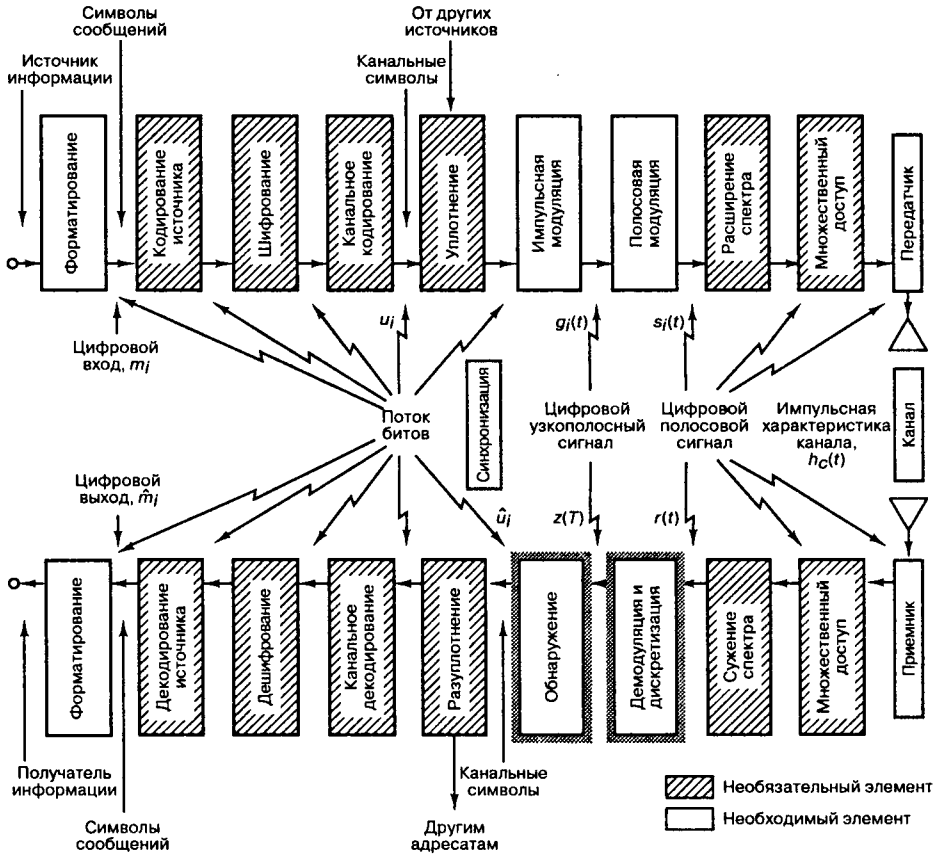
5.3. Какого резерва достаточно для работы канала (см. раздел 5.4.3)?

5.4. Существует два основных источника шума и интерференции на входе приемника. Назовите их (см. раздел 5.5.5).

5.5. Если мы желаем получить справедливое совместное использование *нерегенеративного* спутникового ретранслятора, то какая важная связь должна существовать между пользователями (см. раздел 5.7.1)?

ГЛАВА 6

Канальное кодирование: часть 1



Канальное кодирование (channel coding) представляет собой класс преобразований сигнала, выполняемых для повышения качества связи. В результате этого сигнал становится менее уязвим к таким эффектам ухудшения качества передачи, как шум, помехи и замирание. Канальное кодирование можно считать способом приведения параметров системы к желаемому компромиссу (т.е. соотношению между достоверностью передачи и шириной полосы пропускания или мощностью и шириной полосы пропускания). Как вы думаете, почему канальное кодирование так распространено? Это стало возможно благодаря использованию больших интегральных схем (БИС) и применению высокоскоростной цифровой обработки сигналов. Данный метод позволил более чем на 10 дБ повысить производительность при значительно меньших затратах по сравнению с другими методами, например методами увеличения мощности передатчика или размера антенны.

6.1. Кодирование сигнала и структурированные последовательности

Тему канального кодирования можно условно разделить на два раздела: кодирование (или обработка) сигнала и структурированные последовательности (или структурированная избыточность), как это показано на рис. 6.1. *Кодирование сигнала* означает преобразование сигнала в некий "улучшенный сигнал", позволяющий сделать процесс обнаружения менее подверженным ошибкам. Метод *структурированных последовательностей* — это преобразование последовательности данных в новую, "улучшенную последовательность", обладающую структурной избыточностью (которая вмещает избыточные биты). Эти избыточные разряды служат для определения и исправления ошибок. На выходе процедуры кодирования получается закодированный (формой сигнала или структурированной последовательностью) сигнал, имеющий лучшие пространственные характеристики, чем некодированный. Итак, сначала рассмотрим некоторые методы кодирования сигнала, а затем, начиная с раздела 6.3, обсудим суть структурированных последовательностей.

6.1.1. Антиподные и ортогональные сигналы

Антиподные и ортогональные сигналы уже обсуждались ранее, поэтому мы лишь напомним их основные особенности. В примере, приведенном на рис. 6.2, показано аналитическое представление набора синусоидальных антиподных сигналов ($s_1(t) = -s_2(t) = \sin \omega_0 t$, $0 \le t \le T$), а также его векторное и графическое представление. Какие существуют альтернативные определения антиподных сигналов? О таких сигналах можно сказать, что они либо являются зеркальными отображениями друг друга, либо один сигнал является отрицательным по отношению к другому, либо они различаются между собой на 180°.

Рис. 6.1. Основные преобразования цифровой связи

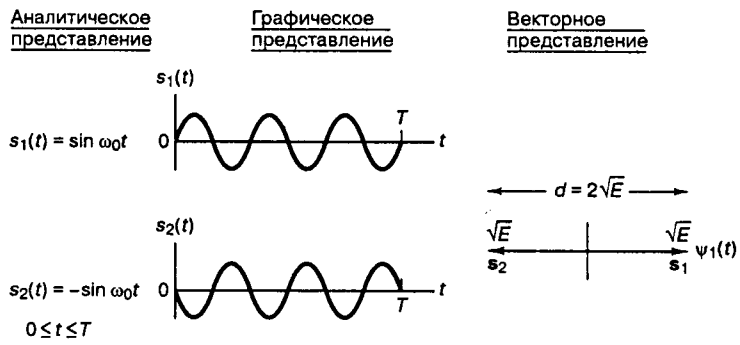


Рис. 6.2. Пример антиподного набора сигналов

В примере, приведенном на рис. 6.3, показан набор ортогональных сигналов, которые имеют вид импульсов, описывающихся следующими выражениями.

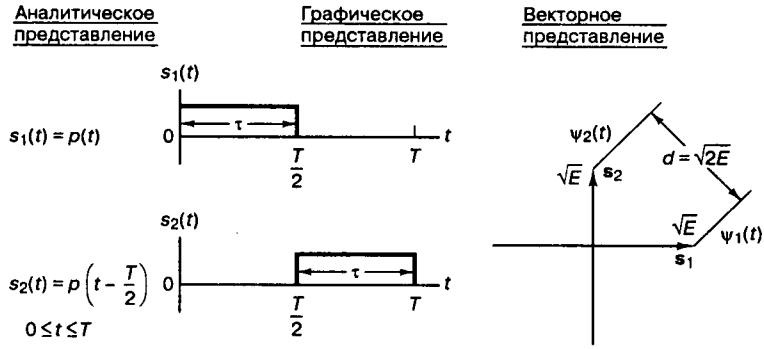


Рис. 6.3. Пример двоичного набора ортогональных сигналов

$$s_1(t) = p(t) \quad 0 \le t \le T$$

и

$$s_2(t) = p\left(t - \frac{T}{2}\right) \quad 0 \le t \le T$$

В данном случае $p(t)$ — импульс длительностью $\tau = T/2$, где T — период. В системах связи возможны и другие наборы ортогональных сигналов, например часто используемые $\sin x$ и $\cos x$. Любой набор равноэнергетических сигналов $s_i(t)$, $i = 1, 2, \ldots, M$, будет ортонормированным (ортогональным и нормированным на 1) тогда и только тогда, когда

$$z_{ij} = \frac{1}{E}\int_0^T s_i(t)s_j(t)dt = \begin{cases} 1 & \text{при } i = j \\ 0 & \text{при } i \ne j \end{cases}, \tag{6.1}$$

где z_{ij} является *коэффициентом взаимной корреляции* (cross-correlation coefficient), а величина E — энергией сигнала, выражаемой следующим образом.

$$E = \int_0 s_i^2(t)dt \tag{6.2}$$

Из графического представления на рис. 6.3 видно, что $s_1(t)$ и $s_2(t)$ не могут взаимодействовать, поскольку они разнесены во времени. Векторное представление показывает, что ортогональные сигналы перпендикулярны. Посмотрим на другие, альтернативные определения ортогональных сигналов или векторов. Можно сказать, например, что скалярное произведение двух разных векторов в ортогональном наборе должно быть равно нулю. В двух- и трехмерных декартовых системах координат векторы сигналов можно представить геометрически, как взаимно ортогональные друг к другу. Можно также сказать, что один вектор имеет нулевую проекцию на другой или один сигнал не может взаимодействовать с другим, поскольку они не принадлежат одному и тому же *пространству сигналов*.

6.1.2. *M*-арная передача сигналов

При M-арной передаче сигналов процессор за один такт работы принимает k бит данных. После этого он указывает модулятору произвести один из $M = 2^k$ сигналов; частным случаем $k = 1$ является двоичная передача сигнала. Для $k > 1$ M-арную передачу сигналов можно рассматривать как процедуру *кодирования формы сигнала*. При ортогональной передаче сигналов (например, сигналов MFSK) увеличение k приводит к повышению достоверности передачи или уменьшению требуемого E_b/N_0 за счет увеличения полосы пропускания; при неортогональной передаче сигналов (например, сигналов MPSK) улучшение эффективности использования полосы пропускания происходит за счет снижения достоверности передачи или возрастания требуемого E_b/N_0. Подходящий выбор формы сигнала позволяет найти компромисс между вероятностью ошибки, E_b/N_0 и эффективностью использования полосы пропускания. Более подробно такие компромиссы рассмотрены в главе 9.

6.1.3. Кодирование сигнала

Процедура кодирования сигнала состоит в преобразовании набора сигналов (представляющих набор сообщений) в усовершенствованный набор сигналов. Этот улучшенный набор можно использовать для получения более приемлемой величины P_B, соответствующей исходному набору. Наиболее популярные из таких *кодов сигнала* называются *ортогональными* (orthogonal) и *биортогональными кодами* (biorthogonal). В процессе кодирования каждый сигнал набора пытаются сделать настолько непохожим на другие, насколько это возможно, чтобы для всех пар сигналов коэффициент взаимной корреляции z_{ij} (см. уравнение 6.1) имел наименьшее возможное значение. Строго это условие выполняется тогда, когда сигналы антикоррелируют ($z_{ij} = -1$); этого можно добиться только в том случае, если в наборе всего два значения ($M = 2$) и они *антиподны* друг другу. Вообще, все коэффициенты взаимной корреляции можно сделать равными нулю [1]. В этом случае набор будет *ортогональным*. Наборы антиподных сигналов являются оптимальными в том смысле, что все сигналы максимально удалены друг от друга, как можно видеть на рис. 6.2. Расстояние d между векторами сигналов определяется как $d = 2\sqrt{E}$, где E — энергия сигнала на интервале T, как показано в уравнении (6.2). Сравнив пространственные характеристики ортогональных сигналов с характеристиками антиподных сигналов, приходим к выводу, что о первых можно сказать нечто вроде "довольно хорошо" (при данном уровне энергии сигнала). На

рис. 6.3 расстояние между векторами ортогональных сигналов составляет $d = \sqrt{2E}$.

Взаимная корреляция между двумя сигналами является мерой *расстояния* между двумя векторами сигналов. Чем меньше взаимная корреляция, тем больше векторы удалены друг от друга. Это можно проверить с помощью рис. 6.2, где антиподные сигналы (для которых $z_{ij} = -1$) представлены векторами, наиболее удаленными друг от друга, и рис. 6.3, где ортогональные сигналы (для которых $z_{ij} = 0$) представлены векторами, расположенными ближе друг к другу, чем антиподные векторы. Очевидно, что расстояние между идентичными сигналами ($z_{ij} = 1$) должно быть равно нулю.

Условие ортогональности в уравнении 6.1 записано через сигналы $s_i(t)$ и $s_j(t)$, где $i, j = 1, 2, \ldots, M$ (M — количество сигналов в наборе). Каждый сигнал набора $\{s_j(t)\}$ может содержать последовательность импульсов с уровнями +1 или −1, которые представляют двоичную 1 или 0. Если выразить набор в таком виде, уравнение (6.1) можно упростить, положив, что $\{s_j(t)\}$ состоит из ортогональных сигналов тогда и только тогда, когда

$$z_{ij} = \frac{\text{(количество совпавших цифр)} - \text{(количество несовпавших цифр)}}{\text{общее количество цифр в последовательности}} = \begin{cases} 1 & \text{для } i = j \\ 0 & \text{для } i \neq j \end{cases} \tag{6.3}$$

6.1.3.1. Ортогональные коды

Набор однобитовых данных можно преобразовать с помощью ортогональных кодовых слов, состоящих из двух разрядов каждое, которые описываются строками показанной ниже матрицы $\mathbf{H}_1$.

Набор данных	Набор ортогональных кодовых слов
0	$\mathbf{H}_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$
1	

(6.4,а)

В этом и следующих примерах проверка ортогональности набора кодовых слов производится с помощью уравнения (6.3). Для кодирования набора двухбитовых данных упомянутый выше набор следует расширить по горизонтали и вертикали, что дает матрицу $\mathbf{H}_2$.

Набор данных	Набор ортогональных кодовых слов
0 0	$\mathbf{H}_2 = \left[\begin{array}{cc:cc} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ \hdashline 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{array}\right] = \begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_1 \\ \mathbf{H}_1 & \overline{\mathbf{H}_1} \end{bmatrix}$
0 1	
1 0	
1 1	

(6.4,б)

Правый нижний квадрант является дополнением к исходному набору кодовых слов. С помощью подобной процедуры можно определить и ортогональный набор $\mathbf{H}_3$ для набора 3-битовых данных.

Набор данных	Набор ортогональных кодовых слов

$$\begin{array}{c} 0\ 0\ 0 \\ 0\ 0\ 1 \\ 0\ 1\ 0 \\ 0\ 1\ 1 \\ \\ 1\ 0\ 0 \\ 1\ 0\ 1 \\ 1\ 1\ 0 \\ 1\ 1\ 1 \end{array} \quad \mathbf{H}_3 = \left[\begin{array}{cccc:cccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cc} \mathbf{H}_2 & \overline{\mathbf{H}_2} \\ \mathbf{H}_2 & \overline{\mathbf{H}_2} \end{array}\right] \qquad (6.4,\text{в})$$

Вообще, для набора k-битовых данных из матрицы $\mathbf{H}_{k-1}$, можно построить набор кодовых слов $\mathbf{H}_k$ размерностью $2^k \times 2^k$, который называется *матрицей Адамара* (Hadamard matrix).

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{H}_{k-1} & \overline{\mathbf{H}_{k-1}} \\ \mathbf{H}_{k-1} & \overline{\mathbf{H}_{k-1}} \end{bmatrix} \qquad (6.4,\text{г})$$

Каждая пара слов в каждом наборе кодовых слов $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{H}_3$, ..., $\mathbf{H}_k$, ... содержит одинаковое количество совпадающих и несовпадающих разрядов [2]. Поэтому, в соответствии с уравнением (6.3), $z_{ij} = 0$ (при $i \neq j$) и каждый из этих наборов ортогонален.

Точно так же, как M-арная передача сигналов с ортогональной модуляцией (такой, как MFSK) понижает P_B, кодирование информации ортогональным набором сигналов при когерентном обнаружении дает *абсолютно такой же* результат. Для одинаковых, равноэнергетических ортогональных сигналов вероятность ошибки в кодовом слове (символе), P_E, можно оценить сверху, как [2]

$$P_E(M) \leq (M-1)(M-1)Q\left(\sqrt{\frac{E_s}{N_0}}\right), \qquad (6.5)$$

где размер набора кодовых слов M равен 2^k, а k — это число информационных бит в кодовом слове. Функция $Q(x)$ определена в уравнении (3.43), а $E_s = kE_b$ является энергией кодового слова. При фиксированном M с ростом E_n/N_0 оценка становится все более точной; уже для $P_E(M) \leq 10^{-3}$ уравнение (6.5) является довольно хорошим приближением. Для выражения вероятности появления ошибочного бита мы будем использовать связь между P_B и P_E, которая дается уравнением (4.112). Приведем ее повторно.

$$\frac{P_B(k)}{P_E(k)} = \frac{2^{k-1}}{2^k - 1} \quad \text{или} \quad \frac{P_B(M)}{P_E(M)} = \frac{M/2}{(M-1)} \qquad (6.6)$$

В результате объединения уравнений (6.5) и (6.6) вероятность появления ошибочного бита можно оценить следующим образом.

$$P_B(k) \leq (2^{k-1})Q\left(\sqrt{\frac{kE_B}{N_0}}\right) \quad \text{или} \quad P_B(M) \leq \frac{M}{2}Q\left(\sqrt{\frac{kE_s}{N_0}}\right) \qquad (6.7)$$

6.1.3.2. Биортогональные коды

Биортогональный набор сигналов, состоящий из M сигналов или кодовых слов, получается из ортогонального набора, состоящего из $M/2$ сигналов, путем дополнения последнего сопряженными значениями каждого сигнала.

$$\mathbf{B}_k = \begin{bmatrix} \mathbf{H}_{k-1} \\ \hline \mathbf{H}_{k-1} \end{bmatrix}$$

Например, набор 3-битовых данных можно преобразовать в биортогональный набор кодовых слов следующим образом.

Набор данных	Набор ортогональных кодовых слов
0 0 0	$\mathbf{B}_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$
0 0 1	
0 1 0	
0 1 1	
1 0 0	
1 0 1	
1 1 0	
1 1 1	

В действительности биортогональный набор состоит из двух ортогональных кодов, таких, что для каждого кодового слова в одном наборе имеется антиподное ему слово в другом. Биортогональный набор состоит из *комбинации ортогональных и антиподных* сигналов. Если использовать коэффициенты z_{ij}, введенные в уравнении (6.1), то биортогональные коды можно представить следующим образом.

$$z_{ij} = \begin{cases} 1 & \text{для } i = j \\ -1 & \text{для } i \neq j, |i-j| = \dfrac{M}{2} \\ 0 & \text{для } i \neq j, |i-j| \neq \dfrac{M}{2} \end{cases} \tag{6.8}$$

Одно из преимуществ биортогональных кодов перед ортогональными заключается в том, что при передаче аналогичной информации размер кодового слова биортогональных кодов *вдвое меньше* размера кодового слова ортогональных кодов (сравните строки матрицы $\mathbf{B}_3$ со строками представленной ранее матрицы $\mathbf{H}_3$). Следовательно, при использовании биортогональных кодов требования к полосе пропускания вдвое слабее, чем при использовании ортогональных кодов. Поскольку антиподные векторы сигналов имеют лучшие пространственные характеристики, чем ортогональные, не должно удивлять, что биортогональные коды лучше ортогональных. Для одинаковых, равноэнергетических биортогональных сигналов вероятность ошибки в кодовом слове (символе) можно оценить [2] следующим образом.

$$P_E(M) \le (M-2)Q\left(\sqrt{\frac{E_s}{N_0}}\right) + Q\left(\sqrt{\frac{2E_s}{N_0}}\right) \tag{6.9}$$

При фиксированном M с ростом E_n/N_0 оценка становится все более точной. Зависимость $P_B(M)$ от $P_E(M)$ является довольно сложной, но ее, согласно [2], можно аппроксимировать следующим образом.

$$P_B(M) \approx \frac{P_E(M)}{2}$$

Это приближение становится достаточно хорошим при $M > 8$. Таким образом, можно записать следующее.

$$P_E(M) \lesssim \frac{1}{2}\left[(M-2)Q\left(\sqrt{\frac{E_s}{N_0}}\right) + Q\left(\sqrt{\frac{2E_s}{N_0}}\right)\right] \tag{6.10}$$

Описанные биортогональные коды значительно снижают P_B по сравнению с ортогональными кодами и требуют только *вдвое меньшей полосы пропускания*, чем аналогичные ортогональные коды.

6.1.3.3. Трансортогональные (симплексные) коды

Код, получаемый из ортогонального путем удаления первого разряда каждого кодового слова, называется *трансортогональным* (transorthogonal), или *симплексным* (simplex) *кодом*. Такой код описывается следующими значениями z_{ij}.

$$z_{ij} = \begin{cases} 1 & \text{для } i = j \\ \dfrac{-1}{M-1} & \text{для } i \ne j \end{cases} \tag{6.11}$$

С точки зрения *минимальной энергии*, необходимой для поддержания заданной вероятности ошибки, симплексный код эквивалентен равновероятному ортогональному набору. Сравнивая достоверность передачи ортогонального, биортогонального и симплексного кодов, можно сказать, что симплексный код имеет наименьшее требуемое E_b/N_0 для получения определенной частоты появления символьных ошибок. Впрочем, *при больших M* все три схемы *очень похожи между собой* в смысле достоверности передачи. При этом биортогональное кодирование, по сравнению с другими методами, требует лишь половины полосы пропускания. В то же время для каждого из этих кодов требования к полосе пропускания (и сложность системы) экспоненциально растут с увеличением M; так что подобные схемы кодирования годятся лишь тогда, когда доступна значительная полоса пропускания.

6.1.4. Примеры системы кодирования сигналов

На рис. 6.4 дается пример присвоения k-битовому сообщению из набора размером $M = 2^k$ кодированной последовательности импульсов из кодового набора аналогичного размера. Каждое из k-битовых сообщений выбирает один генератор, производящий кодированную последовательность или кодовое слово. Последовательности в кодированном наборе, заменяющие исходные сообщения, формируют набор сигналов с хорошими пространственными характеристиками (например, ортогональный, биортогональный). Для ортогонального кода, описанного в разделе 6.1.3.1, каждое кодовое слово состоит из $M = 2^k$ импульсов

(представляющих кодовые биты). Таким образом, 2^k кодовых бит заменяют k информационных бит. Затем выбранная последовательность с использованием двоичной PSK модулируется несущей волной, так что фаза ($\phi_j = 0$ или π) несущей волны в течение каждого интервала передачи кодированного бита, $0 \le t \le T_c$, соответствует амплитуде ($j = -1$ или 1) j-го биполярного импульса в кодовом слове. В приемнике, показанном на рис. 6.5, сигнал демодулируется и подается на M корреляторов (или согласованных фильтров). Для ортогональных кодов, таких как описаны в разделе 6.1.3.1 (которые определяются матрицей Адамара), за период передачи кодового слова ($T = 2^k T_c$) определяются корреляции принятого сигнала. Для систем связи реального времени сообщения не могут опаздывать, поэтому время передачи кодового слова должно совпадать с длительностью сообщения. Следовательно, T также можно выразить как $T = (\log_2 M)T_b = kT_b$, где T_b — длительность битов сообщения. Отметим, что длительность бита сообщения в M/k раз больше, чем у кодового бита. Другими словами, кодовые биты или кодированные импульсы (сигналы PSK) должны перемещаться со скоростью, в M/k раз большей, чем биты сообщения. Для ортогонально кодированных сигналов и каналов с шумом AWGN математическое ожидание выходной мощности для каждого коррелятора в момент времени T равно нулю; исключением является только коррелятор, соответствующий переданному кодовому слову.

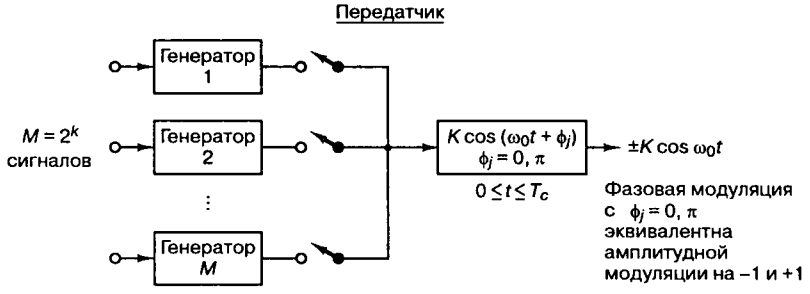


Рис. 6.4. Система кодирования сигналов (передатчик)

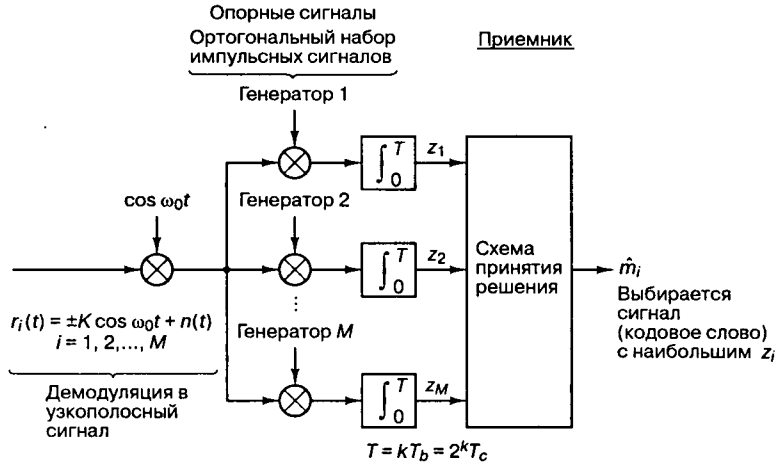


Рис. 6.5. Система кодирования сигналов с когерентным обнаружением (приемник)

Каковы преимущества описанного ортогонального кодирования сигналов по сравнению с обычным поступлением в каждую единицу времени одного бита или одного импульса? Можно оценить достоверность передачи с таким кодированием и без него, сравнив уравнение (4.79) для когерентного обнаружения антиподных сигналов с уравнением (6.7) для когерентного обнаружения ортогональных кодовых слов. При данном размере k-битового сообщения (скажем, $k = 5$) и желаемой вероятности появления ошибочного бита (например, 10^{-5}), обнаружение ортогональных кодовых слов (каждое из которых состоит из 5 бит) может выполняться с приблизительно на 2,9 дБ меньшим отношением E_b/N_0, чем побитовое обнаружение антиподных сигналов. (Проверить этот факт предоставляется читателю в качестве задачи 6.28.) Данный результат можно было предвидеть, сравнив рабочие характеристики ортогональной передачи сигналов на рис. 4.28 с характеристиками бинарной (антиподной) передачи на рис. 4.29. Чем мы платим за такой уровень достоверности передачи? Плата выражается в увеличении полосы пропускания. В приведенном примере передача некодированного сообщения — это посылка 5 бит. Сколько кодированных импульсов необходимо отправить для передачи с кодированием каждой последовательности сообщения? В данном примере каждая 5-битовая последовательность сообщения представлена $M = 2^k = 2^5 = 32$ кодовыми битами или кодированными импульсами. 32 кодированных импульса, составляющих кодовое слово, нужно отправить за то же время, что и соответствующие исходные 5 бит. Таким образом, требуемая ширина полосы пропускания составляет 32/5 от ширины полосы пропускания в случае без кодирования. Вообще, полоса пропускания, необходимая для подобных ортогонально кодированных сигналов, в M/k раз больше требуемой в случае передачи без кодирования. Далее мы рассмотрим более выгодные и эффективные способы получения компромиссов между шириной полосы пропускания и схемой кодирования [3, 4].

6.2. Типы защиты от ошибок

Перед тем как начать обсуждение структурированной избыточности, рассмотрим два основных метода использования избыточности для защиты от ошибок. В первом методе, *обнаружение ошибок и повторная передача*, для проверки на наличие ошибки используется контрольный бит четности (дополнительный бит, присоединяемый к данным). При этом приемное оконечное устройство не предпринимает попыток исправить ошибку, оно просто посылает передатчику запрос на повторную передачу данных. Следует заметить, что для такого диалога между передатчиком и приемником необходима двухсторонняя связь. Второй метод, *прямое исправление ошибок* (forward error correction — FEC), требует лишь односторонней линии связи, поскольку в этом случае контрольный бит четности служит как для обнаружения, так и исправления ошибок. Далее мы увидим, что не все комбинации ошибок можно исправить, так что коды коррекции классифицируются в соответствии с их возможностями исправления ошибок.

6.2.1. Связность оконечных устройств

Оконечные устройства систем связи часто классифицируют согласно их связности с другими оконечными устройствами. Возможные типы соединения, показанные на рис. 6.6, называются *симплексными* (simplex) (не путайте с симплексными, или трансортогональными кодами), *полудуплексными* (half-duplex) и *полнодуплексными* (full-duplex). Симплексное соединение на рис. 6.6, *а* — это односторонняя линия связи.

Передача сигналов производится *только* от оконечного устройства А к оконечному устройству В. Полудуплексное соединение на рис. 6.6, *б* — это линия связи, посредством которой можно осуществлять передачи сигналов в обоих направлениях, но не одновременно. И наконец, полнодуплексное соединение (рис. 6.6, *в*) — это двусторонняя связь, где передача сигналов происходит одновременно в обоих направлениях.

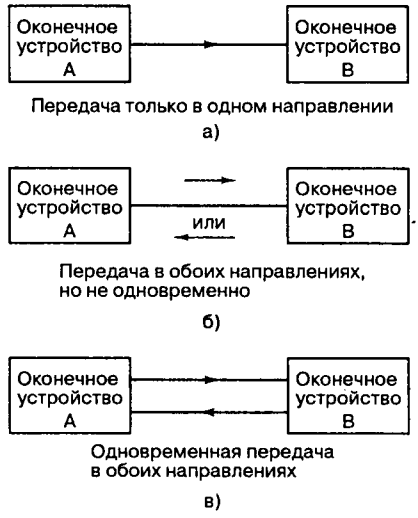


Рис. 6.6. Классификация связности оконечных устройств: а) симплексная; б) полудуплексная; в) полнодуплексная

6.2.2. Автоматический запрос повторной передачи

Если защита от ошибок заключается только в их обнаружении, система связи должна обеспечить средства предупреждения передатчика об опасности, сообщающие, что была обнаружена ошибка и требуется повторная передача. Подобные процедуры защиты от ошибок известны как методы *автоматического запроса повторной передачи* (Automatic Repeat Request — ARQ). На рис. 6.7 показаны три наиболее распространенные процедуры ARQ. На каждой схеме ось времени направлена слева направо. Первая процедура ARQ, *запрос ARQ с остановками* (stop-and-wait ARQ), показана на рис. 6.7, *а*. Ее реализация требует только полудуплексного соединения, поскольку передатчик перед началом очередной передачи ожидает подтверждения об успешном приеме (acknowledgement — ACK) предыдущей. В примере, приведенном на рисунке, третий блок передаваемых данных принят с ошибкой. Следовательно, приемник передает отрицательное подтверждение приема (negative acknowledgment — NAK); передатчик повторяет передачу третьего блока сообщения и только после этого передает следующий по очередности блок. Вторая процедура ARQ, *непрерывный запрос ARQ с возвратом* (continuous ARQ with pullback), показана на рис. 6.7, *б*. Здесь требуется полнодуплексное соединение. Оба оконечных устройства начинают передачу одновременно: передатчик отправляет информацию, а приемник передает подтверждение о приеме данных. Следует отметить, что каждому блоку передаваемых данных присваивается порядковый номер. Кроме того, номера кадров ACK и NAK должны быть согласованы; иначе говоря, задержка распространения сигнала должна быть известна *априори*,

чтобы передатчик знал, к какому блоку сообщения относится данный кадр подтверждения приема. В примере на рис. 6.7, *б* время подобрано так, что между отправленным блоком сообщений и полученным подтверждением о приеме существует постоянный интервал в четыре блока. Например, после отправки сообщения 8, приходит сигнал NAK, сообщающий об ошибке в блоке 4. При использовании процедуры ARQ передатчик "возвращается" к сообщению с ошибкой и снова передает всю информацию, начиная с поврежденного сообщения. И наконец, третья процедура, именуемая *непрерывным запросом ARQ с выборочным повторением* (continuous ARQ with selective repeat), показана на рис. 6.7, *в*. Здесь, как и во второй процедуре, требуется полнодуплексное соединение. Впрочем, в этой процедуре повторно передается только искаженное сообщение; затем передатчик продолжает передачу с того места, где она прервалась, не выполняя повторной передачи правильно принятых сообщений.

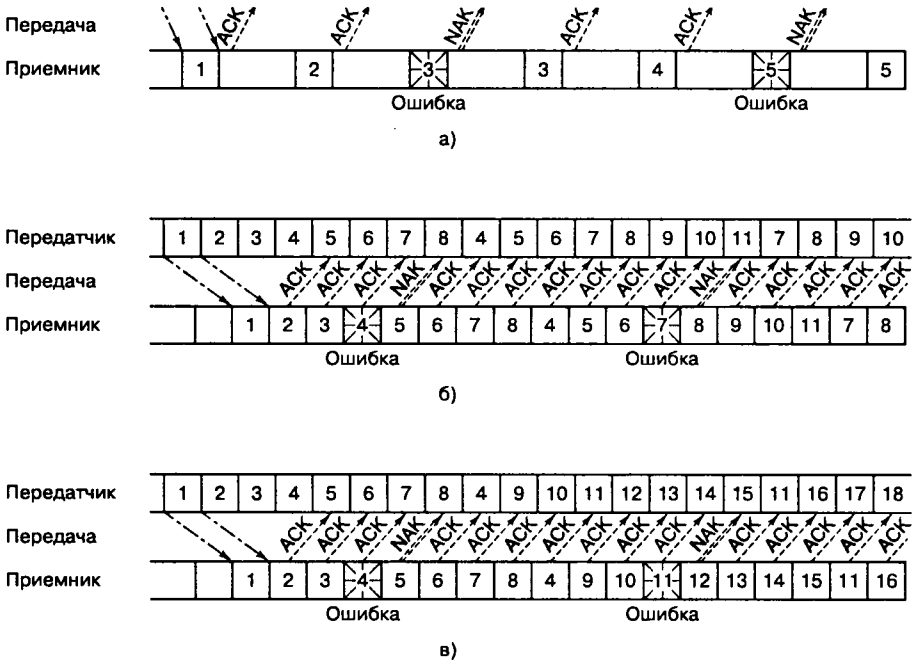


Рис. 6.7. Автоматический запрос повторной передачи (ARQ): а) запрос ARQ с остановками (полудуплексная связь); б) непрерывный запрос ARQ с возвратом (полнодуплексная связь); в) непрерывный запрос ARQ с выборочным повторением (полнодуплексная связь)

Выбор конкретной процедуры ARQ является компромиссом между требованиями эффективности применения ресурсов связи и необходимостью полнодуплексной связи. Полудуплексная связь (рис. 6.7, *а*) требует меньших затрат, нежели полнодуплексная; в то же время она менее эффективна, что можно определить по количеству пустых временных интервалов. Более эффективная работа, показанная на рис. 6.7, *б*, требует более дорогой полнодуплексной связи.

Главное преимущество схем ARQ перед схемами прямого исправления ошибок (forward error correction — FEC) заключается в том, что обнаружение ошибок требует более простого декодирующего оборудования и меньшей избыточности, чем коррекция ошибок. Кроме того, она гибче; информация передается повторно только при об-

наружении ошибки. С другой стороны, метод FEC может оказаться более приемлемым (или дополняющим) по какой-либо из следующих причин.

1. Обратный канал недоступен или задержка при использовании ARQ слишком велика.
2. Алгоритм повторной передачи нельзя реализовать удобным образом.
3. При ожидаемом количестве ошибок потребуется слишком много повторных передач.

6.3. Структурированные последовательности

В разделе 4.8 мы рассмотрели цифровую передачу данных посредством $M = 2^k$ сигналов (M-арная передача сигнала), где каждый сигнал содержит k бит информации. Было показано, что при ортогональной M-арной передаче сигналов уменьшения вероятности ошибки P_B можно добиться путем увеличения M (расширения полосы пропускания). В разделе 6.1 мы показали, что P_B можно уменьшить за счет кодирования k двоичных битов в одно из M ортогональных кодовых слов. Одним из основных недостатков ортогонального кодирования является неэффективное использование полосы пропускания. При наборе ортогональных кодов, включающем $M = 2^k$ сигналов, требуемая ширина полосы пропускания в M/k раз больше необходимой для передачи некодированного сигнала. В этом и последующих разделах мы отойдем от рассмотрения ортогональных или антиподных свойств сигналов и сосредоточим внимание на классе процедур кодирования, известных как *коды с контролем четности* (parity-check codes). Эти процедуры канального кодирования относятся к *структурированным последовательностям*, поскольку они представляют методы введения в исходные данные структурированной избыточности таким образом, что это позволяет обнаруживать или исправлять ошибки. Как показано на рис. 6.1, структурированные последовательности делятся на три подкатегории: *блочные*, *сверточные* и *турбокоды*. Блочное кодирование рассматривается в этой главе, а другие описываются в главах 7 и 8.

6.3.1. Модели каналов

6.3.1.1. Дискретный канал без памяти

Дискретный канал без памяти (discrete memoryless channel — DMC) характеризуется дискретным входным алфавитом, дискретным выходным алфавитом и набором условных вероятностей $P(i \mid j)$ $(1 \le i \le M,\ 1 \le j \le Q)$, где i представляет модулятор M-арного входного символа, j — демодулятор Q-арного выходного символа, а $P(i \mid j)$ — это вероятность приема символа j при переданном символе i. Каждый выходной символ канала зависит только от соответствующего входного символа, так что для данной входной последовательности $\mathbf{U} = u_1, u_2, u_3, \ldots, u_m, \ldots, u_N$ условную вероятность соответствующей выходной последовательности $\mathbf{Z} = z_1, z_2, \ldots, z_m, \ldots, z_N$ можно записать следующим образом.

$$P(\mathbf{Z}|\mathbf{U}) = \prod_{m=1}^{N} P(z_m|u_m) \tag{6.12}$$

Если же канал *имеет память* (т.е. в пакете данных имеются помехи или канал подвергается воздействию замирания), условную вероятность последовательности $\mathbf{Z}$ нужно выражать как *совместную* вероятность всех элементов последовательности. Уравнение (6.12) — это условие *отсутствия памяти* у канала. Поскольку считается, что шум в

канале без памяти влияет на каждый символ независимо от других, то в этом случае условная вероятность **Z** является произведением вероятностей независимых элементов.

6.3.1.2. Двоичный симметричный канал

Двоичный симметричный канал (binary symmetric channel — BSC) является частным случаем дискретного канала без памяти, входной и выходной алфавиты которого состоят из двоичных элементов (0 и 1). Условные вероятности имеют симметричный вид.

$$P(0|1) = P(1|0) = p$$

и (6.13)

$$P(1|1) = P(0|0) = 1 - p$$

Уравнение (6.13) выражает так называемые *вероятности перехода*. Иными словами, при передаче канального символа вероятность принятия его с ошибкой равна p (относительно значения энергии), а вероятность того, что он передан без ошибки, — $(1-p)$. Поскольку на выход демодулятора поступают дискретные элементы 0 или 1, говорят, что по отношению к каждому символу демодулятор принимает *жесткое решение* (hard decision). Рассмотрим наиболее распространенную схему кодирования — данные в формате BPSK плюс демодуляция по принципу жесткого решения. Вероятность появления ошибки в канальном символе находится с использованием метода, обсуждавшегося в разделе 4.7.1, и дается уравнением (4.79).

$$p = Q\left(\sqrt{\frac{2E_c}{N_0}}\right)$$

Здесь E_c/N_0 — отношение энергии канального символа к плотности помех, а функция $Q(x)$ была определена в уравнении (3.43).

Если описанная схема жестких решений применяется в системах с бинарными кодировками, то с демодулятора на декодер поступают двоичные *кодовые символы* или *биты канала*. Поскольку декодер работает на основе жестких решений, определяемых демодулятором, декодирование в двоичном симметричном канале называется также *жестким декодированием*.

6.3.1.3. Гауссов канал

Определение двоичного симметричного канала можно использовать и для каналов с недискретным алфавитом. Пример — *гауссов канал* с дискретным входным алфавитом и непрерывным выходным алфавитом, лежащим в диапазоне $(-\infty, \infty)$. Этот канал добавляет шум ко всем передаваемым символам. Поскольку шум — это гауссова случайная переменная с нулевым средним и дисперсией σ^2, результирующую функцию плотности вероятности принятой случайной величины z при условии передачи символа u_k (правдоподобие u_k) можно записать следующим образом.

$$p(z|u_k) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(z-u_k)^2}{2\sigma^2}\right] \tag{6.14}$$

для всех z, где $k = 1, 2, \ldots, M$.

В этом случае *отсутствие памяти* имеет то же значение, что и в разделе 6.3.1.1, а само уравнение (6.12) можно использовать при вычислении условной вероятности для последовательности **Z**.

Если на выходе демодулятора находится непрерывный алфавит или его квантовое приближение (с более чем двумя квантовыми уровнями), говорят, что демодулятор принимает *мягкое решение* (soft decision). Если в системе используется кодирование, демодулятор подает такие квантовые кодовые символы на декодер. Поскольку декодер работает на основе мягких решений, определяемых демодулятором, декодирование в гауссовом канале называется *мягким*.

В канале с жестким решением процесс обнаружения можно описать через вероятность символьной ошибки. Но в канале с мягкими решениями выбор детектора нельзя однозначно отнести к верному или неверному. Таким образом, поскольку определенного решения не существует, не может быть и выражения для вероятности ошибки; детектор может только определять семейство условных вероятностей или правдоподобий разных типов символов.

В принципе, декодеры с мягкими решениями можно сделать, но для блочных кодов они будут значительно сложнее декодеров с жесткими решениями; поэтому, как правило, блочные коды реализуются в системах с декодерами, работающими по принципу жесткого решения. Для сверточных кодов реализация и жестких, и мягких решений одинаково популярна. В этой главе мы предполагаем, что каналы являются двоичными симметричными и, следовательно, декодеры используют жесткие решения. В главе 7 мы перейдем к обсуждению жесткого и мягкого декодирования для сверточных кодов, а также продолжим обсуждение моделей канала.

6.3.2. Степень кодирования и избыточность

При использовании блочных кодов исходные данные делятся на блоки из k бит, которые иногда называют информационными битами, или битами сообщения; каждый блок может представлять любое из 2^k отдельных сообщений. В процессе кодирования каждый k-битовый блок данных преобразуется в больший блок из n бит, который называется кодовым битом, или канальным символом. К каждому блоку данных кодирующее устройство прибавляет $(n - k)$ бит, которые называются *избыточными битами* (redundant bits), *битами четности* (parity bits), или *контрольными битами* (check bits); новой информации они не несут. Для обозначения описанного кода используется запись (n, k). Отношение числа избыточных бит к числу информационных бит, $(n - k)/k$, называется *избыточностью* (redundancy) кода; отношение числа бит данных к общему числу бит, k/n, именуется *степенью кодирования* (code rate). Под степенью кодирования подразумевается доля кода, которая приходится на полезную информацию. Например, в коде со степенью 1/2, каждый кодовый бит несет 1/2 бит информации.

В этой главе и в главах 7 и 8 мы рассмотрим методы кодирования, получающие избыточность за счет увеличения необходимой ширины полосы. Например, метод защиты от ошибок, использующий код со степенью 1/2 (100%-ная избыточность), будет требовать двойной, по сравнению с некодированной передачей, полосы пропускания. В то же время, если использовать код со степенью 3/4, то избыточность составит 33%, и увеличение полосы пропускания будет всего 4/3. В главе 9 мы рассмотрим методы модуляции/кодирования для узкополосных каналов, где защита от ошибок происходит не за счет увеличения полосы пропускания, а за счет усложнения метода (и, как следствие, его аппаратной реализации).

6.3.2.1. Терминология в кодировании

Разные авторы по-разному называют элементы на выходе кодирующего устройства: кодовые биты (code bits), канальные биты (channel bits), кодовые символы (code symbols), ка-

нальные символы (channel symbols), биты четности (parity bits), символы четности (parity symbols). Вообще, по смыслу эти термины очень похожи между собой. В этой книге для двоичных кодов термины "кодовые биты", "канальные биты", "кодовые символы" и "канальные символы" употребляются как синонимы. Следует уточнить, что названия "кодовые биты" и "канальные биты" подходят для описания только двоичных кодов. Такие общие названия, как "кодовые символы" и "канальные символы", зачастую более предпочтительны, поскольку они могут означать как двоичное, так и любое другое кодирование. Отметим, что эти понятия не следует путать с тем, что получается при группировке битов в передаваемые символы, о которых шла речь в предыдущей главе. Термины "биты четности" и "символы четности" применяются только к тем составляющим кода, которые представляют избыточные компоненты, прибавляемые к исходным данным.

6.3.3. Коды с контролем четности

6.3.3.1. Код с одним контрольным битом

Коды с контролем четности (parity-check code) для обнаружения или исправления ошибок используют линейные суммы информационных битов, которые называются *символами* (parity symbols), или *битами четности* (parity bits). Код с одним контрольным битом — это прибавление к блоку информационных битов одного контрольного бита. Этот бит (бит четности) может быть равен нулю или единице, причем его значение выбирается так, чтобы сумма всех битов в кодовом слове была четной или нечетной. В операции суммирования используется арифметика по модулю 2 (операция исключающего ИЛИ), описанная в разделе 2.9.3. Если бит четности выбран так, что результат четный, то говорят, что схема имеет *положительную четность* (even parity); если при добавлении бита четности результирующий блок данных является нечетным, то говорят, что он имеет *отрицательную четность* (odd parity). На рис. 6.8, *а* показана последовательная передача данных (первым является крайний справа бит). К каждому блоку добавляется один бит четности (крайний слева бит в каждом блоке), дающий положительную четность.

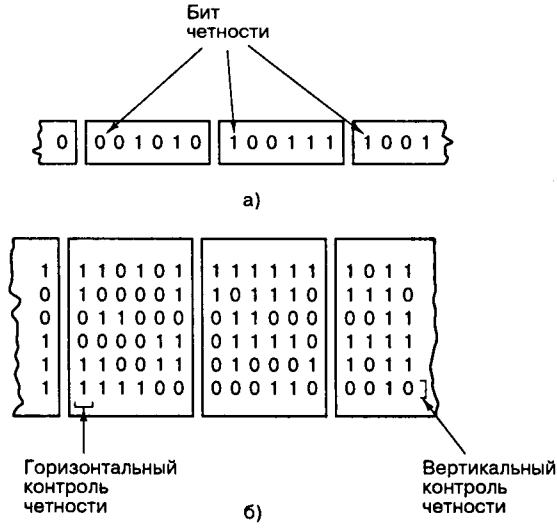


Рис. 6.8. Проверка четности для последовательной и параллельной структуры кода: а) последовательная структура; б) параллельная структура

В приемном оконечном устройстве производится декодирование, заключающееся в проверке, дают ли нуль суммы принятых битов кодового слова по модулю 2 (положительная четность). Если полученный результат равен 1, то кодовое слово заведомо содержит ошибки. Степень кодирования такого кода можно записать как $k/(k+1)$. Как вы думаете, может ли декодер автоматически *исправить* цифру, полученную с ошибкой? Нет, это невозможно. Можно только *обнаружить*, что в кодовом символе присутствует нечетное количество ошибок. (Если ошибка была внесена в четное число битов, то проверка четности покажет отсутствие ошибок; данный случай — это пример *необнаруженной ошибки*.) Предполагая, что ошибки во всех разрядах равновероятны и появляются независимо, можно записать вероятность появления j ошибок в блоке, состоящем из n символов.

$$P(j,n)=\binom{n}{j}p^{j}(1-p)^{n-j}. \tag{6.15}$$

Здесь p — вероятность получения *канального символа* с ошибкой, а через

$$\binom{n}{i}=\frac{n!}{j!(n-j)!} \tag{6.16}$$

— обозначается число различных способов выбора из n бит j ошибочных. Таким образом, для кода с одним битом четности вероятность необнаруженной ошибки P_{nd} в блоке из n бит вычисляется следующим образом.

$$P_{\text{nd}}=\sum_{j=1}^{\substack{n/2\,(\text{при } n=\text{четное})\\(n-1)/2\,(\text{при } n=\text{нечетное})}}\binom{n}{2j}p^{2j}(1-p)^{n-2j} \tag{6.17}$$

Пример 6.1. Код положительной четности

Нужно создать код обнаружения ошибок (4, 3) положительной четности, причем символ четности должен располагаться на крайней левой позиции кодового слова. Какие ошибки может обнаружить код? Вычислите вероятность необнаруженной ошибки сообщения, предполагая, что все символьные ошибки являются независимыми событиями и вероятность ошибки в канальном символе равна $p = 10^{-3}$.

Решение

Сообщение	Четность	Кодовое слово	
000	0	0	000
100	1	1	100
010	1	1	010
110	0	0	110
001	1	1	001
101	0	0	101
011	0	0	011
111	1	1	111
		↗	↖
		Четность	Сообщение

Код может выявлять все комбинации с одной или тремя ошибками. Вероятность необнаруженной ошибки равна вероятности появления где-либо в кодовом слове двух или четырех ошибок.

$$
\begin{aligned}
P_{\text{nd}} &= \binom{4}{2}p^2(1-p)^2 + \binom{4}{4}p^4 = \\
&= 6p^2(1-p)^2 + p^4 = \\
&= 6p^2 - 12p^3 + 7p^4 = \\
&= 6(10^{-3})^2 - 12(10^{-3})^3 + 7(10^{-3})^4 \approx 6\times 10^{-6}
\end{aligned}
$$

6.3.3.2. Прямоугольный код

Прямоугольный код (rectangular code), называемый также *композиционным* (product code), можно представить в виде параллельной структуры кода, изображенной на рис. 6.8, *б*. Код создается следующим образом. Вначале из битов сообщения строятся прямоугольники, состоящие из M строк и N столбцов; затем к каждой строке и каждому столбцу прибавляется бит четности, что в результате дает матрицу размером $(M+1)\times(N+1)$. Степень кодирования прямоугольного кода, k/n, может быть записана следующим образом.

$$\frac{k}{n} = \frac{MN}{(M+1)(N+1)}$$

Насколько прямоугольный код мощнее кода, который имеет один контрольный бит и предоставляет только возможность обнаружить ошибку? Отметим, что любая отдельная ошибка в разряде приведет к нарушению четности в одном столбце *и* в одной из строк матрицы. Следовательно, прямоугольный код может исправить любую единичную ошибку, поскольку расположение такой ошибки однозначно определяется пересечением строки и столбца, в которых была нарушена четность. В примере, показанном на рис. 6.8, *б*, размеры матрицы равны $M = N = 5$; следовательно, на рисунке отображен код (36, 25), способный исправлять единичные ошибки, расположенные в любом из 36 двоичных разрядов. Вычислим для такого блочного кода с коррекцией ошибок вероятность появления неисправленной ошибки, для чего учтем все способы появления *ошибки сообщения*. Исходя из вероятности наличия j ошибок в блоке из n символов, записанной в выражении (6.5), можно записать вероятность ошибки сообщения, называемой также *блочной ошибкой* или *ошибочным словом*, для кода, который может исправить ошибочные комбинации, состоящие из t или менее ошибочных битов.

$$P_M = \sum_{j=t+1}^{n}\binom{n}{j}p^j(1-p)^{n-j} \tag{6.18}$$

Здесь p — вероятность получения ошибочного *канального символа*. В примере на рис. 6.8, *б* код может исправить все однобитовые ошибки ($t = 1$) в прямоугольном блоке, состоящем из $n = 36$ бит. Следовательно, суммирование в уравнении (6.18) начинается с $j = 2$.

$$P_M = \sum_{j=2}^{36}\binom{36}{j}p^j(1-p)^{36-j}$$

При достаточно малом p, наибольший вклад дает первое слагаемое суммы. Следовательно, для примера с прямоугольным кодом (36, 25) можно записать следующее.

$$P_M \approx \binom{36}{2} p^2 (1-p)^{34}$$

Точная *вероятность битовой ошибки* P_B зависит от конкретного кода и используемого декодера. Приближенные значения P_B приводятся в разделе 6.5.3.

6.3.4. Зачем используется кодирование с коррекцией ошибок

Кодирование с коррекцией ошибок можно рассматривать как инструмент, реализующий различные компромиссы системы. На рис. 6.9 приведен сравнительный вид двух кривых, описывающих зависимость достоверности передачи от отношения E_b/N_0. Одна кривая соответствует обычной схеме модуляции без кодирования, а вторая представляет такую же модуляцию, но уже с использованием кодирования. Ниже подробно рассмотрено четыре компромисса, имеющие место при канальном кодировании.

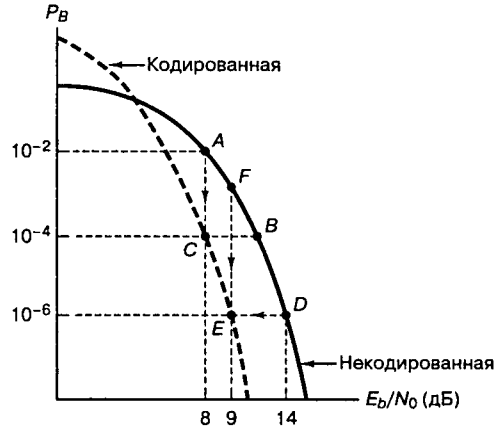


Рис. 6.9. Сравнение типичной достоверности передачи при использовании схемы с кодированием и схемы без кодирования

6.3.4.1. Компромисс 1: достоверность или полоса пропускания

Представим себе, что разработана простая, недорогая система речевой связи, которая была установлена у заказчика. Система не использует кодирование с коррекцией ошибок. Пусть рабочая точка системы совпадает с точкой A на рис. 6.9 (E_b/N_0 = 8 дБ, $P_B = 10^{-2}$). После нескольких испытаний у заказчика появляются жалобы на качество связи; он полагает, что вероятность появления битовой ошибки должна быть не выше 10^{-4}. Обычным способом удовлетворения требования заказчика является сдвиг рабочей точки из точки *A*, например, в точку *B* (рис. 6.9). В то же время допустим, что E_b/N_0, равное 8 дБ, — это максимальное значение, возможное в данной системе. Из рис. 6.9 видим, что один из возможных выходов из ситуации (компромиссов) — это сдвиг рабочей точки из точки *A* в точку *C*. Иными словами, "съехав" по вертикали вниз в точку *C* на кривой, отвечающей кодированному случаю, можно предоставить заказчику более высокую достоверность передачи данных. Чего это будет стоить? Помимо введения новых компонентов (кодера и декодера), это приведет к увеличению

необходимой полосы пропускания. Кодирование с коррекцией ошибок требует избыточности. Если предположить, что связь будет происходить в реальном времени (так что сообщения не могут задерживаться), добавление избыточных битов потребует увеличения скорости передачи и, конечно же, большей полосы пропускания.

6.3.4.2. Компромисс 2: мощность или полоса пропускания

Допустим, заказчику установлена система без кодирования с рабочей точкой, совпадающей с точкой D на рис. 6.9 ($E_b/N_0 = 14$ дБ, $P_B = 10^{-6}$). Заказчик не имеет претензий к качеству связи, но с помощью данного оборудования затруднительно получить требуемые $E_b/N_0 = 14$ дБ. Иными словами, оборудование постоянно работает на грани отказа. Если снизить требования к E_b/N_0 или мощности, то проблем с надежностью оборудования можно избежать. В контексте рис. 6.9 данные меры выглядят как сдвиг рабочей точки из D в E. Другими словами, требуемое значение E_b/N_0 можно получить, если применить кодирование с коррекцией ошибок. Таким образом, при фиксированном качестве связи компромисс заключается в получении большей производительности при снижении требований к мощности или E_b/N_0. Чем за это приходится платить? Тем же, чем и в прошлый раз, — большей полосой пропускания.

Заметим, что в системах, где не используется *связь в реальном времени*, применение кодирования с коррекцией ошибок даст несколько отличные результаты. Повышение достоверности передачи или понижение потребляемой мощности (подобное описанным выше случаям 1 или 2) будет достигаться за счет увеличения *времени задержки*, а не за счет расширения полосы пропускания.

6.3.4.3. Эффективность кодирования

Пример компромиссных решений, рассмотренный в предыдущем разделе, позволяет понизить E_b/N_0 с 14 до 9 дБ при поддержании той же достоверности передачи. В контексте этого примера и с помощью рис. 6.9 мы можем ввести понятие *эффективность кодирования* (coding gain). Итак, при *данной вероятности битовой ошибки* эффективность кодирования определяется как уменьшение E_b/N_0, которое достигается при использовании кодирования. Эффективность кодирования G, как правило, выражается в децибелах.

$$G(\text{дБ}) = \left(\frac{E_b}{N_0}\right)_u (\text{дБ}) - \left(\frac{E_b}{N_0}\right)_c (\text{дБ}) \tag{6.19}$$

Здесь $(E_b/N_0)_u$ и $(E_b/N_0)_c$ — требуемые некодированное и кодированное значения E_b/N_0.

6.3.4.4. Компромисс 3: скорость передачи данных или полоса пропускания

Пусть разработана система без кодирования с рабочей точкой, совпадающей с точкой D на рис. 6.9 ($E_b/N_0 = 14$ дБ, $P_B = 10^{-6}$). Допустим, что с качеством данных нет никаких проблем и нет особой нужды в снижении мощности. Однако у заказчика возросли требования к скорости передачи данных. Напомним в связи с этим уравнение (5.20,б).

$$\frac{E_b}{N_0} = \frac{P_r}{N_0}\left(\frac{1}{R}\right)$$

Если в системе ничего не менять, кроме скорости передачи данных R, то из приведенного выше выражения видно, что это приведет к уменьшению значения E_b/N_0 и

перемещению рабочей точки вверх, например из точки D в некоторую точку F. А теперь представим, что она "съезжает" вниз по вертикали в точку E на кривую, которая представляет кодированную модуляцию. Возрастание скорости передачи данных плохо отражается на качестве их передачи. В то же время применение кодирования с коррекцией ошибок восстанавливает утраченное качество, сохраняя при этом прежний уровень мощности (P_r/N_0). Итак, значение E_b/N_0 понижено, но код способствует получению той же вероятности ошибки при сниженном значении E_b/N_0. Какова цена такого увеличения скорости передачи данных или увеличения емкости? Как и раньше, это увеличение полосы пропускания.

6.3.4.5. Компромисс 4: пропускная способность или ширина полосы пропускания

Компромисс 4 сходен с компромиссом 3 в том, что оба дают возрастание пропускной способности. Метод множественного доступа, именуемый множественным доступом с кодовым разделением каналов (code-division multiple access — CDMA), который описывается в главе 12, — это один из стандартов, используемых в сотовой связи. При CDMA, когда все клиенты совместно используют общий спектр частот, каждый клиент является источником помех для других пользователей в той же ячейке или соседних. Поэтому пропускная способность (максимальное число клиентов) ячейки обратно пропорциональна значению E_b/N_0 (см. раздел 12.8). При этом снижение E_b/N_0 дает в итоге увеличение пропускной способности; код позволяет снизить мощности, используемые каждым клиентом, что, в свою очередь, приводит к увеличению общего числа клиентов. И снова платой за это является увеличение полосы пропускания. Но в этом случае увеличение полосы сигнала, получаемое при переходе к кодированию с коррекцией ошибок, незначительно, по сравнению с существенным увеличением полосы пропускания, получаемым при расширении спектра сигнала; поэтому при передаче данных оно не оказывает влияния на полосу пропускания.

В каждом из упомянутых выше компромиссов предполагалось использование "традиционного" кода с избыточными битами и более быстрая передача сигналов (для систем связи реального времени); следовательно, в каждом случае платой было расширение полосы передачи. В то же время существуют методы коррекции ошибок, называемые *решетчатым кодированием* (trellis-coded modulation), которые не требуют увеличения скорости передачи сигналов или расширения полосы частот для систем связи реального времени. (Эти методы рассмотрены в разделе 9.10.)

Пример 6.2. Связь вероятности ошибки с использованием кодирования

Сравните вероятность ошибки в сообщении для двух каналов связи — обычного и использующего кодирование с коррекцией ошибок. Пусть некодированная передача имеет следующие характеристики: модуляция BPSK, гауссов шум, $P_r/N_0 = 43\,776$, скорость передачи данных $R = 4800$ бит/с. Для случая с кодированием предполагается использование кода с коррекцией ошибок (15, 11), предоставляющего возможность исправления любых однобитовых ошибочных комбинаций кода в блоке из 15 бит. Будем считать, что демодулятор принимает жесткие решения и передает демодулированный код прямо на декодер, который, в свою очередь, определяет исходное сообщение.

Решение

Используем уравнение (4.79). Пусть $p_u = Q\sqrt{2E_b/N_0}$ и $p_c = Q\sqrt{2E_c/N_0}$ — вероятности символьных ошибок в канале без кодирования и в канале с кодированием, где E_b/N_0 — от-

ношение энергии бита к спектральной плотности мощности шума, а E_c/N_0 — отношение энергии кодированного бита к спектральной плотности мощности шума.

Без кодирования

$$\frac{E_b}{N_0} = \frac{P_r}{N_0}\left(\frac{1}{R}\right) = 9{,}12\ (9{,}6\ \text{дБ})$$

и

$$p_u = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) = Q(\sqrt{18{,}24}) = 1{,}02 \times 10^{-5} \tag{6.20}$$

Для $Q(x)$ используется следующее приближение, приведенное в уравнении (3.44).

$$Q(x) \approx \frac{1}{x\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) \quad \text{для } x > 3$$

Вероятность того, что некодированный блок сообщений P_M^u будет принят с ошибкой, равна 1 минус произведение вероятностей того, что каждый бит будет обнаружен правильно. Таким образом,

$$\begin{aligned} P_M^u &= 1-(1-p_u)^k = \\ &= \underbrace{1-(1-p_u)^{11}}_{\substack{\text{Вероятность правильности} \\ \text{всех 11 бит в некодированных} \\ \text{блоках}}} = \underbrace{1{,}12\times10^{-4}}_{\substack{\text{Вероятность ошибки,} \\ \text{по крайней мере, в одном} \\ \text{из 11 бит}}} \end{aligned} \tag{6.21}$$

С кодированием

Допустим, рассматриваемая система — это система связи реального времени, где задержки недопустимы, а скорость передачи канальных символов, или скорость передачи кодированных битов, равна $R_c = 15/11$ скорости некодированной передачи.

$$R_c = 4800 \times \tfrac{15}{11} \approx 6545 \text{ бит/с}$$

и

$$\frac{E_c}{N_0} = \frac{P_r}{N_0}\left(\frac{1}{R_c}\right) = 6{,}69\ (8{,}3\ \text{дБ})$$

Для каждого кодового бита значение E_c/N_0 меньше, чем в случае с некодированными битами данных. Это объясняется тем, что скорость передачи канальных битов возросла, а мощность передатчика при этом не изменилась.

$$p_c = Q\left(\sqrt{\frac{2E_c}{N_0}}\right) = Q(\sqrt{13{,}38}) = 1{,}36 \times 10^{-4} \tag{6.22}$$

Сравнивая выражения (6.20) и (6.22), можно видеть, что вследствие внесения избыточности вероятность ошибки в канальном бите уменьшилась. За то же время и с теми же номинальными мощностями нужно обнаружить большее число бит; повышение производительности в результате кодирования *еще не очевидно*. Вычислим теперь с помощью уравнения (6.18) частоту появления ошибок в кодированном сообщении P_M^c.

$$P_M^c = \sum_{j=2}^{n=15} \binom{15}{j} (p_c)^j (1-p_c)^{15-j}$$

Суммирование начинается с $j = 2$, поскольку код позволяет исправлять все однобитовые ошибки в блоках из $n = 15$ бит. Достаточно хорошее приближение можно получить, используя только первый член суммы. Для p_c используем значение, полученное из уравнения (6.22).

$$P_M^c \approx \binom{15}{2}(p_c)^2(1-p_c)^{13} = 1{,}94 \times 10^{-6} \tag{6.23}$$

Сравнивая выражения (6.21) и (6.23), можно видеть, что вследствие применения кода с коррекцией ошибок вероятность ошибки сообщения была уменьшена примерно в 58 раз. Данный пример иллюстрирует типичное поведение систем связи реального времени при использовании кодирования с коррекцией ошибок. Введение избыточности означает увеличение скорости передачи сигналов, уменьшение энергии, приходящейся на канальный символ, и увеличение числа ошибок вне демодулятора. Преимуществом такого подхода является то, что декодер (при разумном значении E_b/N_0) позволяет с лихвой компенсировать слабую производительность демодулятора.

6.3.4.6. Характеристики кода при низком значении E_b/N_0

В конце данной главы читателю предлагается решить задачу 6.5, сходную с примером 6.2. В п. а задачи 6.5, где значение E_b/N_0 принимается равным 14 дБ, кодирование дает повышение достоверности передачи сообщения. В то же время в п. б, где значение E_b/N_0 снижается до 10 дБ, кодирование не дает улучшения; фактически происходит ухудшение. Может возникнуть вопрос, почему в п. б происходит такое ухудшение? По сути, в обоих пунктах задачи применяется одна и та же процедура. Ответ можно найти на рис. 6.9, который наглядно показывает связь между кодированными и некодированными вероятностями ошибки. Хотя в задаче 6.5 речь идет о вероятности ошибки сообщения, а на рис. 6.9 приведен график битовой ошибки, следующее объяснение остается в силе. Итак, на подобных графиках кривые пересекаются (как правило, при низких значениях E_b/N_0). Смысл этого пересечения (порога) в том, что у всех систем кодирования имеется ограниченная способность к коррекции ошибок. Если в блоке имеется больше ошибок, чем способен исправить код, система будет работать плохо. Представим себе, что значение E_b/N_0 снижается непрерывно. Что мы увидим на выходе демодулятора? Демодулятор будет допускать все больше и больше ошибок. Следовательно, такое постепенное уменьшение E_b/N_0 должно в конце концов создать пороговую ситуацию, когда декодер будет переполнен ошибками. При достижении этого порога снижение производительности можно объяснить поглощением энергии избыточными битами, которые не дают никакого выигрыша. Не удивляет ли читателя то, что в области (низких значений E_b/N_0), где больше всего следовало бы ожидать улучшения достоверности передачи, код имеет наименьшую эффективность? Впрочем, существует класс мощных кодов, называемых *турбокодами* (turbo code), которые позволяют повысить надежность передачи при низких значениях E_b/N_0; у турбокодов точка пересечения графиков находится значительно ниже, чем у сверточных кодов. (Турбокоды рассматриваются в разделе 8.4.)

6.4. Линейные блочные коды

Линейные блочные коды (подобные коду, описанному в примере 6.2) — это класс кодов с контролем четности, которые можно описать парой чисел (n, k) (объяснение этой формы записи приводилось выше). В процессе кодирования блок из k символов сообщения (вектор сообщения) преобразуется в больший блок из n символов кодового слова (кодовый

вектор), образованного с использованием элементов данного алфавита. Если алфавит состоит только из двух элементов (0 и 1), код является двоичным и включает двоичные разряды (биты). Если не будет оговорено противное, наше последующее обсуждение линейных блочных кодов будет подразумевать именно двоичные коды.

k-битовые сообщения формируют набор из 2^k последовательностей сообщения, называемых *k-кортежами* (k-tuple) (последовательностями k цифр). n-битовые блоки могут формировать 2^n последовательности, также именуемые *n-кортежами*. Процедура кодирования сопоставляет с каждым из 2^k k-кортежей сообщения *один* из 2^n n-кортежей. Блочные коды представляют взаимно однозначное соответствие, в силу чего 2^k k-кортежей сообщения *однозначно* отображаются в множество из 2^k n-кортежей кодовых слов; отображение производится согласно таблице соответствия. Для *линейных кодов* преобразование отображения является, конечно же, *линейным*.

6.4.1. Векторные пространства

Множество всех двоичных n-кортежей, V_n, называется *векторным пространством* над двоичным полем двух элементов (0 и 1). В двоичном поле определены две операции, сложение и умножение, причем результат этих операций принадлежит этому же множеству двух элементов. Арифметические операции сложения и умножения определяются согласно обычным правилам для алгебраического поля [4]. Например, в двоичном поле правила сложения и умножения будут следующими.

Сложение	Умножение
$0 \oplus 0 = 0$	$0 \cdot 0 = 0$
$0 \oplus 1 = 1$	$0 \cdot 1 = 0$
$1 \oplus 0 = 1$	$1 \cdot 0 = 0$
$1 \oplus 1 = 0$	$1 \cdot 1 = 1$

Операция сложения, обозначаемая символом "$\oplus$", — это та же операция сложения по модулю 2, которая описывалась в разделе 2.9.3. Суммирование двоичных n-кортежей всегда производится путем сложения по модулю 2. Хотя для простоты мы чаще будем использовать для этой операции обычный знак +.

6.4.2. Векторные подпространства

Подмножество S векторного пространства V_n называется *подпространством*, если для него выполняются следующие условия.

1. Множеству S принадлежит нулевой вектор.
2. Сумма любых двух векторов в S также принадлежит S (*свойство замкнутости*).

При алгебраическом описании *линейных блочных кодов* данные свойства являются фундаментальными. Допустим, $\mathbf{V}_i$ и $\mathbf{V}_j$ — два кодовых слова (или кодовых вектора) в двоичном блочном коде (n, k). Код называется *линейным* тогда и только тогда, когда $(\mathbf{V}_i \oplus \mathbf{V}_j)$ также является кодовым вектором. Линейный блочный код — это такой код, в котором вектор, не принадлежащий подпространству, нельзя получить путем сложения любых кодовых слов, принадлежащих этому подпространству.

Например, векторное пространство $\mathbf{V}_4$ состоит из следующих шестнадцати 4-кортежей.

0000 0001 0010 0011 0100 0101 0110 0111
1000 1001 1010 1011 1100 1101 1110 1111

Примером подмножества $\mathbf{V}_4$, являющегося подпространством, будет следующее.

0000 0101 1010 1111

Легко проверить, что сложение любых двух векторов подпространства может дать в итоге лишь один из векторов подпространства. Множество из 2^k n-кортежей называется *линейным блочным кодом* тогда и только тогда, когда оно является подпространством векторного пространства $\mathbf{V}_n$ всех n-кортежей. На рис. 6.10 показана простая геометрическая аналогия, представляющая структуру линейного блочного кода. Векторное пространство $\mathbf{V}_n$ можно представить как составленное из 2^n n-кортежей. Внутри этого векторного пространства существует подмножество из 2^k n-кортежей, образующих подпространство. Эти 2^k вектора или точки показаны разбросанными среди более многочисленных 2^n точек, представляющих допустимые или возможные кодовые слова. Сообщение кодируется одним из 2^k возможных векторов кода, после чего передается. Вследствие наличия в канале шума приниматься может измененное кодовое слово (один из 2^n векторов пространства n-кортежей). Если измененный вектор не слишком отличается (лежит на небольшом расстоянии) от действительного кодового слова, декодер может обнаружить сообщение правильно. Основная задача выбора конкретной части кода подобна цели выбора семейства модулирующих сигналов, и в контексте рис. 6.10 ее можно определить следующим образом.

Рис. 6.10. Структура линейного блочного кода

1. Наполняя пространство $\mathbf{V}_n$ максимальным количеством кодовых слов, мы боремся за эффективность кодирования. Это равносильно утверждению, что мы хотим ввести лишь *небольшую избыточность* (избыток полосы).
2. Мы хотим, чтобы кодовые слова были *максимально удалены друг от друга*, так что даже если векторы будут искажены в ходе передачи, их все еще можно будет с высокой вероятностью правильно декодировать.

6.4.3. Пример линейного блочного кода (6, 3)

Приведем необходимые предварительные замечания относительно кода (6, 3). Он состоит из $2^k = 2^3 = 8$ векторов сообщений и, следовательно, восьми кодовых слов. В векторном пространстве $\mathbf{V}_6$ имеется 2^n (2^6 = шестьдесят четыре) 6-кортежей.

Нетрудно убедиться, что восемь кодовых слов, показанных в табл. 6.1, образуют в $\mathbf{V}_6$ подпространство (есть нулевой вектор, сумма любых двух кодовых слов дает кодовое слово этого же подпространства). Таким образом, эти кодовые слова представляют *линейный блочный код*, определенный в разделе 6.4.2. Может возникнуть естественный вопрос о соответствии кодовых слов и сообщений для этого кода (6, 3). Однозначного соответствия для отдельных кодов (n, k) не существует; хотя, впрочем, здесь нет полной свободы выбора. Подробнее о требованиях и ограничениях, сопровождающих разработку кода, будет рассказано в разделе 6.6.3.

Таблица 6.1. Соответствие кодовых слов и сообщений

Вектор сообщения	Кодовое слово
000	000000
100	110100
010	011010
110	101110
001	101001
101	011101
011	110011
111	000111

6.4.4. Матрица генератора

При больших k реализация *таблицы соответствия* кодера становится слишком громоздкой. Для кода (127, 92) существует 2^{92} или приблизительно 5×10^{27} кодовых векторов. Если кодирование выполняется с помощью простой таблицы соответствия, то представьте, какое количество памяти нужно для такого огромного числа кодовых слов! К счастью, задачу можно значительно упростить, по мере необходимости генерируя необходимые кодовые слова, вместо того чтобы хранить их в памяти постоянно.

Поскольку множество кодовых слов, составляющих линейный блочный код, является k-мерным подпространством n-мерного двоичного векторного пространства ($k < n$), всегда можно найти такое множество n-кортежей (с числом элементов, меньшим 2^k), которое может генерировать все 2^k кодовых слова подпространства. О генерирующем множестве векторов говорят, что оно *охватывает* подпространство. Наименьшее *линейно независимое* множество, охватывающее подпространство, называется *базисом* подпространства, а число векторов в этом базисном множестве является размерностью подпространства. Любое базисное множество k линейно независимых n-кортежей $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_k$ можно использовать для генерации нужных векторов линейного блочного кода, поскольку каждый вектор кода является линейной комбинацией $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_k$. Иными словами, каждое из множества 2^k кодовых слов $\{\mathbf{U}\}$ можно представить следующим образом.

$$\mathbf{U} = m_1 \mathbf{V}_1 + m_2 \mathbf{V}_2 + \ldots + m_k \mathbf{V}_k$$

Здесь m_i = (0 или 1) — цифры сообщения, а $i = 1, \ldots, k$.

Вообще, *матрицу генератора* можно определить как массив размером $k \times n$.

$$\mathbf{G} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \vdots \\ \mathbf{V}_{k1} \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & & & \\ v_{k1} & v_{k2} & \cdots & v_{km} \end{bmatrix} \tag{6.24}$$

Кодовые векторы принято представлять векторами-строками. Таким образом, сообщение **m** (последовательность k бит сообщения) представляется как вектор-строка (матрица $1 \times k$, в которой 1 строка и k столбцов).

$$\mathbf{m} = m_1, m_2, \ldots, m_k$$

В матричной записи генерация кодового слова **U** будет выглядеть как произведение **m** и **G**

$$\mathbf{U} = \mathbf{mG}, \tag{6.25}$$

где умножение матриц $\mathbf{C} = \mathbf{AB}$ выполняется по следующему правилу.

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \quad i = 1,\ldots,l \quad j = 1,\ldots,m$$

Здесь **A** — матрица размером $l \times n$, **B** — матрица размером $n \times m$, а результирующая матрица **C** имеет размер $l \times m$. Для примера, рассмотренного в предыдущем разделе, матрица генератора имеет следующий вид.

$$\mathbf{G} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \mathbf{V}_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{6.26}$$

Здесь $\mathbf{V}_1$, $\mathbf{V}_2$ и $\mathbf{V}_3$ — три *линейно независимых вектора* (подмножество восьми кодовых векторов), которые могут сгенерировать все кодовые векторы. Отметим, что сумма любых двух генерирующих векторов в результате не дает ни одного генерирующего вектора (противоположность свойству замкнутости). Покажем, как с использованием матрицы генератора, приведенной в выражении (6.26), генерируется кодовое слово $\mathbf{U}_4$ для четвертого вектора сообщения 110 в табл. 6.1.

$$\begin{aligned} \mathbf{U}_4 &= \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \mathbf{V}_3 \end{bmatrix} = 1 \cdot \mathbf{V}_1 + 1 \cdot \mathbf{V}_2 + 0 \cdot \mathbf{V}_3 = \\ &= 110100 + 011010 + 000000 = \\ &= 101110 \text{ (кодовое слово для вектора сообщения 1 1 0)} \end{aligned}$$

Таким образом, кодовый вектор, соответствующий вектору сообщения, является линейной комбинацией строк матрицы **G**. Поскольку код полностью определяется матрицей **G**, кодеру нужно помнить лишь k строк матрицы **G**, а не все 2^k кодовых вектора. Из приведенного примера можно видеть, что матрица генератора размерностью 3×6, приведенная в уравнении (6.26), полностью заменяет исходный массив кодовых слов размерностью 8×6, приведенный в табл. 6.1, что значительно упрощает систему.

6.4.5. Систематические линейные блочные коды

Систематический линейный блочный код (systematic linear block code) (n, k) — это такое отображение k-мерного вектора сообщения в n-мерное кодовое слово, что часть генерируемой последовательности совмещается с k символами сообщения. Остальные $(n-k)$ бит — это биты четности. Матрица генератора систематического линейного блочного кода имеет следующий вид.

$$\mathbf{G} = \left[\begin{array}{c:c} \mathbf{P} & \mathbf{I}_k \end{array}\right] = \tag{6.27}$$

$$= \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1(n-k)} & 1 & 0 & \cdots & 0 \\ p_{21} & p_{22} & \cdots & p_{2(n-k)} & 0 & 1 & \cdots & 0 \\ \vdots & & & & & & \vdots & \\ p_{k1} & p_{k2} & \cdots & p_{k(n-k)} & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Здесь $\mathbf{P}$ — массив четности, входящий в матрицу генератора, $p_{ij} = (0$ или $1)$, а $\mathbf{I}_k$ — единичная матрица размерностью $k \times k$ (у которой диагональные элементы равны 1, а все остальные — 0). Заметим, что при использовании этого систематического генератора процесс кодирования еще больше упрощается, поскольку нет необходимости хранить ту часть массива, где находится единичная матрица. Объединяя выражения (6.26) и (6.27), можно представить каждое кодовое слово в следующем виде.

$$u_1, u_2, \ldots, u_n = [m_1, m_2, \ldots, m_k] \times \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1(n-k)} & 1 & 0 & \cdots & 0 \\ p_{21} & p_{22} & \cdots & p_{2(n-k)} & 0 & 1 & \cdots & 0 \\ \vdots & & & & & & \vdots & \\ p_{k1} & p_{k2} & \cdots & p_{k(n-k)} & 0 & 0 & \cdots & 1 \end{bmatrix},$$

где

$$\begin{aligned} u_i &= m_1 p_{1i} + m_2 p_{2i} + \ldots + m_k p_{ki} && \text{для } i = 1, \ldots, (n-k) \\ &= m_{i-n+k} && \text{для } i = (n-k+1), \ldots, n \end{aligned}$$

Для данного k-кортежа сообщения

$$\mathbf{m} = m_1, m_2, \ldots, m_k$$

и k-кортежа кодовых векторов

$$\mathbf{U} = u_1, u_2, \ldots, u_k$$

систематический кодовый вектор можно записать в следующем виде.

$$U = \underbrace{p_1, p_2, \ldots, p_{n-k}}_{\text{биты четности}}, \underbrace{m_1, m_2, \ldots, m_k}_{\text{биты сообщения}}, \tag{6.28}$$

где

$$\begin{aligned} p_1 &= m_1 p_{11} + m_2 p_{21} + \ldots + m_k p_{k1} \\ p_2 &= m_1 p_{12} + m_2 p_{22} + \ldots + m_k p_{k2} \\ p_{n-k} &= m_1 p_{1(n-k)} + m_2 p_{2(n-k)} + \ldots + m_k p_{k(n-k)} \end{aligned} \tag{6.29}$$

Систематические кодовые слова иногда записываются так, чтобы биты сообщения занимали левую часть кодового слова, а биты четности — правую. Такая перестановка

не влияет на свойства кода, связанные с процедурами обнаружения и исправления ошибок, поэтому далее рассматриваться не будет.

Для кода (6, 3), рассмотренного в разделе 6.4.3, кодовое слово выглядит следующим образом.

$$\mathbf{U} = [m_1, m_2, m_3] \underbrace{\left[\begin{array}{ccc} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{array}\right.}_{\mathbf{P}} \underbrace{\left.\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right]}_{\mathbf{I}_3} = \tag{6.30}$$

$$= \underbrace{m_1 + m_3}_{u_1},\ \underbrace{m_1 + m_2}_{u_2},\ \underbrace{m_2 + m_3}_{u_3},\ \underbrace{m_1}_{u_4},\ \underbrace{m_2}_{u_5},\ \underbrace{m_3}_{u_6} \tag{6.31}$$

Выражение (6.31) позволяет получить некоторое представление о структуре линейных блочных кодов. Видно, что избыточные биты имеют разное происхождение. Первый бит четности является суммой первого и третьего битов сообщения; второй бит четности — это сумма первого и второго битов сообщения; а третий бит четности — сумма второго и третьего битов сообщения. Интуитивно понятно, что, по сравнению с контролем четности методом дублирования разряда или с помощью одного бита четности, описанная структура может предоставлять более широкие возможности обнаружения и исправления ошибок.

6.4.6. Проверочная матрица

Определим матрицу **H**, именуемую *проверочной*, которая позволит нам декодировать полученные вектора. Для каждой матрицы $(k \times n)$ генератора **G** существует матрица **H** размером $(n-k) \times n$, такая, что строки матрицы **G** ортогональны к строкам матрицы **H**. Иными словами, $\mathbf{GH}^T = \mathbf{0}$, где $\mathbf{H}^T$ — *транспонированная* матрица **H**, а **0** — нулевая матрица размерностью $k \times (n-k)$. $\mathbf{H}^T$ — это матрица размером $n \times (n-k)$, строки которой являются столбцами матрицы **H**, а столбцы — строками матрицы **H**. Чтобы матрица **H** удовлетворяла требованиям ортогональности систематического кода, ее компоненты записываются в следующем виде.

$$\mathbf{H} = \left[\mathbf{I}_{n-k} \ \vdots \ \mathbf{P}^T\right] \tag{6.32}$$

Следовательно, матрица $\mathbf{H}^T$ имеет следующий вид.

$$\mathbf{H}^T = \left[\begin{array}{c} \mathbf{I}_{n-k} \\ \cdots\cdots \\ \mathbf{P} \end{array}\right] = \tag{6.33,а}$$

$$= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 1 \\ p_{11} & p_{12} & \cdots & p_{1,(n-k)} \\ p_{21} & p_{22} & \cdots & p_{2,(n-k)} \\ \vdots & & & \\ p_{k1} & p_{12} & \cdots & p_{k,(n-k)} \end{bmatrix} \tag{6.33,б}$$

Нетрудно убедиться, что произведение $\mathbf{UH}^T$ любого кодового слова **U**, генерируемого **G**, и матрицы $\mathbf{H}^T$ дает следующее.

$$\mathbf{UH}^T = p_1 + p_1, p_2 + p_2, \ldots, p_{n-k} + p_{n-k} = \mathbf{0},$$

где биты четности $p_1, p_2, \ldots, p_{n-k}$ определены в уравнении (6.29). Таким образом, поскольку *проверочная матрица* **H** создана так, чтобы удовлетворять условиям ортогональности, она позволяет проверять принятые векторы на предмет их принадлежности заданному набору кодовых слов. **U** будет кодовым словом, генерируемым матрицей **G**, тогда и только тогда, когда $\mathbf{UH}^T = 0$.

6.4.7. Контроль с помощью синдромов

Пусть $\mathbf{r} = r_1, r_2, \ldots, r_n$ — принятый вектор (один из 2^n n-кортежей), полученный после передачи $\mathbf{U} = u_1, u_2, \ldots, u_n$ (один из 2^k n-кортежей). Тогда **r** можно представить в следующем виде.

$$\mathbf{r} = \mathbf{U} + \mathbf{e} \tag{6.34}$$

Здесь $\mathbf{e} = e_1, e_2, \ldots, e_n$ — вектор ошибки или ошибочная комбинация, внесенная каналом. Всего в пространстве из 2^n n-кортежей существует $2^n - 1$ возможных ненулевых ошибочных комбинаций. *Синдром* сигнала **r** определяется следующим образом.

$$\mathbf{S} = \mathbf{rH}^T \tag{6.35}$$

Синдром — это результат проверки четности, выполняемой над сигналом **r** для определения его принадлежности заданному набору кодовых слов. При положительном результате проверки синдром **S** равен **0**. Если **r** содержит ошибки, которые можно исправить, то синдром (как и симптом болезни) имеет определенное ненулевое значение, что позволяет отметить конкретную ошибочную комбинацию. Декодер, в зависимости от того, производит ли он прямое исправление ошибок или использует запрос ARQ, участвует в локализации и исправлении ошибки (прямое исправление ошибок) или посылает запрос на повторную передачу (ARQ). Используя уравнения (6.34) и (6.35), мы можем представить синдром **r** в следующем виде.

$$\begin{aligned} \mathbf{S} &= (\mathbf{U} + \mathbf{e})\mathbf{H}^T = \\ &= \mathbf{UH}^T + \mathbf{eH}^T \end{aligned} \tag{6.36}$$

Но для всех элементов набора кодовых слов $\mathbf{UH}^T = \mathbf{0}$. Поэтому

$$\mathbf{S} = \mathbf{eH}^T \tag{6.37}$$

Из сказанного выше очевидно, что контроль с помощью синдромов, проведенный над искаженным вектором кода или над ошибочной комбинацией, вызвавшей его появление, даст один и тот же синдром. Важной особенностью линейных блочных кодов (весьма важной в процессе декодирования) является взаимно однозначное соответствие между синдромом и исправимой ошибочной комбинацией.

Интересно также отметить два необходимых свойства проверочной матрицы.

1. В матрице **H** не может быть столбца, состоящего из одних нулей, иначе ошибка в соответствующей позиции кодового слова не отразится в синдроме и не будет обнаружена.
2. Все столбцы матрицы **H** должны быть различными. Если в матрице **H** найдется два одинаковых столбца, ошибки в соответствующих позициях кодового слова будут неразличимы.

Пример 6.3. Контроль с помощью синдромов

Пусть передано кодовое слово $\mathbf{U} = 1\ 0\ 1\ 1\ 1\ 0$ из примера в разделе 6.4.3 и принят вектор $\mathbf{r} = 0\ 0\ 1\ 1\ 1\ 0$, т.е. крайний левый бит принят с ошибкой. Нужно найти вектор синдрома $\mathbf{S} = \mathbf{r}\mathbf{H}^T$ и показать, что он равен $\mathbf{e}\mathbf{H}^T$.

Решение

$$\mathbf{S} = \mathbf{r}\mathbf{H}^T = [0\ 0\ 1\ 1\ 1\ 0]\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\\ 1 & 1 & 0\\ 0 & 1 & 1\\ 1 & 0 & 1\end{bmatrix} =$$

$$= [1,\ 1+1,\ 1+1] = [1\ \ 0\ \ 1] \text{(синдром искаженного вектора кода)}$$

Далее проверим, что синдром искаженного вектора кода равен синдрому ошибочной комбинации, которая вызвала эту ошибку.

$$\mathbf{S} = \mathbf{e}\mathbf{H}^T = [1\ 0\ 0\ 0\ 0\ 0]\ \mathbf{H}^T = [1\ 0\ 0] \text{ (синдром ошибочной комбинации)}$$

6.4.8. Исправление ошибок

Итак, мы обнаружили отдельную ошибку и показали, что контроль с помощью синдромов, выполняемый как на искаженном кодовом слове, так и на соответствующей ошибочной комбинации, дает один и тот же синдром. Этот момент является ключевым, поскольку мы имеем возможность не только определить ошибку, но и (поскольку существует взаимно однозначное соответствие между исправимой ошибочной комбинацией и синдромом) исправить подобные ошибочные комбинации. Давайте так расположим 2^n n-кортежей, которые представляют собой возможные принимаемые векторы, в так называемой *нормальной* матрице, чтобы первый ряд содержал все кодовые слова, начиная с кодового слова с одними нулями, а первый столбец — все исправимые ошибочные комбинации. Напомним, что в число основных свойств линейного кода входит то, что набор кодовых слов должен содержать член в виде вектора, состоящего из одних нулей. Каждая строка сформированной матрицы, именуемая *классом смежности*, состоит из ошибочной комбинации в первом столбце, называемой *образующим элементом класса смежности*, за которой следуют кодовые слова, подвергающиеся воздействию этой ошибочной комбинации. Нормальная матрица для кода (n, k) имеет следующий вид.

$$\begin{array}{ccccccc}
\mathbf{U}_1 & \mathbf{U}_2 & \cdots & \mathbf{U}_i & \cdots & \mathbf{U}_{2^k} \\
\mathbf{e}_2 & \mathbf{U}_2+\mathbf{e}_2 & \cdots & \mathbf{U}_i+\mathbf{e}_2 & \cdots & \mathbf{U}_{2^k}+\mathbf{e}_2 \\
\mathbf{e}_3 & \mathbf{U}_2+\mathbf{e}_3 & \cdots & \mathbf{U}_i+\mathbf{e}_3 & \cdots & \mathbf{U}_{2^k}+\mathbf{e}_3 \\
\vdots & \vdots & & & & \\
\mathbf{e}_j & \mathbf{U}_2+\mathbf{e}_j & \cdots & \mathbf{U}_i+\mathbf{e}_j & \cdots & \mathbf{U}_{2^k}+\mathbf{e}_j \\
\vdots & \vdots & & & & \\
\mathbf{e}_{2^{n-k}} & \mathbf{U}_2+\mathbf{e}_{2^{n-k}} & \cdots & \mathbf{U}_i+\mathbf{e}_{2^{n-k}} & \cdots & \mathbf{U}_{2^k}+\mathbf{e}_{2^{n-k}}
\end{array} \tag{6.38}$$

Отметим, что кодовое слово $\mathbf{U}_1$ (кодовое слово со всеми нулями) имеет два значения. Оно является кодовым словом, а также может рассматриваться как ошибочная комбинация $\mathbf{e}_1$ — комбинация, означающая отсутствие ошибки, так что $\mathbf{r} = \mathbf{U}$. Матрица содержит все 2^n n-кортежей, имеющихся в пространстве V_n. Каждый n-кортеж упомянут *только один раз*, причем ни один не пропущен и не продублирован. Каждый класс смежности содержит 2^k n-кортежей. Следовательно, всего классов смежности будет $(2^n/2^k) = 2^{n-k}$.

Алгоритм декодирования предусматривает замену искаженного вектора (любого n-кортежа, за исключением указанного в первой строке) правильным кодовым словом, указанным вверху столбца, содержащего искаженный вектор. Предположим, что кодовое слово $\mathbf{U}_i$ ($i = 1, \ldots, 2^k$) передано по каналу с помехами, в результате чего принят (искаженный) вектор $\mathbf{U}_i + \mathbf{e}_j$. Если созданная каналом ошибочная комбинация $\mathbf{e}_j$ является образующим элементом класса смежности с индексом $j = 1, \ldots, 2^{n-k}$, принятый вектор будет правильно декодирован в переданное кодовое слово $\mathbf{U}_i$. Если ошибочная комбинация не является образующим элементом класса, то декодирование даст ошибочный результат.

6.4.8.1. Синдром класса смежности

Если $\mathbf{e}_j$ является образующим элементом класса смежности или ошибочной комбинацией j-го класса смежности, то вектор $\mathbf{U}_i + \mathbf{e}_j$ является n-кортежем в этом классе смежности. Синдром этого n-кортежа можно записать в следующем виде.

$$\mathbf{S} = (\mathbf{U}_i + \mathbf{e}_j)\mathbf{H}^T = \mathbf{U}_i\mathbf{H}^T + \mathbf{e}_j\mathbf{H}^T$$

Поскольку $\mathbf{U}_i$ — это вектор кода и $\mathbf{U}_i\mathbf{H}^T = \mathbf{0}$, то, как и в уравнении (6.37), мы можем записать следующее.

$$\mathbf{S} = (\mathbf{U}_i + \mathbf{e}_j)\mathbf{H}^T = \mathbf{e}_j\mathbf{H}^T \tag{6.39}$$

Вообще, название *класс смежности* (или *сомножество*) — это сокращение от "*множество* чисел, имеющих *со*вместные свойства". Что же все-таки общего между членами каждой данной строки (класса смежности)? Из уравнения (6.39) видно, что каждый член класса смежности имеет *один и тот же синдром*. Синдром каждого класса смежности отличается от синдромов других классов смежности; именно этот синдром используется для определения ошибочных комбинаций.

6.4.8.2. Декодирование с исправлением ошибок

Процедура декодирования с исправлением ошибок состоит из следующих этапов.

1. С помощью уравнения $\mathbf{S} = \mathbf{r}\mathbf{H}^T$ вычисляется синдром $\mathbf{r}$.
2. Определяются образующие элементы класса смежности (ошибочные комбинации) $\mathbf{e}_j$, синдром которых равен $\mathbf{r}\mathbf{H}^T$.

3. Полагается, что эти ошибочные комбинации вызываются искажениями в канале.
4. Полученный исправленный вектор, или кодовое слово, определяется как $\mathbf{U} = \mathbf{r} + \mathbf{e}_j$. Можно сказать, что в результате вычитания определенных ошибок мы восстановили верное кодовое слово. (*Замечание:* в арифметических операциях по модулю 2 операция вычитания равносильна операции сложения.)

6.4.8.3. Локализация ошибочной комбинации

Возвращаясь к примеру из раздела 6.4.3, мы составляем матрицу из 2^6 = шестидесяти четырех 6-кортежей, как это показано на рис. 6.11. Правильные кодовые слова — это восемь векторов в первой строке, а *исправимые ошибочные комбинации* — это семь ненулевых *образующих элементов классов смежности* в первом столбце. Заметим, что все однобитовые ошибочные комбинации являются исправимыми. Отметим также, что после того, как исчерпываются все однобитовые ошибочные комбинации, еще остаются некоторые возможности для исправления ошибок, поскольку учтены еще не все шестьдесят четыре 6-кортежа. Имеется один образующий элемент класса смежности, с которым ничего не сопоставлено; а значит, остается возможность исправления еще одной ошибочной комбинации. Эту ошибочную комбинацию (один из n-кортежей в оставшемся образующем элементе класса смежности) можно выбрать произвольным образом. На рис. 6.11 эта последняя исправимая ошибочная комбинация выбрана равной комбинации с двумя ошибочными битами 010001. Декодирование будет правильным тогда и только тогда, когда ошибочная комбинация, введенная каналом, будет одним из образующих элементов классов смежности.

000000	110100	011010	101110	101001	011101	110011	000111
000001	110101	011011	101111	101000	011100	110010	000110
000010	110110	011000	101100	101011	011111	110001	000101
000100	110000	011110	101010	101101	011001	110111	000011
001000	111100	010010	100110	100001	010101	111011	001111
010000	100100	001010	111110	111001	001101	100011	010111
100000	010100	111010	001110	001001	111101	010011	100111
010001	100101	001011	111111	111000	001100	100010	010110

Рис. 6.11. Пример нормальной матрицы для кода (6, 3)

Определим синдром, соответствующий каждой последовательности исправимых ошибок, вычислив $\mathbf{e}_j\mathbf{H}^T$ для каждого образующего элемента.

$$\mathbf{S} = \mathbf{e}_j \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Результаты приводятся в табл. 6.2. Поскольку все синдромы в таблице различны, декодер может определить ошибочную комбинацию **e**, которой соответствует каждый синдром.

Таблица 6.2. Таблица соответствия синдромов

Ошибочная комбинация	Синдром
000000	000
000001	101
000010	011
000100	110
001000	001
010000	010
100000	100
010001	111

6.4.8.4. Пример исправления ошибки

Как говорилось в разделе 6.4.8.2, мы принимаем вектор $\mathbf{r}$ и рассчитываем его синдром с помощью выражения $\mathbf{S} = \mathbf{r}\mathbf{H}^T$. Затем, используя таблицу соответствия синдромов (табл. 6.2), составленную в предыдущем разделе, находим соответствующую ошибочную комбинацию, которая является оценкой ошибки (далее будем обозначать ее через $\hat{\mathbf{e}}$). Затем декодер прибавляет $\hat{\mathbf{e}}$ к $\mathbf{r}$ и оценивает переданное кодовое слово $\hat{\mathbf{U}}$.

$$\hat{\mathbf{U}} = \mathbf{r} + \hat{\mathbf{e}} = (\mathbf{U} + \mathbf{e}) + \hat{\mathbf{e}} = \mathbf{U} + (\mathbf{e} + \hat{\mathbf{e}}) \tag{6.40}$$

Если правильно вычислили ошибку: $\hat{\mathbf{e}} = \mathbf{e}$, тогда оценка $\hat{\mathbf{U}}$ совпадает с переданным кодовым словом $\mathbf{U}$. С другой стороны, если оценка ошибки неверна, декодер неверно определит переданное кодовое слово и мы получим *необнаружимую ошибку декодирования*.

Пример 6.4. Исправление ошибок

Пусть передано кодовое слово $\mathbf{U} = 101110$ из примера в разделе 6.4.3 и принят вектор $\mathbf{r} = 001110$. Нужно показать, как декодер, используя таблицу соответствия синдромов (табл. 6.2), может исправить ошибку.

Решение

Рассчитывается синдром $\mathbf{r}$.

$$\mathbf{S} = [0\,0\,1\,1\,1\,0]\,\mathbf{H}^T = [1\,0\,0]$$

С помощью табл. 6.2 оценивается ошибочная комбинация, соответствующая приведенному выше синдрому.

$$\hat{\mathbf{e}} = 1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0$$

Исправленный вектор равен следующему.

$$\begin{aligned}\hat{\mathbf{U}} &= \mathbf{r} + \hat{\mathbf{e}} = \\ &= 0\ \ 0\ \ 1\ \ 1\ \ 1\ \ 0 + 1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0 = \\ &= 1\ \ 0\ \ 1\ \ 1\ \ 1\ \ 0\end{aligned}$$

Поскольку оцененная ошибочная комбинация в этом примере совпадает с действительной ошибочной комбинацией, процедура исправления ошибки дает $\hat{\mathbf{U}} = \mathbf{U}$.

Можно видеть, что процесс декодирования искаженного кодового слова путем предварительного обнаружения и последующего исправления ошибки можно сравнить с аналогичной

медицинской процедурой. Пациент (потенциально искаженный вектор) приходит в медицинское учреждение (декодер). Врач проводит серию тестов (умножение на $\mathbf{H}^T$), чтобы определить симптомы болезни (синдром). Допустим, врач нашел характерные пятна на рентгенограмме пациента. Опытный врач может непосредственно установить связь между симптомом и болезнью (ошибочной комбинацией). Начинающий врач может обратиться к медицинскому справочнику (табл. 6.2) для определения соответствия между симптомом (синдромом) и болезнью (ошибочной комбинацией). Последний шаг заключается в назначении соответствующего лечения, которое устранит болезнь (уравнение (6.40)). Продолжая аналогию двоичных кодов и медицины, можно сказать, что уравнение (6.40) — это несколько необычный способ лечения. Пациент излечивается в результате повторного заболевания той же болезнью.

6.4.9. Реализация декодера

Если код небольшой, например рассмотренный ранее код (6, 3), декодер может быть реализован в виде довольно простой схемы. Рассмотрим шаги, которые должны быть предприняты декодером: (1) вычислить синдром, (2) локализовать ошибочную комбинацию и (3) осуществить сложение по модулю 2 ошибочной комбинации и принятого вектора (что приводит к устранению ошибки). В примере 6.4, имея искаженный вектор, мы покажем, как с помощью последовательности этих шагов можно получить исправленное кодовое слово. Сейчас мы рассмотрим схему, показанную на рис. 6.12, где реализованы логические элементы исключающего ИЛИ и И, которые позволяют получить тот же результат для любой *комбинации с одним ошибочным битом* в коде (6, 3). Из табл. 6.2 и уравнения (6.39) можно записать все разряды синдрома через разряды принятых кодовых слов.

$$\mathbf{S} = \mathbf{r}\mathbf{H}^T$$

$$\mathbf{S} = [r_1 \quad r_2 \quad r_3 \quad r_4 \quad r_5 \quad r_6] \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

и

$$\begin{aligned} s_1 &= r_1 + r_4 + r_6 \\ s_2 &= r_2 + r_4 + r_5 \\ s_3 &= r_3 + r_5 + r_6 \end{aligned}$$

Мы используем эти выражения для синдромов при связывании схемы на рис. 6.12. Логический элемент "исключающее ИЛИ" — это и есть реализация той самой операции сложения (или вычитания) по модулю 2, поэтому он обозначен тем же символом "+". Маленький кружок в конце каждой линии, входящей в элемент И, означает операцию логического дополнения сигнала.

Искаженный сигнал подается на декодер одновременно в верхней части схемы, где происходит вычисление синдрома, и в нижней, где синдром преобразуется в соответствующую ошибочную комбинацию. Ошибка устраняется путем повторного добавления ее к принятому вектору, что дает в итоге исправленное кодовое слово.

Заметим, что с методической точки зрения рис. 6.12 составлен так, чтобы выделить алгебраические этапы декодирования — вычисление синдрома и ошибочной комби-

нации, а также выдачу исправленных выходных данных. В реальной ситуации код (n, k) обычно конфигурируется в систематическом виде.

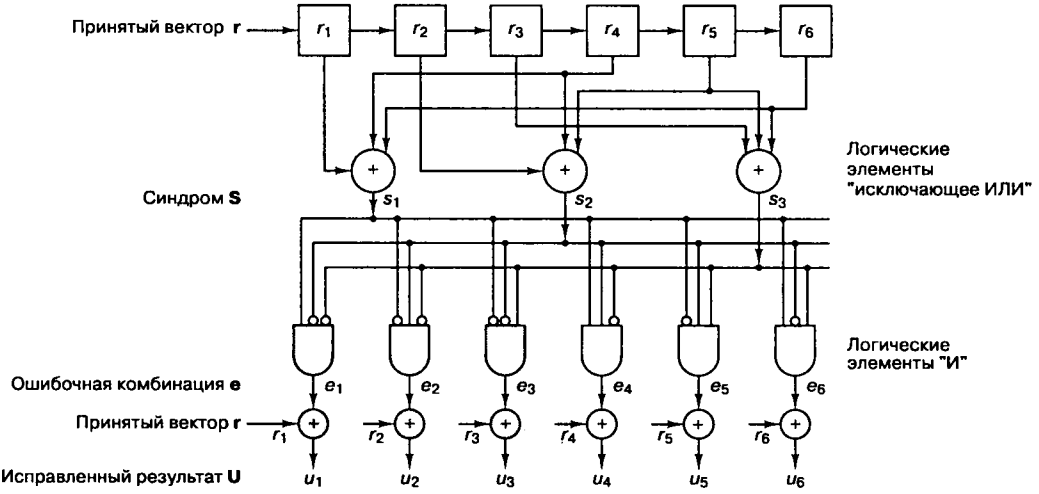


Рис. 6.12. Схема реализации декодера для кода (6, 3)

Декодеру не нужно выдавать полное кодовое слово; на выходе у него должны быть только биты данных. Поэтому схема на рис. 6.12 упрощается за счет удаления заштрихованных элементов. Для более длинных кодов такая реализация намного сложнее; в данной ситуации более предпочтительной методикой декодирования является последовательная схема, а не рассмотренный здесь параллельный метод [4]. Важно также подчеркнуть, что схема на рис. 6.12 позволяет определять и исправлять только комбинации кода (6, 3) с одним ошибочным битом. Исправление комбинаций с двумя ошибочными битами потребует дополнительной схемы.

6.4.9.1. Векторные обозначения

Выше кодовые слова, ошибочные комбинации, принятые векторы и синдромы обозначались как векторы **U**, **e**, **r** и **S**. Для упрощения записи индексы, сопутствующие конкретному вектору, в основном, опускались. Хотя, если быть точным, каждый из векторов **U**, **e**, **r** и **S** должен записываться в следующем виде.

$$\mathbf{x}_j = \{x_1, x_2, \ldots, x_i, \ldots\}$$

Рассмотрим диапазон индексов j и i в контексте кода (6, 3), приведенного в табл. 6.1. Для кодового слова $\mathbf{U}_j$ индекс $j = 1, \ldots, 2^k$ показывает, что имеется $2^3 = 8$ отдельных кодовых слов, а индекс $i = 1, \ldots, n$ демонстрирует, что каждое кодовое слово составлено из $n = 6$ бит. Для исправимой ошибочной комбинации $\mathbf{e}_j$ индекс $j = 1, \ldots, 2^{n-k}$ означает, что имеется $2^3 = 8$ образующих элементов классов смежности (7 ненулевых исправимых ошибочных комбинаций), а индекс $i = 1, \ldots, n$ указывает, что каждая ошибочная комбинация составлена из $n = 6$ бит. Для принятого вектора $\mathbf{r}_j$ индекс $j = 1, \ldots, 2^n$ показывает, что имеется $2^6 = 64$ n-кортежей, прием которых возможен, а индекс $i = 1, \ldots, n$ указывает, что каждый принятый n-кортеж состоит из $n = 6$ бит. И наконец, для синдрома $\mathbf{S}_j$ индекс $j = 1, \ldots, n - k$ означает, что каждый синдром состоит из $n - k = 3$ бит. В этой главе индексы часто опускаются, и векторы $\mathbf{U}_j$, $\mathbf{e}_j$, $\mathbf{r}_j$ и $\mathbf{S}_j$ зачастую обозначаются

как **U**, **e**, **r** и **S**. Читателю следует помнить, что для этих векторов индексы всегда подразумеваются, даже в тех случаях, когда они опущены для простоты записи.

6.5. Возможность обнаружения и исправления ошибок

6.5.1. Весовой коэффициент двоичных векторов и расстояние между ними

Конечно же, понятно, что правильно декодировать можно не все ошибочные комбинации. Возможности кода для исправления ошибок в первую очередь определяются его структурой. *Весовой коэффициент Хэмминга* (Hamming weight) $w(\mathbf{U})$ кодового слова **U** определяется как число ненулевых элементов в **U**. Для двоичного вектора это эквивалентно числу единиц в векторе. Например, если $\mathbf{U} = 1\,0\,0\,1\,0\,1\,1\,0\,1$, то $w(\mathbf{U}) = 5$. *Расстояние Хэмминга* (Hamming distance) между двумя кодовыми словами **U** и **V**, обозначаемое как $d(\mathbf{U}, \mathbf{V})$, определяется как количество элементов, которыми они отличаются.

$$\begin{aligned} \mathbf{U} &= 1\,0\,0\,1\,0\,1\,1\,0\,1 \\ \mathbf{V} &= 0\,1\,1\,1\,1\,0\,1\,0\,0 \\ d(\mathbf{U}, \mathbf{V}) &= 6 \end{aligned}$$

Согласно свойствам сложения по модулю 2, можно отметить, что сумма двух двоичных векторов является другим двоичным вектором, двоичные единицы которого расположены на тех позициях, которыми эти векторы отличаются.

$$\mathbf{U} + \mathbf{V} = 1\,1\,1\,0\,1\,1\,0\,0\,1$$

Таким образом, можно видеть, что расстояние Хэмминга между двумя векторами равно весовому коэффициенту Хэмминга их суммы, т.е. $d(\mathbf{U}, \mathbf{V}) = w(\mathbf{U} + \mathbf{V})$. Также видно, что весовой коэффициент Хэмминга кодового слова равен его расстоянию Хэмминга до нулевого вектора.

6.5.2. Минимальное расстояние для линейного кода

Рассмотрим множество расстояний между всеми парами кодовых слов в пространстве $\mathbf{V}_n$. Наименьший элемент этого множества называется *минимальным расстоянием* кода и обозначается $d_{\min}$. Как вы думаете, почему нас интересует именно минимальное расстояние, а не максимальное? Минимальное расстояние подобно наиболее слабому звену в цепи, оно дает нам меру минимальных возможностей кода и, следовательно, характеризует его мощность.

Как обсуждалось ранее, сумма двух произвольных кодовых слов дает другой элемент пространства кодовых слов. Это свойство линейных кодов формулируется просто: если **U** и **V** — кодовые слова, то и $\mathbf{W} = \mathbf{U} + \mathbf{V}$ тоже должно быть кодовым словом. Следовательно, расстояние между двумя кодовыми словами равно весовому коэффициенту третьего кодового слова, т.е. $d(\mathbf{U}, \mathbf{V}) = w(\mathbf{U} + \mathbf{V}) = w(\mathbf{W})$. Таким образом, минимальное расстояние линейного кода можно определить, не прибегая к изучению расстояний между всеми комбинациями пар кодовых слов. Нам нужно лишь определить вес каждого кодового слова (за исключением нулевого вектора) в подпространстве; минимальный вес соответствует минимальному расстоянию $d_{\min}$. Иными словами, $d_{\min}$ соответствует наименьшему из множества расстояний между нулевым кодовым словом и всеми остальными кодовыми словами.

6.5.3. Обнаружение и исправление ошибок

Задача декодера после приема вектора $\mathbf{r}$ заключается в оценке переданного кодового слова $\mathbf{U}_i$. Оптимальная стратегия декодирования может быть выражена в терминах алгоритма *максимального правдоподобия* (см. приложение Б); считается, что передано было слово $\mathbf{U}_i$, если

$$P(\mathbf{r}\,|\,\mathbf{U}_i) = \max_{\text{по всем } \mathbf{U}_j} P(\mathbf{r}\,|\,\mathbf{U}_j). \tag{6.41}$$

Поскольку для двоичного симметричного канала (binary symmetric channel — BSC) правдоподобие $\mathbf{U}_i$ относительно $\mathbf{r}$ обратно пропорционально расстоянию между $\mathbf{r}$ и $\mathbf{U}_i$, можно сказать, что передано было слово $\mathbf{U}_i$, если

$$d(\mathbf{r}\,|\,\mathbf{U}_i) = \min_{\text{по всем } \mathbf{U}_j} d(\mathbf{r}\,|\,\mathbf{U}_j). \tag{6.42}$$

Другими словами, декодер определяет расстояние между $\mathbf{r}$ и всеми возможными переданными кодовыми словами $\mathbf{U}_j$, после чего выбирает наиболее правдоподобное $\mathbf{U}_i$, для которого

$$d(\mathbf{r},\mathbf{U}_i) \le d(\mathbf{r},\mathbf{U}_j) \quad \text{для } i,j = 1,\ldots,M \quad \text{и} \quad i \ne j, \tag{6.43}$$

где $M = 2^k$ — это размер множества кодовых слов. Если минимум не один, выбор между минимальными расстояниями является произвольным. Наше обсуждение метрики расстояний будет продолжено в главе 7.

На рис. 6.13 расстояние между двумя кодовыми словами $\mathbf{U}$ и $\mathbf{V}$ показано как *расстояние Хэмминга*. Каждая черная точка обозначает искаженное кодовое слово. На рис. 6.13, *а* проиллюстрирован прием вектора $\mathbf{r}_1$, находящегося на расстоянии 1 от кодового слова $\mathbf{U}$ и на расстоянии 4 от кодового слова $\mathbf{V}$. Декодер с коррекцией ошибок, следуя стратегии максимального правдоподобия, выберет при принятом векторе $\mathbf{r}_1$ кодовое слово $\mathbf{U}$. Если $\mathbf{r}_1$ получился в результате появления одного ошибочного бита в переданном векторе кода $\mathbf{U}$, декодер успешно исправит ошибку. Но если же это произошло в результате 4-битовой ошибки в векторе кода $\mathbf{V}$, декодирование будет ошибочным. Точно так же, как показано на рис. 6.13, *б*, двойная ошибка при передаче $\mathbf{U}$ может привести к тому, что в качестве переданного вектора будет ошибочно определен вектор $\mathbf{r}_2$, находящийся на расстоянии 2 от вектора $\mathbf{U}$ и на расстоянии 3 от вектора кода $\mathbf{V}$. На рис. 6.13 показана ситуация, когда в качестве переданного вектора ошибочно определен вектор $\mathbf{r}_3$, который находится на расстоянии 3 от вектора кода $\mathbf{U}$ и на расстоянии 2 от вектора $\mathbf{V}$. Из рис. 6.13 видно, что если задача состоит только в обнаружении ошибок, а не в их исправлении, то можно определить искаженный вектор — изображенный черной точкой и представляющий одно-, двух-, трех- и четырехбитовую ошибку. В то же время пять ошибок при передаче могут привести к приему кодового слова $\mathbf{V}$, когда в действительности было передано кодовое слово $\mathbf{U}$; такую ошибку невозможно будет *обнаружить*.

Из рис. 6.13 можно видеть, что способность кода к обнаружению и исправлению ошибок связана с *минимальным расстоянием* между кодовыми словами. Линия решения на рисунке служит той же цели, что и в процессе демодуляции, — для разграничения областей решения.

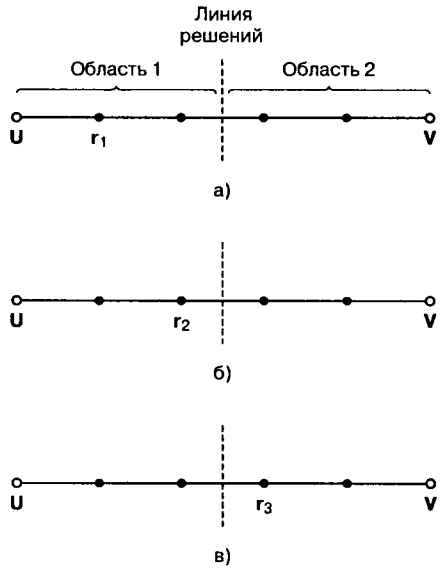


Рис. 6.13. Возможности определения и исправления ошибок: а) принятый вектор $\mathbf{r}_1$*; б) принятый вектор* $\mathbf{r}_2$*; в) принятый вектор* $\mathbf{r}_3$

В примере, приведенном на рис. 6.13, критерий принятия решения может быть следующим: выбрать **U**, если **r** попадает в область 1, и выбрать **V**, если **r** попадает в область 2. Выше показывалось, что такой код (при $d_{\min} = 5$) может исправить две ошибки. Вообще, *способность кода к исправлению ошибок* t определяется, как максимальное число гарантированно исправимых ошибок на кодовое слово, и записывается следующим образом [4].

$$t = \left\lfloor \frac{d_{\min} - 1}{2} \right\rfloor \tag{6.44}$$

Здесь $\lfloor x \rfloor$ означает наибольшее целое, не превышающее x. Часто код, который исправляет все искаженные символы, содержащие ошибку в t или меньшем числе бит, также может исправлять символы, содержащие $t+1$ ошибочных бит. Это можно увидеть на рис. 6.11. В этом случае $d_{\min} = 3$, поэтому из уравнения (6.44) можно видеть, что исправимы *все* ошибочные комбинации из $t = 1$ бит. Также исправима *одна* ошибочная комбинация, содержащая $t+1$ (т.е. 2) ошибочных бит. Вообще, линейный код (n, k), способный исправлять все символы, содержащие t ошибочных бит, может исправить всего 2^{n-k} ошибочных комбинаций. Если блочный код с возможностью исправления символов, имеющих ошибки в t бит, применяется для исправления ошибок в двоичном симметричном канале с вероятностью перехода p, то вероятность ошибки сообщения P_M (вероятность того, что декодер совершит неправильное декодирование и n-битовый блок содержит ошибку) можно оценить сверху, используя уравнение (6.18).

$$P_M \le \sum_{n=t+1}^{n} \binom{n}{j} p^j (1-p)^{n-j} \tag{6.45}$$

Оценка переходит в равенство, если декодер исправляет все ошибочные комбинации, содержащие до t ошибочных бит включительно, но не комбинации с числом ошибочных бит, большим t. Такие декодеры называются *декодерами с ограниченным расстоянием*. Вероятность ошибки в декодированном бите P_B зависит от конкретного кода и декодера. Приближенно ее можно выразить следующим образом [5].

$$P_B \approx \frac{1}{n} \sum_{n=t+1}^{n} \binom{n}{j} p^j (1-p)^{n-j} \tag{6.46}$$

В блочном коде, прежде чем исправлять ошибки, необходимо их обнаружить. (Или же код может использоваться только для определения наличия ошибок.) Из рис. 6.13 видно, что любой полученный вектор, который изображается черной точкой (искаженное кодовое слово), можно определить как ошибку. Следовательно, возможность определения наличия ошибки дается следующим выражением.

$$E = d_{\min} - 1 \tag{6.47}$$

Блочный код с минимальным расстоянием $d_{\min}$ гарантирует обнаружение всех ошибочных комбинаций, содержащих $d_{\min} - 1$ или меньшее число ошибочных бит. Такой код также способен обнаружить и большую ошибочную комбинацию, содержащую $d_{\min}$ или более ошибок. Фактически код (n, k) может обнаружить $2^n - 2^k$ ошибочных комбинаций длины n. Объясняется это следующим образом. Всего в пространстве 2^n n-кортежей существует $2^n - 1$ возможных ненулевых ошибочных комбинаций. Даже правильное кодовое слово — это потенциальная ошибочная комбинация. Поэтому всего существует $2^k - 1$ ошибочных комбинаций, которые идентичны $2^k - 1$ ненулевым кодовым словам. При появлении любая из этих $2^k - 1$ ошибочных комбинаций изменяет передаваемое кодовое слово $\mathbf{U}_i$ на другое кодовое слово $\mathbf{U}_j$. Таким образом, принимается кодовое слово $\mathbf{U}_j$, и его синдром равен нулю. Декодер принимает $\mathbf{U}_j$ за переданное кодовое слово, и поэтому декодирование даст неверный результат. Следовательно, существует $2^k - 1$ необнаружимых ошибочных комбинаций. Если ошибочная комбинация не совпадает с одним из 2^k кодовых слов, проверка вектора $\mathbf{r}$ с помощью синдромов дает ненулевой синдром и ошибка успешно обнаруживается. Отсюда следует, что существует ровно $2^n - 2^k$ выявляемых ошибочных комбинаций. При больших n, когда $2^k << 2^n$, необнаружимой будет только незначительная часть ошибочных комбинаций.

6.5.3.1. Распределение весовых коэффициентов кодовых слов

Пусть A_j — количество кодовых слов с весовым коэффициентом j в линейном коде (n, k). Числа $A_0, A_1, \ldots, A_n$ называются *распределением весовых коэффициентов* этого кода. Если код применяется только для обнаружения ошибок в двоичном симметричном канале, то вероятность того, что декодер не сможет определить ошибку, можно рассчитать, исходя из распределения весовых коэффициентов кода [5].

$$P_{\text{nd}} = \sum_{j=1}^{n} A_j p^j (1-p)^{n-j}, \tag{6.48}$$

где p — вероятность перехода в двоичном симметричном канале. Если минимальное расстояние кода равно $d_{\min}$, значения от A_1 до $A_{d_{\min}-1}$ равны нулю.

Пример 6.5. Вероятность необнаруженной ошибки в коде

Пусть код (6, 3), введенный в разделе 6.4.3, используется только для обнаружения наличия ошибок. Рассчитайте вероятность необнаруженной ошибки, если применяется двоичный симметричный канал, а вероятность перехода равна 10^{-2}.

Решение

Распределение весовых коэффициентов этого кода выглядит следующим образом: $A_0 = 1$, $A_1 = A_2 = 0$, $A_3 = 4$, $A_5 = 0$, $A_6 = 0$. Следовательно, используя уравнение (6.48), можно записать следующее.

$$P_{nd} = 4p^3(1-p)^3 + 3p^4(1-p)^2$$

Для $p = 10^{-2}$ вероятность необнаруженной ошибки будет равна $3{,}9 \times 10^{-6}$.

6.5.3.2. Одновременное обнаружение и исправление ошибок

Возможностями исправления ошибок с максимальным гарантированным (t), где t определяется уравнением (6.44), можно пожертвовать в пользу определения класса ошибок. Код можно использовать для одновременного исправления α и обнаружения β ошибок, причем $\alpha \leq \beta$, а минимальное расстояние кода дается следующим выражением [4].

$$d_{min} \geq \alpha + \beta + 1 \tag{6.49}$$

При появлении t или меньшего числа ошибок код способен обнаруживать и исправлять их. Если ошибок больше t, но меньше $e+1$, где e определяется уравнением (6.47), код может определять наличие ошибок, но исправить может только некоторые из них. Например, используя код с $d_{min} = 7$, можно выполнить обнаружение и исправление со следующими значениями α и β.

Обнаружение (β)	Исправление (α)
3	3
4	2
5	1
6	0

Заметим, что исправление ошибки подразумевает ее предварительное обнаружение. В приведенном выше примере (с тремя ошибками) все ошибки можно обнаружить и исправить. Если имеется пять ошибок, их можно обнаружить, но исправить можно только одну из них.

6.5.4. Визуализация пространства 6-кортежей

На рис. 6.14 визуально представлено восемь кодовых слов, фигурирующих в примере из раздела 6.4.3. Кодовые слова образованы посредством линейных комбинаций из трех независимых 6-кортежей, приведенных в уравнении (6.26); сами кодовые слова образуют трехмерное подпространство. На рисунке показано, что такое подпространство полностью занято восемью кодовыми словами (большие черные круги); координаты подпространства умышленно выбраны неортогональными. На рис. 6.14 предпринята попытка изобразить все пространство, содержащее шестьдесят четыре 6-кортежей, хотя точно нарисовать или составить такую модель невозможно. Каждое кодовое слово окружают сферические слои или оболочки. Радиус внутренних непересекающихся слоев — это расстояние Хэмминга, равное 1; радиус внешнего слоя — это расстояние Хэмминга, равное 2. Большие расстоя-

ния в этом примере не рассматриваются. Для каждого кодового слова два показанных слоя заняты искаженными кодовыми словами. На каждой внутренней сфере существует шесть таких точек (всего 48 точек), представляющих шесть возможных однобитовых ошибок в векторах, соответствующих каждому кодовому слову. Эти кодовые слова с однобитовыми возмущениями могут быть соотнесены только с одним кодовым словом; следовательно, такие ошибки могут быть исправлены. Как видно из нормальной матрицы, приведенной на рис. 6.11, существует также одна двухбитовая ошибочная комбинация, которая также поддается исправлению. Всего существует $\binom{6}{2} = 15$ разных двухбитовых ошибочных комбинаций, которыми может быть искажено любое кодовое слово, но исправить можно только одну из них (в нашем примере это ошибочная комбинация 0 1 0 0 0 1). Остальные четырнадцать двухбитовых ошибочных комбинаций описываются векторами, которые нельзя однозначно сопоставить с каким-либо одним кодовым словом; эти не поддающиеся исправлению ошибочные комбинации дают векторы, которые эквивалентны искаженным векторам двух или большего числа кодовых слов. На рисунке все (56) исправимые кодовые слова с одно- и двухбитовыми искажениями показаны маленькими черными кругами. Искаженные кодовые слова, не поддающиеся исправлению, представлены маленькими прозрачными кругами.

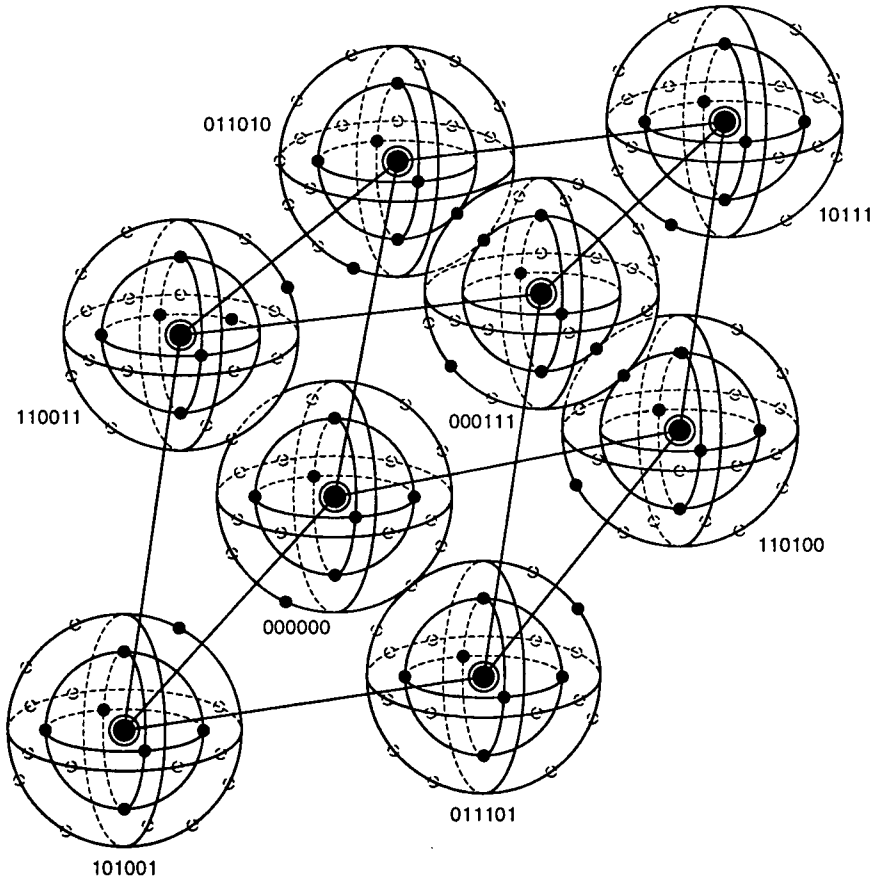


Рис. 6.14. Пример восьми кодовых слов в пространстве 6-кортежей

При представлении свойств класса кодов, известных как *совершенные коды* (perfect code), рис. 6.14 весьма полезен. Код, исправляющий ошибки в t битах, называется совершенным, если нормальная матрица содержит все ошибочные комбинации из t или меньшего числа ошибок и не содержит иных образующих элементов классов смежности (отсутствует возможность исправления остаточных ошибок). В контексте рис. 6.14 совершенный код с коррекцией ошибок в t битах — это такой код, который (при использовании обнаружения по принципу максимального правдоподобия) может исправить все искаженные кодовые слова, находящиеся на расстоянии Хэмминга t (или ближе) от исходного кодового слова, и не способен исправить ни одну из ошибок, находящихся на расстоянии, превышающем t.

Кроме того, рис. 6.14 способствует пониманию основной цели поиска хороших кодов. Предпочтительным является пространство, максимально заполненное кодовыми словами (эффективное использование введенной избыточности), а также желательно, чтобы кодовые слова были по возможности максимально удалены друг от друга. Очевидно, что эти цели противоречивы.

6.5.5. Коррекция со стиранием ошибок

Приемник можно сконструировать так, чтобы он объявлял символ *стертым*, если последний принят неоднозначно либо обнаружено наличие помех или кратковременных сбоев. Размер входного алфавита такого канала равен Q, а выходного — $Q+1$; лишний выходной символ называется *меткой стирания* (erasure flag), или просто *стиранием* (erasure). Если демодулятор допускает символьную ошибку, то для ее исправления необходимы два параметра, определяющие ее *расположение* и *правильное* значение символа. В случае двоичных символов эти требования упрощаются — нам необходимо только расположение ошибки. В то же время, если демодулятор объявляет символ *стертым* (при этом правильное значение символа неизвестно), расположение этого символа известно, поэтому декодирование стертого кодового слова может оказаться проще исправления ошибки. Код защиты от ошибок можно использовать для исправления стертых символов или одновременного исправления ошибок и стертых символов. Если минимальное расстояние кода равно $d_{\min}$, любая комбинация из ρ или меньшего числа стертых символов может быть исправлена при следующем условии [6].

$$d_{\min} \geq \rho + 1 \tag{6.50}$$

Предположим, что ошибки появляются вне позиций стирания. Преимущество исправления посредством стираний качественно можно выразить так: если минимальное расстояние кода равно $d_{\min}$, согласно уравнению (6.50), можно восстановить $d_{\min} - 1$ стирание. Поскольку число ошибок, которые можно исправить без стирания информации, не превышает $(d_{\min} - 1)/2$, то преимущество исправления ошибок посредством стираний очевидно. Далее, любую комбинацию из α ошибок и γ стираний можно исправить одновременно, если, как показано в работе [6],

$$d_{\min} \geq 2\alpha + \gamma + 1. \tag{6.51}$$

Одновременное исправление ошибок и стираний можно осуществить следующим образом. Сначала позиции из γ стираний замещаются нулями, и получаемое кодовое слово декодируется обычным образом. Затем позиции из γ стираний замещаются единицами, и декодирование повторяется для этого варианта кодового слова. После об-

работки обоих кодовых слов (одно с подставленными нулями, другое — с подставленными единицами) выбирается то из них, которое соответствует наименьшему числу ошибок, исправленных вне позиций стирания. Если удовлетворяется неравенство (6.51), то описанный метод всегда дает верное декодирование.

Пример 6.6. Коррекция со стиранием ошибок

Рассмотрим набор кодовых слов, представленный в разделе 6.4.3.

000000 110100 011010 101110 101001 011101 110011 000111

Пусть передано кодовое слово 110011, в котором два крайних слева разряда приемник объявил стертыми. Проверьте, что поврежденную последовательность xx0011 можно исправить.

Решение

Поскольку $d_{\min} = \rho + 1 = 3$, код может исправить $\rho = 2$ стирания. В этом легко убедиться из рис. 6.11 или приведенного выше перечня кодовых слов, сравнивая 4 крайних правых разряда xx0011 с каждым из допустимых кодовых слов. Действительно переданное кодовое слово — это ближайшее (с точки зрения расстояния Хэмминга) к искаженной последовательности.

6.6. Полезность нормальной матрицы

6.6.1. Оценка возможностей кода

Нормальную матрицу можно представлять как организационный инструмент, картотеку, содержащую все возможные 2^n записи в пространстве n-кортежей, в которой ничего не упущено и не продублировано. На первый взгляд может показаться, что выгода от использования этого инструмента ограничена *малыми* блочными кодами, поскольку для кодов длиной более $n = 20$ пространство n-кортежей насчитывает миллионы элементов. Впрочем, даже для больших кодов нормальная матрица позволяет определить важные исходные характеристики, такие как возможные компромиссы между обнаружением и исправлением ошибок и пределы возможностей кода в коррекции ошибок. Одно из таких ограничений, называемое *пределом Хэмминга* [7], описывается следующим образом.

$$\text{Количество бит четности: } n - k \geq \log_2 \left[1 + \binom{n}{1} + \binom{n}{2} + \ldots + \binom{n}{t} \right] \qquad (6.52,\text{а})$$

или

$$\text{Количество классов смежности: } 2^{n-k} \geq \left[1 + \binom{n}{1} + \binom{n}{2} + \ldots + \binom{n}{t} \right] \qquad (6.52,\text{б})$$

Здесь величина $\binom{n}{j}$, определяемая уравнением (6.16), представляет число способов выбора из n бит j ошибочных. Заметим, что сумма членов уравнения (6.52), находящихся в квадратных скобках, дает минимальное количество строк, которое должно присутствовать в нормальной матрице для исправления всех комбинаций ошибок, вплоть до t-битовых ошибок. Неравенство определяет нижнюю границу числа $n - k$ бит четности (или 2^{n-k} классов смежности) как функцию возможностей кода в коррекции t-битовых ошибок. Аналогичным образом можно сказать, что неравенство дает верхнюю границу возможностей кода в коррекции t-битовых ошибок как

функцию числа $n-k$ бит четности (или 2^{n-k} классов смежности). Для обеспечения возможности коррекции t-битовых ошибок произвольных линейных блочных кодов (n, k) необходимым условием является удовлетворение предела Хэмминга.

Чтобы показать, как нормальная матрица может обеспечить визуальное представление этого предела, возьмем в качестве примера код БХЧ (127, 106). Матрица содержит все $2^k = 2127 \approx 1{,}70 \times 10^{38}$ n-кортежей пространства. Верхняя строка матрицы содержит $2^k = 2106 \approx 8{,}11 \times 10^{31}$ кодовых слов; следовательно, это число столбцов в матрице. Крайний левый столбец содержит $2^{n-k} = 2^{21} = 2\,097\,152$ образующих элемента классов смежности; следовательно, это количество строк в матрице. Несмотря на то что число n-кортежей и кодовых слов просто огромно, нас не интересует конкретный вид каждого элемента матрицы. Основной интерес представляет количество классов смежности. Существует 2 097 152 класса смежности и, следовательно, 2 097 151 ошибочная комбинация, которую способен исправить этот код. Далее показано, каким образом это число классов смежности определяет верхний предел возможностей кода в коррекции t-битовых ошибок.

Поскольку каждое кодовое слово содержит 127 бит, существует 127 возможностей допустить ошибку в одном бите. Рассчитываем количество возможностей появления двух ошибок — $\binom{127}{2} = 8\,001$. Затем переходим к трехбитовым ошибкам, поскольку ошибки, упомянутые выше, — это лишь незначительная часть всех 2 097 151 ошибочных комбинаций. Итак, существует $\binom{127}{3} = 333\,375$ возможностей совершить трехбитовую ошибку. Эти расчеты приведены в табл. 6.3; там же показано, что нулевая ошибочная комбинация требует наличия первого класса смежности. Затем перечислены требования одно-, двух- и трехбитовых ошибок. Также показывается количество классов смежности, необходимое для коррекции каждого типа ошибок, и общее количество классов смежности, необходимых для коррекции ошибок всех типов, вплоть до требуемого типа ошибки. Из этой таблицы можно видеть, что код (127, 106) способен исправить все комбинации, содержащие 1, 2 или 3 ошибочных бита, причем это составляет только 341 504 из 2 097 152 возможных классов смежности. Неиспользованные 1 755 648 строк говорят о больших потенциальных возможностях в коррекции ошибок, чем было использовано. Действительно, в матрицу можно попытаться втиснуть все возможные 4-битовые ошибки. Но при взгляде на табл. 6.3 становится совершенно ясно, что это невозможно, поскольку, как показывает последняя строка таблицы, число оставшихся в матрице классов смежности значительно меньше общего числа классов смежности, требуемого для коррекции 4-битовых ошибок. Следовательно, предел Хэмминга описанного кода (127, 106) гарантирует исправление всех ошибок вплоть до 3-битовых.

Таблица 6.3. Предел возможностей коррекции для кода (127, 106)

Количество битовых ошибок	Количество необходимых классов смежности	Общее число необходимых классов смежности
0	1	1
1	127	128
2	8001	8129
3	333375	341504
4	10334625	10676129

6.6.2. Пример кода (n, k)

Нормальная матрица дает возможность взглянуть на возможные компромиссы между исправлением и обнаружением ошибок. Рассмотрим пример кода (n, k) и факторы, определяющие выбор конкретных значений (n, k).

1. Для получения нетривиального соотношения между исправлением и обнаружением ошибок желательно, чтобы код имел возможности коррекции ошибок, по крайней мере, с $t = 2$. Согласно уравнению (6.44), минимальное расстояние при этом равно $d_{\min} = 2t + 1 = 5$.
2. Чтобы кодовая система была нетривиальной, желательно, чтобы количество бит данных было не менее $k = 2$. Следовательно, число кодовых слов $2^k = 4$. Далее будем считать наш код следующим: (n, 2).
3. Нас интересует минимальное значение n, которое позволит исправлять все одно- и двухбитовые ошибки. В этом примере каждый из 2^n n-кортежей в матрице будет табулирован. Минимальное значение n нас интересует потому, что при каждом увеличении n на единицу число n-кортежей в нормальной матрице удваивается. Это условие, разумеется, диктуется только соображениями удобства использования таблицы. Для реальных прикладных кодов минимальное значение n выбирается по разным причинам — эффективность использования полосы пропускания и простота системы. Если при выборе n используется предел Хэмминга, то n следует выбрать равным 7. В то же время размерность полученного кода (7, 2) не соответствует указанным выше требованиям $t = 2$ и $d_{\min} = 5$. Чтобы увидеть это, следует ввести другую верхнюю границу возможностей кода в коррекции t-битовых ошибок (или $d_{\min}$). Эта граница, называемая *предел Плоткина* [7], определяется следующим образом.

$$d_{\min} \le \frac{n \times 2^{k-1}}{2^k - 1} \tag{6.54}$$

Вообще, линейный код (n, k) должен удовлетворять всем перечисленным выше условиям, включая возможности коррекции ошибок (или минимальное расстояние). Для высокоскоростных кодов из удовлетворения предела Хэмминга следует удовлетворение предела Плоткина; это справедливо, например, для рассмотренного ранее кода (127, 106). Для кодов с низкими скоростями передачи существует обходной путь удовлетворения названных требований [7]. Поскольку в нашем примере речь идет именно о таких кодах, важно оценить их возможности в коррекции ошибок с помощью предела Плоткина. Поскольку $d_{\min} = 5$, из уравнения (6.53) получаем, что n должно быть равно 8; следовательно, для удовлетворения всех требований, поставленных в этом примере, минимальная размерность кода равна (8, 2). Можно ли практически использовать подобный код (8, 2)? Этого делать не стоит, поскольку это потребует слишком большой полосы пропускания; лучше выбрать более эффективный код. Данный код мы используем только с методической целью, единственным его преимуществом являются удобные размеры его нормальной матрицы.

6.6.3. Разработка кода (8, 2)

Ответ на вопрос, *как выбираются кодовые слова из пространства 2^8 8-кортежей*, неоднозначен, хотя определенные ограничения выбора все же существуют. Ниже перечислены некоторые моменты, которые могут указать наилучшее решение.

1. Количество кодовых слов $2^k = 2^2 = 4$.
2. Среди кодовых слов должен быть нулевой вектор.
3. Следует учесть свойство замкнутости — сумма двух любых кодовых слов в пространстве должна давать кодовое слово из этого же пространства.
4. Каждое кодовое слово содержит 8 двоичных разрядов.
5. Поскольку $d_{min} = 5$, весовой коэффициент каждого кодового слова (за исключением нулевого) также должен быть не менее 5 (в силу свойства замкнутости). Весовой коэффициент вектора определяется как число ненулевых компонентов этого вектора.
6. Предположим, что код является систематическим; значит, 2 крайних правых бита каждого кодового слова являются соответствующими битами сообщения.

Далее предлагается вариант набора кодовых слов, удовлетворяющих всем перечисленным выше требованиям.

Сообщения	Кодовые слова
00	00000000
01	11110001
10	00111110
11	11001111

Создание набора кодовых слов может выполняться совершенно произвольно; нужно только неуклонно следовать свойствам весовых коэффициентов и придерживаться систематической формы кода. Выбор первых нескольких кодовых слов обычно очень прост. Далее процесс, как правило, усложняется и возможность выбора все больше ограничивается за счет свойства замкнутости.

6.6.4. Соотношение между обнаружением и исправлением ошибок

Для кодовой системы (8, 2), выбранной в предыдущем разделе, матрицу генератора $(k \times n) = (2 \times 8)$ можно записать в следующем виде.

$$\mathbf{G} = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Декодирование начинается с расчета синдрома, что можно представлять как изучение "симптомов" ошибки. Для кода (n, k) $(n-k)$-битовый синдром $\mathbf{S}$ является произведением принятого n-битового вектора $\mathbf{r}$ и транспонированной проверочной матрицы $\mathbf{H}$ размерностью $(n-k) \times n$. Проверочная матрица $\mathbf{H}$ построена таким образом, что строки матрицы $\mathbf{G}$ ортогональны строкам матрицы $\mathbf{H}$, т.е. $\mathbf{GH}^T = \mathbf{0}$. В нашем примере кода (8, 2) $\mathbf{S}$ — это 6-битовый вектор, а $\mathbf{H}$ — матрица размером 6×8, где

$$\mathbf{H}^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

Синдром для каждой ошибочной комбинации можно рассчитать, исходя из уравнения (6.37), а именно

$$\mathbf{S}_i = \mathbf{r}_i\mathbf{H}^T \quad i = 1, \ldots, 2^{n-k},$$

где $\mathbf{S}_i$ — один из $2^{n-k} = 64$ синдромов, а $\mathbf{e}_i$ — один из 64 образующих элементов классов смежности (ошибочных комбинаций) в нормальной матрице. На рис. 6.15, помимо самой нормальной матрицы, показаны все 64 синдрома для кода (8, 2). Набор синдромов рассчитывался с помощью уравнения (6.37); позиции произвольной строки (смежный класс) нормальной матрицы имеют один и тот же синдром. Исправление искаженного кодового слова осуществляется путем расчета его синдрома и локализации ошибочной комбинации, соответствующей этому синдрому. В заключение ошибочная комбинация прибавляется (по модулю 2) к поврежденному кодовому слову, что и дает правильное кодовое слово. Из уравнения (6.49), повторно приведенного ниже, видно, что между возможностями обнаружения и исправления ошибок существует некий компромисс, ограничиваемый расстоянием.

$$d_{\min} \geq \alpha + \beta + 1$$

Здесь α представляет количество исправляемых битовых ошибок, а β — количество обнаруживаемых битовых ошибок, причем $\beta \geq \alpha$. В коде (8, 2) возможны следующие компромиссы между этими двумя величинами.

Обнаружение (β)	Исправление (α)
2	2
3	1
4	0

Из данной таблицы видно, что код (8, 2) можно использовать только для исправления ошибок; это означает, что код вначале обнаруживает $\beta = 2$ ошибки, после чего они исправляются. Если пожертвовать возможностью исправления и использовать код для исправления только однобитовых ошибок, то возможность обнаружения ошибки возрастает до $\beta = 3$ ошибок. И наконец, если целиком отказаться от исправления ошибок, то декодер сможет обнаруживать ошибки с $\beta = 4$. В случае, если ошибки только обнаруживаются, реализация декодера будет очень простой: производится вычисление синдрома и обнаруживается ошибка при появлении любого ненулевого синдрома.

Для исправления однобитовых ошибок декодер может реализовываться с логическими элементами [4], подобными приведенным на рис. 6.12, где принятый вектор кода $\mathbf{r}$ поступал в схему в двух точках. В верхней части рисунка принятые символы поступают на логический элемент исключающего ИЛИ, который и определяет синдром. Для любого принятого вектора синдром рассчитывается согласно уравнению (6.35).

$$\mathbf{S}_i = \mathbf{r}_i\mathbf{H}^T \quad i = 1, \ldots, 2^{n-k}$$

Синдромы			Нормальная матрица		
000000	1.	00000000	11110001	00111110	11001111
111100	2.	00000001	11110000	00111111	11001110
001111	3.	00000010	11110011	00111100	11001101
000001	4.	00000100	11110101	00111010	11001011
000010	5.	00001000	11111001	00110110	11000111
000100	6.	00010000	11100001	00101110	11011111
001000	7.	00100000	11010001	00011110	11101111
010000	8.	01000000	10110001	01111110	10001111
100000	9.	10000000	01110001	10111110	01001111
110011	10.	00000011	11110010	00111101	11001100
111101	11.	00000101	11110100	00111011	11001010
111110	12.	00001001	11111000	00110111	11000110
111000	13.	00010001	11100000	00101111	11011110
110100	14.	00100001	11010000	00011111	11101110
101100	15.	01000001	10110000	01111111	10001110
011100	16.	10000001	01110000	10111111	01001110
001110	17.	00000110	11110111	00111000	11001001
001101	18.	00001010	11111011	00110100	11000101
001011	19.	00010010	11100011	00101100	11011101
000111	20.	00100010	11010011	00011100	11101101
011111	21.	01000010	10110011	01111100	10001101
101111	22.	10000010	01110011	10111100	01001101
000011	23.	00001100	11111101	00110010	11000011
000101	24.	00010100	11100101	00101010	11011011
001001	25.	00100100	11010101	00011010	11101011
010001	26.	01000100	10110101	01111010	10001011
100001	27.	10000100	01110101	10111010	01001011
000110	28.	00011000	11101111	00100110	11010111
001010	29.	00101000	11011001	00010110	11100111
010010	30.	01001000	10111001	01110110	10000111
100010	31.	10001000	01111001	10110110	01000111
001100	32.	00110000	11000001	00001110	11111111
010100	33.	01010000	10100001	01101110	10011111
100100	34.	10010000	01100001	10101110	01011111
011000	35.	01100000	10010001	01011110	10101111
101000	36.	10100000	01010001	10011110	01101111
110000	37.	11000000	00110001	11111110	00001111
110010	38.	00000111	11100010	00111001	11101000
110111	39.	00010011	11100010	00101101	11011100
111011	40.	00100011	11010010	00011101	11101100
100011	41.	01000011	10110010	01111101	10001100
010011	42.	10000011	01110010	10111101	01001100
111111	43.	00001101	11111100	00110011	11000010
111001	44.	00010101	11100100	00101011	11011010
110101	45.	00100101	11010100	00011011	11101010
101101	46.	01000101	10110100	01111011	10001010
011101	47.	10000101	01110100	10111011	01001010
011110	48.	01000110	10110111	01111000	10001001
101110	49.	10000110	01110111	10111000	01001001
100101	50.	10010100	01100101	10101010	01011011
011001	51.	01100100	10010101	01011010	10101011
110001	52.	11000100	00110101	11111010	00001011
011010	53.	01101000	10011001	01010110	10100111
010110	54.	01011000	10101001	01100110	10010111
100110	55.	10011000	01101001	10100110	01010111
101010	56.	10101000	01011001	10010110	01100111
101001	57.	10100100	01010101	10011010	01101011
100111	58.	10100010	01010011	10011100	01101101
010111	59.	01100010	10010011	01011100	10101101
010101	60.	01010100	10100101	01101010	10011011
011011	61.	01010010	10100011	01101100	10011101
110110	62.	00101001	11011000	00010111	11100110
111010	63.	00011001	11101000	00100111	11010110
101011	64.	10010010	01100011	10101100	01011101

Рис. 6.15. Синдромы и нормальная матрица для кода (8, 2)

С помощью значений $\mathbf{H}^T$ для кода (8, 2), необходимо так соединить элементы схемы (подобно тому, как это было сделано на рис. 6.12), чтобы вычислялось следующее.

$$\mathbf{S}_i = [r_1 \quad r_2 \quad r_3 \quad r_4 \quad r_5 \quad r_6 \quad r_7 \quad r_8] \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

Каждая из цифр s_j ($j = 1, \ldots, 6$) определяет синдром $\mathbf{S}_i$ ($i = 1, \ldots, 64$), связанный с входным принятым вектором кода следующим образом.

$$s_1 = r_1 + r_8 \qquad s_2 = r_2 + r_8 \qquad s_3 = r_3 + r_7 + r_8$$
$$s_4 = r_4 + r_7 + r_8 \qquad s_5 = r_5 + r_7 \qquad s_6 = r_6 + r_7$$

Для реализации схемы декодера для кода (8, 2), подобной представленной на рис. 6.12, необходимо, чтобы восемь принятых разрядов соединялись с шестью сумматорами по модулю 2 (см. выше), выдающими цифры синдрома. Соответственно, потребуются и другие модификации схемы, приведенной на рисунке.

Если декодер реализован так, чтобы исправлять только однобитовые ошибки (т.е. $\alpha = 1$ и $\beta = 3$), это эквивалентно ограничению матрицы на рис. 6.15 девятью первыми классами смежности, а исправление ошибок происходит, только когда один из восьми синдромов соответствует появлению однобитовой ошибки. Затем схема декодирования (подобная изображенной на рис. 6.12) преобразует синдром в соответствующую ошибочную комбинацию. Далее ошибочная комбинация прибавляется по модулю 2 к "потенциально" искаженному принятому вектору, т.е. происходит исправление ошибки. Для проверки ситуаций, когда синдром не равен нулю, а схемы коррекции нет, нужно вводить дополнительные логические элементы (например, для однобитовых ошибок, соответствующих синдромам 10–64).

Если декодер реализован так, чтобы исправлять одно- и двухбитовые ошибки (а это означает, что обнаруживается, а затем исправляется $\beta = 2$ ошибки), это эквивалентно ограничению матрицы (рис. 6.15) 37 классом смежности. Хотя код (8, 2) может исправлять некоторые комбинации трехбитовых ошибок, соответствующие образующим элементам классов смежности под номерами 38–64, декодер чаще всего реализуется как *декодер с ограниченным расстоянием*; это означает, что он исправляет все искаженные символы, содержащие ошибку только в t или меньшем числе бит. Нереализованные возможности используются для некоторого улучшения процесса обнаружения ошибок. Как и ранее, реализация декодера подобна схеме, изображенной на рис. 6.12.

6.6.5. Взгляд на код сквозь нормальную матрицу

В контексте рис. 6.15 код (8, 2) удовлетворяет пределу Хэмминга. Иными словами, из нормальной матрицы можно видеть, что код (8, 2) способен исправлять все комбинации одно- и двухбитовых ошибок. Рассмотрим следующее: пусть передача происходит

по каналу, который всегда вносит ошибки в виде пакета 3-битовых ошибок, и, следовательно, нет необходимости в исправлении одно- или двухбитовых ошибок. Можно ли настроить образующие элементы классов смежности так, чтобы они соответствовали только трехбитовым ошибкам? Нетрудно определить, что в последовательности из 8 бит существует $\binom{8}{3} = 56$ возможностей произвести трехбитовую ошибку. Если единственным нашим желанием является коррекция только этих 56 комбинаций трехбитовых ошибок, то кажется, что в нормальной матрице достаточно места (достаточное количество классов смежности), поскольку всего в ней имеется 64 строки. Будет ли это работать? Однозначно, нет. Для любого кода главным параметром, определяющим способности кода к коррекции ошибок, является d_{min}. Для кода (8, 2) $d_{min} = 5$, а это означает, что возможно исправление только 2-битовых ошибок.

Как стандартная матрица может помочь разобраться, почему эта схема не будет работать? Чтобы осуществить исправление x-битовых ошибок для группы x-битовых ошибочных комбинаций, полная группа векторов с весовым коэффициентом x должна быть классом смежности, т.е. они должны находиться только в крайнем левом столбце. На рис. 6.15 можно видеть, что все векторы с весовым коэффициентом 1 и 2 находятся в крайнем левом столбце нормальной матрицы и нигде более. Если мы даже и втиснем все векторы с весовым коэффициентом 3 в строку со второго номера по 57-й, увидим, что некоторые из этих векторов снова появятся в матрице где-нибудь еще (что нарушит основное свойство нормальной матрицы). На рис. 6.15 затененные блоки обозначают те 56 векторов, которые имеют весовой коэффициент 3. Взгляните на образующие элементы классов смежности, представляющие 3-битовые ошибочные комбинации, в строках 38, 41–43, 46–49 и 52 нормальной матрицы. Потом посмотрите на позиции в тех же строках в крайнем правом столбце, где затененные блоки показывают другие векторы с весовым коэффициентом 3. Видите некоторую неопределенность, существующую для каждой строки, о которых говорилось выше, и понятно ли теперь, почему нельзя исправить все 3-битовые ошибочные комбинации с помощью кода (8, 2)? Допустим, декодер принял вектор с весовым коэффициентом 3 — 1 1 0 0 1 0 0 0, размещенный в строке 38 в крайнем правом столбце. Это искаженное кодовое слово могло появиться, во-первых, при передаче кодового слова 1 1 0 0 1 1 1 1, искаженного 3-битовой ошибочной комбинацией 0 0 0 0 0 1 1 1, а во-вторых, при передаче кодового слова 0 0 0 0 0 0 0 0, искаженного 3-битовой ошибочной комбинацией 1 1 0 0 1 0 0 0.

6.7. Циклические коды

Важным подклассом линейных блочных кодов являются двоичные циклические коды (cyclic codes). Код легко реализуется на регистре сдвига с обратной связью; на подобных регистрах сдвига с обратной связью вычисляется синдром; алгебраическая структура циклического кода естественным образом позволяет эффективно реализовать методы декодирования. Итак, линейный код (n, k) называется *циклическим*, если он обладает следующим свойством. Если n-кортеж $\mathbf{U} = (u_0, u_1, u_2, \ldots, u_{n-1})$ является кодовым словом в подпространстве **S**, тогда $\mathbf{U}(1) = (u_{n-1}, u_0, u_1, u_2, \ldots, u_{n-1})$, полученный из **U** с помощью циклического сдвига, также является кодовым словом в **S**. Или, вообще, $\mathbf{U}(i) = (u_{n-i}, u_{n-i+1}, \ldots, u_{n-1}, u_0, u_1, \ldots, u_{n-i-1})$, полученный i циклическими сдвигами, является кодовым словом в **S**.

Компоненты кодового слова $\mathbf{U} = (u_0, u_1, u_2, \ldots, u_{n-1})$ можно рассматривать как коэффициенты полинома $\mathbf{U}(X)$.

$$\mathbf{U}(X) = u_0 + u_1X + u_2X^2 + \ldots + u_{n-1}X^{n-1} \tag{6.54}$$

Полиномиальную функцию $\mathbf{U}(X)$ можно рассматривать как "заполнитель" разрядов кодового слова $\mathbf{U}$, т.е. вектор n-кортежа описывается полиномом степени $n-1$ или меньше. Наличие или отсутствие каких-либо членов в полиноме означает наличие 1 или 0 в соответствующем месте n-кортежа. Если u_{n-1}-й компонент отличен от нуля, порядок полинома равен $n-1$. Удобство такого полиномиального представления кодового слова станет более понятным по мере дальнейшего обсуждения алгебраических свойств циклических кодов.

6.7.1. Алгебраическая структура циклических кодов

В кодовых словах, выраженных в полиномиальной форме, циклическая природа кода проявляется следующим образом. Если $\mathbf{U}(X)$ является кодовым словом, представленным полиномом порядка $(n-1)$, то $\mathbf{U}^{(i)}(X)$ — остаток от деления $X^i\mathbf{U}(X)$ на X^n+1 — также является кодовым словом.
Иными словами,

$$\frac{X^i\mathbf{U}(X)}{X^n+1} = \mathbf{q}(X) + \frac{\mathbf{U}^{(i)}(X)}{X^n+1} \tag{6.55,а}$$

или, умножая обе части уравнения на X^n+1,

$$X^i\mathbf{U}(X) = \mathbf{q}(X)(X^n+1) + \underbrace{\mathbf{U}^{(i)}(X)}_{\text{остаток}}, \tag{6.55,б}$$

что в модульной арифметике можно описать следующим образом.

$$\mathbf{U}^{(i)}(X) = X^i\mathbf{U}(X) \text{ по модулю } (X^n+1) \tag{6.56}$$

Здесь "x по модулю y" означает остаток от деления x на y. Ниже справедливость выражения (6.56) демонстрируется для случая $i=1$.

$$\mathbf{U}(X) = u_0 + u_1X + u_2X^2 + \ldots + u_{n-2}X^{n-2} + u_{n-1}X^{n-1}$$

$$X\mathbf{U}(X) = u_0X + u_1X^2 + u_2X^3 + \ldots + u_{n-2}X^{n-1} + u_{n-1}X^n$$

К последнему выражению прибавим и вычтем u_{n-1} или, поскольку мы пользуемся арифметическими операциями по модулю 2, можем прибавить u_{n-1} дважды.

$$X\mathbf{U}(X) = \underbrace{u_{n-1} + u_0X + u_1X^2 + u_2X^3 + \cdots + u_{n-2}X^{n-1}}_{\mathbf{U}^{(1)}(X)} + u_{n-1}X^n + u_{n-1} = \mathbf{U}^{(1)}(X) + u_{n-1}(X^n+1)$$

Поскольку порядок $\mathbf{U}^{(1)}(X)$ равен $n-1$, этот полином не делится на X^n+1. Таким образом, используя уравнение (6.55,а), можно записать следующее.

$$\mathbf{U}^{(1)}(X) = X\mathbf{U}(X) \text{ по модулю } (X^n+1)$$

Обобщая, приходим к уравнению (6.56).

$$\mathbf{U}^{(i)}(X) = X^i\mathbf{U}(X) \text{ по модулю } (X^n+1)$$

Пример 6.7. Циклический сдвиг вектора кода

Пусть $\mathbf{U} = 1\ 1\ 0\ 1$ для $n = 4$. Выразите кодовое слово в полиномиальной форме и, используя уравнение (6.56), выполните третий циклический сдвиг кодового слова.

Решение

$\mathbf{U}(X) = 1 + X + X^3$ полином записан в порядке возрастания степени

$X^i\mathbf{U}(X) = X^3 + X^4 + X^6$, где $i = 3$

Разделим $X^3\mathbf{U}(X)$ на $X^4 + 1$ и найдем остаток, используя полиномиальное деление.

$$
\begin{array}{l|l}
X^6 + X^4 + X^3 & X^4 + 1 \\
\hline
X^6 \qquad\quad + X^2 & X^2 + 1 \\
\hline
\quad X^4 + X^3 + X^2 & \\
\quad X^4 \qquad\quad + 1 & \\
\hline
\qquad X^3 + X^2 + 1 & \text{остаток } \mathbf{U}^{(3)}(X)
\end{array}
$$

Записываем остаток в порядке возрастания степеней: $1 + X^2 + X^3$, кодовое слово $\mathbf{U}^{(3)} = 1\ 0\ 1\ 1$ представляет собой три циклических сдвига $\mathbf{U} = 1\ 1\ 0\ 1$. Напомним, что для двоичных кодов операция сложения выполняется по модулю 2, так что $+1 = -1$, и, следовательно, в расчетах знаки "минус" не отражены.

6.7.2. Свойства двоичного циклического кода

С помощью *полиномиального генератора* можно создать циклический код, почти так же как создавались блочные коды с использованием матрицы генератора. Полиномиальный генератор $\mathbf{g}(X)$ для циклического кода (n, k) является единственным и имеет следующий вид.

$$\mathbf{g}(X) = g_0 + g_1X + g_2X^2 + \ldots + g_pX^p \tag{6.57}$$

Здесь g_0 и g_p должны быть равны 1. Каждый полином кодового слова в подпространстве имеет вид $\mathbf{U}(X) = \mathbf{m}(X)\mathbf{g}(X)$, где $\mathbf{U}(X)$ — полином степени $n - 1$ или меньше. Следовательно, полином сообщения $\mathbf{m}(X)$ будет иметь следующий вид.

$$\mathbf{m}(X) = m_0 + m_1X + m_2X^2 + \ldots + m_{n-p-1}X^{n-p-1} \tag{6.58}$$

Всего в коде (n, k) существует 2^{n-p} полинома кодовых слов и 2^k вектора кода. Поэтому на каждый вектор кода должен приходиться один полином кодового слова.

$$n - p = k$$

или

$$p = n - k$$

Отсюда следует, что $\mathbf{g}(X)$, как показано в уравнении (6.57), должен иметь степень $n - k$, и каждый полином кодового слова в коде (n, k) можно выразить следующим образом.

$$\mathbf{U}(X) = (m_0 + m_1X + m_2X^2 + \ldots + m_{n-1}X^{n-1})\,\mathbf{g}(X) \tag{6.59}$$

$\mathbf{U}$ будет считаться действительным кодовым словом из подпространства $\mathbf{S}$ *тогда и только тогда*, когда $\mathbf{U}(X)$ делится на $\mathbf{g}(X)$ без остатка.

Полиномиальный генератор $\mathbf{g}(X)$ циклического кода (n, k) является множителем $X^n + 1$, т.е. $X^n + 1 = \mathbf{g}(X)\mathbf{h}(X)$. Например,

$$X^7 + 1 = (1 + X + X^3)(1 + X + X^2 + X^4)$$

Используя $\mathbf{g}(X) = 1 + X + X^3$ как полиномиальный генератор степени $n - k = 3$, можно получить циклический код $(n, k) = (7, 4)$. Или же с помощью $\mathbf{g}(X) = 1 + X + X^2 + X^4$, где $n - k = 4$, можно получить циклический код (7, 3). Итак, если $\mathbf{g}(X)$ является полиномом степени $n - k$ и множителем $X^n + 1$, то $\mathbf{g}(X)$ однозначным образом генерирует циклический код (n, k).

6.7.3. Кодирование в систематической форме

В разделе 6.4.5 мы ввели понятие *систематическая форма* и рассмотрели уменьшение сложности, которое делает эту форму кодирования более привлекательной. Теперь мы хотим использовать некоторые алгебраические свойства циклического кода для развития процедуры систематического кодирования. Итак, вектор сообщения можно записать в полиномиальной форме следующим образом.

$$\mathbf{m}(X) = m_0 = m_1 X + m_2 X^2 + \ldots + m_{k-1} X^{k-1} \tag{6.60}$$

В систематической форме символы сообщения используются как часть кодового слова. Мы можем сдвинуть символы сообщения в k крайних правых разряда кодового слова, а затем прибавить биты четности, разместив их в крайние левые $n - k$ разряды. Таким образом, осуществляется алгебраическая манипуляция полиномом сообщения, и он оказывается сдвинутым вправо на $n - k$ позиций. Если теперь умножить $\mathbf{m}(X)$ на X^{n-k}, мы получим сдвинутый вправо полином сообщения.

$$X^{n-k}\mathbf{m}(X) = m_0 X^{n-k} + m_1 X^{n-k+1} + \ldots + m_{k-1} X^{n-1} \tag{6.61}$$

Если далее разделить уравнение (6.61) на $\mathbf{g}(X)$, результат можно представить в следующем виде.

$$X^{n-k}\mathbf{m}(X) = \mathbf{q}(X)\mathbf{g}(X) + \mathbf{p}(X) \tag{6.62}$$

Здесь остаток $\mathbf{p}(X)$ записывается следующим образом.

$$\mathbf{p}(X) = p_0 + p_1 X + p_2 X^2 + \ldots + p_{n-k-1} X^{n-k-1}$$

Также можно записать следующее.

$$\mathbf{p}(X) = X^{n-k}\mathbf{m}(X) \text{ по модулю } \mathbf{g}(X) \tag{6.63}$$

Прибавляя $\mathbf{p}(X)$ к обеим частям уравнения (6.62) и используя сложение по модулю 2, получаем следующее.

$$\mathbf{p}(X) + X^{n-k}\mathbf{m}(X) = \mathbf{q}(X)\mathbf{g}(X) = \mathbf{U}(X) \tag{6.64}$$

Левая часть уравнения (6.64) является действительным полиномом кодового слова, так как это полином степени $n - 1$ или менее, который при делении на $\mathbf{g}(X)$ дает нулевой остаток. Это кодовое слово можно записать через все члены полинома.

$$\mathbf{p}(X) + X^{n-k}\mathbf{m}(X) = p_0 + p_1 X + p_2 X^2 + \ldots + p_{n-k-1} X^{n-k-1} + m_0 X^{n-k} + m_1 X^{n-k+1} + \ldots + m_{k-1} X^{n-1}$$

Полином кодового слова соответствует вектору кода.

$$\mathbf{U} = \underbrace{(p_0, p_1, \ldots, p_{n-k-1}}_{\substack{n-k \\ \text{бит четности}}}, \underbrace{m_0, m_1, \ldots, m_{k-1})}_{\substack{k \\ \text{бит сообщения}}} \tag{6.65}$$

Пример 6.8. Циклический код в систематической форме

С помощью полиномиального генератора $\mathbf{g}(X) = 1 + X + X^3$ получите систематическое кодовое слово из набора кодовых слов (7, 4) для вектора сообщения $\mathbf{m} = 1\ 0\ 0\ 1\ 1$.

Решение

$$\mathbf{m}(X) = 1 + X^2 + X^3,\ n = 7,\ k = 4,\ n - k = 3;$$

$$X^{n-k}\mathbf{m}(X) = X^3(1 + X^2 + X^3) = X^3 + X^5 + X^6$$

Разделив $X^{n-k}\mathbf{m}(X)$ на $\mathbf{g}(X)$, можно записать следующее.

$$X^3 + X^5 + X^6 = \underbrace{(1 + X + X^2 + X^3)}_{\text{частное } \mathbf{q}(X)}\ \underbrace{(1 + X + X^3)}_{\text{генератор } \mathbf{g}(X)} + \underbrace{1}_{\text{остаток } \mathbf{p}(X)}$$

Используя уравнение (6.64), получаем следующее.

$$\mathbf{U}(X) = \mathbf{p}(X) + X^3\mathbf{m}(X) = 1 + X^3 + X^5 + X^6$$

$$\mathbf{U} = \underbrace{1\ 0\ 0}_{\text{биты четности}}\ \underbrace{1\ 0\ 1\ 1}_{\text{биты сообщения}}$$

6.7.4. Логическая схема для реализации полиномиального деления

Выше показывалось, что при циклическом сдвиге полинома кодового слова и кодировании полинома сообщения применяется операция деления полиномов друг на друга. Такие операции легко реализуются в *схеме деления* (регистр сдвига с обратной связью). Итак, пусть даны два полинома $\mathbf{V}(X)$ и $\mathbf{g}(X)$, где

$$\mathbf{V}(X) = v_0 + v_1X + v_2X^2 + \ldots + v_mX^m$$

и

$$\mathbf{g}(X) = g_0 + g_1X + g_2X^2 + \ldots + g_pX^p,$$

причем $m \geq p$. Схема деления, приведенная на рис. 6.16, выполняет полиномиальное деление $\mathbf{V}(X)$ на $\mathbf{g}(X)$, определяя, таким образом, частное и остаточное слагаемое.

$$\frac{\mathbf{V}(X)}{\mathbf{g}(X)} = \mathbf{q}(X) + \frac{\mathbf{p}(X)}{\mathbf{g}(X)}$$

В исходном состоянии разряды регистра содержат нули. Коэффициенты $\mathbf{V}(X)$ поступают и продвигаются по регистру сдвига по одному за такт, начиная с коэффициентов более высокого порядка. После p-го сдвига частное на выходе равно $g_p^{-1}v_m$; это слагаемое наивысшего порядка в частном. Далее для каждого коэффициента частного q_i из делимого нужно вычитать полином $q_ig(X)$. Это вычитание обеспечивает обратная связь, отображенная на рис. 6.16. Разность крайних слева p слагаемых остается в делимом, а слагаемое обратной связи $q_ig(X)$ формируется при каждом сдвиге схемы и отображается в виде содержимого регистра. При каждом сдвиге регистра разность смещается на один разряд; слагаемое наивысшего порядка (которое по построению

схемы равно нулю) удаляется, в то время как следующий значащий коэффициент в $\mathbf{V}(X)$ перемещается на его место. После всех $m+1$ сдвигов регистра, на выход последовательно выдается частное, а остаток остается в регистре.

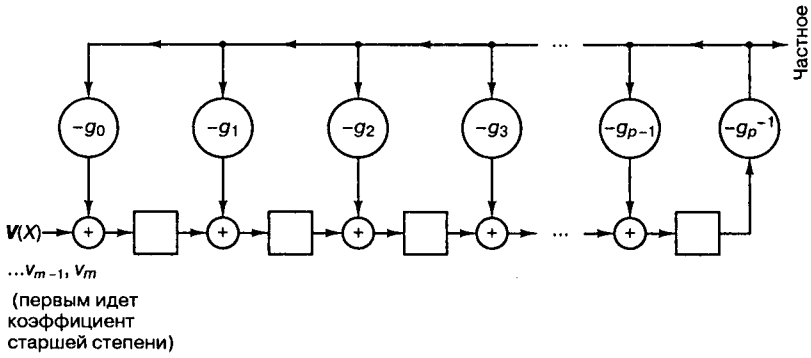


Рис. 6.16. Логическая схема для реализации полиномиального деления

Пример 6.9. Схема полиномиального деления

Используя схему деления, показанную на рис. 6.16, разделите $\mathbf{V}(X) = X^3 + X^5 + X^6$ ($\mathbf{V} = 0\,0\,0\,1\,0\,1\,1$) на $\mathbf{g}(X) = (1 + X + X^3)$. Найдите частное и остаточное слагаемое. Сравните реализацию схемы и действия, происходящие при прямом делении полиномов.

Решение

Схема деления должна выполнить следующее действие.

$$\frac{X^3 + X^5 + X^6}{1 + X + X^3} = \mathbf{q}(X) + \frac{\mathbf{p}(X)}{1 + X + X^3}$$

Необходимый регистр сдвига с обратной связью показан на рис. 6.17. Предположим, что первоначально регистр содержит нули. Схема выполнит следующие шаги.

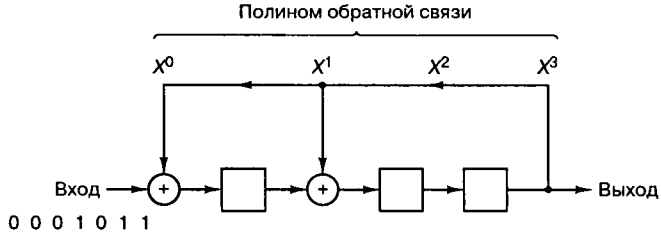


Рис. 6.17. Схема деления для примера 6.9

Входная очередь	Номер сдвига	Содержимое регистра	Выход и обратная связь
0001011	0	000	–
000101	1	100	0
00010	2	110	0
0001	3	011	0
000	4	011	1
00	5	111	1
0	6	101	1
	7	100	1

После четвертого сдвига коэффициенты частного $\{q_i\}$, последовательно поступающие с выхода, выглядят как 1 1 1 1 или же полином частного имеет вид $\mathbf{q}(X) = 1 + X + X^2 + X^3$. Коэффициенты остатка $\{p_i\}$ имеют вид 1 0 0 либо полином остатка имеет вид $\mathbf{p}(X) = 1$. Таким образом, схема выполнила следующие вычисления.

$$\frac{X^3 + X^5 + X^6}{1 + X + X^3} = 1 + X + X^2 + X^3 + \frac{1}{1 + X + X^3}$$

Прямое деление полиномов дает результат, показанный ниже.

$X^6 + X^5 + X^3 \quad | X^3 + X + 1$

обратная связь после 4-го сдвига → $X^6 + X^4 + X^3$ $\quad X^3 + X^2 + X + 1$

регистр после 4-го сдвига → $X^5 + X^4$ $\quad \uparrow \ \uparrow \ \uparrow \ \uparrow$

обратная связь после 5-го сдвига → $X^5 + X^3 + X^2$ $\quad 4 \ \ 5 \ \ 6 \ \ 7$

регистр после 5-го сдвига → $X^4 + X^3 + X^2$

обратная связь после 6-го сдвига → $X^4 + X^2 + X$

регистр после 6-го сдвига → $X^3 + X$

обратная связь после 7-го сдвига → $X^3 + X + 1$

регистр после 7-го сдвига → 1

(остаток)

6.7.5. Систематическое кодирование с ($n - k$)-разрядным регистром сдвига

Как было показано в разделе 6.7.3, кодирование с помощью циклического кода в систематической форме включает в себя, как результат вычисления $X^{n-k}\mathbf{m}(X)$ по модулю $\mathbf{g}(X)$, вычисление битов четности; иными словами, *деление смещенного вверх* (смещенного вправо) полиномиального сообщения на полиномиальный генератор $\mathbf{g}(X)$. Сдвиг вверх приводит к освобождению места для битов четности, которые прибавляются к разрядам сообщения, что в результате дает вектор кода в систематической форме. Сдвиг вверх на $n - k$ разрядов сообщения является тривиальной операцией и в действительности не выполняется в схеме деления. На самом деле вычисляются только биты четности; затем они помещаются на соответствующие места рядом с битами сообщения. Полином четности — это остаток от деления на полиномиальный генератор; он находится в регистре n сдвигов $(n - k)$-разрядного регистра сдвига с обратной связью, показанного на рис. 6.17. Отметим, что первые $n - k$ сдвигов по разрядам — это просто заполнение регистра. У нас не может появиться никакой обратной связи, пока не будет заполнен крайний справа разряд; следовательно, мы можем сократить цикл деления, загружая входящие данные с выхода последнего разряда, как показано на рис. 6.18. Слагаемое обратной связи в крайнем левом разряде является суммой входящих данных и крайнего правого разряда. Гарантия создания этой суммы — обеспечение $g_0 = g_{n-k} = 1$ для произвольного полиномиального генератора $\mathbf{g}(X)$. Соединения схемы обратной связи соответствуют коэффициентам полиномиального генератора, которые записываются в следующем виде.

$$\mathbf{g}(X) = 1 + g_1 X + g_2 X^2 + \ldots + g_{n-k-1} X^{n-k-1} + X^{n-k} \tag{6.66}$$

Следующие шаги описывают процедуру кодирования, использующую устройство, изображенное на рис. 6.18.

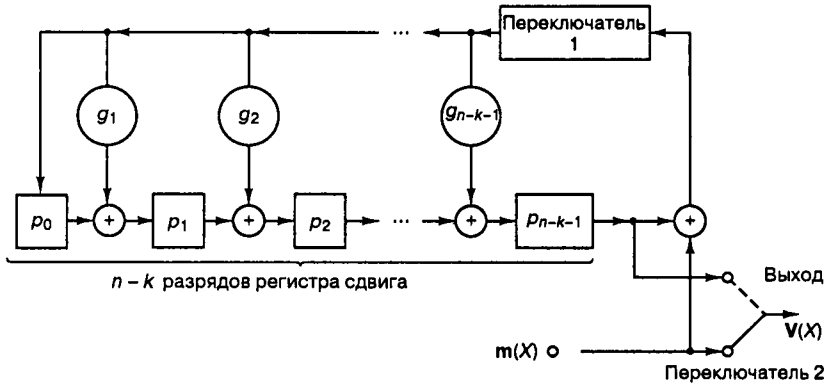


Рис. 6.18. Кодирование с помощью $(n-k)$-разрядного регистра сдвига

1. При первых k сдвигах ключ 1 закрыт для передачи битов сообщения в $(n-k)$-разрядный регистр сдвига.
2. Ключ 2 установлен в нижнее положение для передачи битов сообщения на выходной регистр в течение первых k сдвигов.
3. После передачи k-го бита сообщения ключ 1 открывается, а ключ 2 переходит в верхнее положение.
4. При остальных $n-k$ сдвигах происходит очищение кодирующих регистров, биты четности перемещаются на выходной регистр.
5. Общее число сдвигов равно n, и содержимое выходного регистра представляет собой полином кодового слова $\mathbf{p}(X) + X^{n-k}\mathbf{m}(X)$.

Пример 6.10. Систематическое кодирование циклического кода

Используя регистр сдвига с обратной связью, показанный на рис. 6.18, кодируйте вектор сообщения $\mathbf{m} = 1\ 0\ 1\ 1$ в кодовое слово (7, 4). Полиномиальный генератор $\mathbf{g}(X) = 1 + X + X^3$.

Решение

$$\mathbf{m} = 1\ 0\ 1\ 1$$

$$\mathbf{m}(X) = 1 + X^2 + X^3$$

$$X^{n-k}\mathbf{m}(X) = X^3\mathbf{m}(X) = X^3 + X^5 + X^6$$

$$X^{n-k}\mathbf{m}(X) = \mathbf{q}(X)\mathbf{g}(X) + \mathbf{p}(X)$$

$$\mathbf{p}(X) = (X^3 + X^5 + X^6) \text{ по модулю } (1 + X + X^3)$$

Для $(n-k) = 3$-разрядного регистра сдвига, показанного на рис. 6.19, действия будут следующими.

Входная очередь	Номер сдвига	Содержимое регистра	Выход и обратная связь
1011	0	000	–
101	1	110	1
10	2	101	1
1	3	100	0
	4	100	1

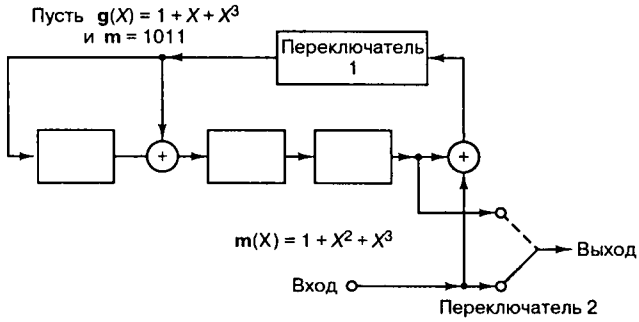


Рис. 6.19. Пример кодирования циклического кода (7, 4) с помощью (n – k)-разрядного регистра сдвига

После четвертого сдвига ключ 1 открывается, ключ 2 переходит в верхнее положение, а биты четности переходят в выходной регистр. Выходное кодовое слово **U** = 1 0 0 1 0 1 1 или, в полиномиальной форме, $\mathbf{U}(X) = 1 + X^3 + X^5 + X^6$.

6.7.6. Обнаружение ошибок с помощью (*n* – *k*)-разрядного регистра сдвига

Передаваемое кодовое слово может быть искажено помехами, и, следовательно, принятый вектор будет искаженным вариантом переданного кодового слова. Допустим, что передается кодовое слово, имеющее в полиномиальном представлении вид $\mathbf{U}(X)$, а принимается вектор, в полиномиальном представлении имеющий вид $\mathbf{Z}(X)$. Поскольку $\mathbf{U}(X)$ — это полином кодового слова, он должен без остатка делиться на полиномиальный генератор $\mathbf{g}(X)$.

$$\mathbf{U}(X) = \mathbf{m}(X)\mathbf{g}(X) \tag{6.67}$$

$\mathbf{Z}(X)$, искаженную версию $\mathbf{U}(X)$, можно представить следующим образом.

$$\mathbf{Z}(X) = \mathbf{U}(X) + \mathbf{e}(X) \tag{6.68}$$

Здесь $\mathbf{e}(X)$ — полином ошибочной комбинации. Декодер проверяет, является ли $\mathbf{Z}(X)$ полиномом кодового слова, т.е. делится ли он на $\mathbf{g}(X)$ без остатка. Это осуществляется путем вычисления синдрома принятого полинома. Синдром $\mathbf{S}(X)$ равен остатку от деления $\mathbf{Z}(X)$ на $\mathbf{g}(X)$.

$$\mathbf{Z}(X) = \mathbf{q}(X)\mathbf{g}(X) + \mathbf{S}(X) \tag{6.69}$$

Здесь $\mathbf{S}(X)$ — полином степени $n - k - 1$ или меньше. Соответственно, синдром — это $(n - k)$-кортеж. Используя уравнения (6.67) и (6.69), получаем следующее.

$$\mathbf{e}(X) = [\mathbf{m}(X) + \mathbf{q}(X)]\ \mathbf{g}(X) + \mathbf{S}(X) \tag{6.70}$$

Сравнивая уравнения (6.69) и (6.70), видим, что синдром $\mathbf{S}(X)$, полученный как $\mathbf{Z}(X)$ по модулю $\mathbf{g}(X)$, аналогичен остатку деления $\mathbf{e}(X)$ на $\mathbf{g}(X)$. Таким образом, синдром принятого полинома $\mathbf{Z}(X)$ содержит информацию, необходимую для исправления ошибочной комбинации. Расчет синдрома выполняется с помощью схемы деления, почти аналогичной схеме кодирования, используемой в передатчике. Пример вычисления синдрома со сдвигом на $(n - k)$ разрядов регистра приведен на рис. 6.20 с использованием вектора кода, полученного в примере 6.10. В исходном состоянии ключ 1 закрыт, а ключ 2 от-

крыт. Принятый вектор подается во входной регистр, в котором в исходном состоянии все разряды имеют нулевое значение. После того как весь принятый вектор будет занесен в регистр сдвига, содержимое регистра — это и есть синдром. Теперь ключ 1 открывается, а ключ 2 закрывается, так что вектор синдрома теперь можно извлечь из регистра. Описанная последовательность действий имеет следующий вид.

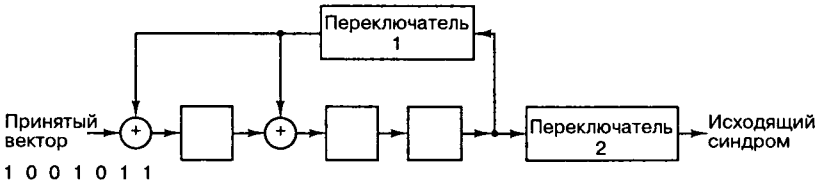


Рис. 6.20. Пример вычисления синдрома с помощью (n – k)-разрядного регистра сдвига

Входная очередь	Номер сдвига	Содержимое регистра
1001011	0	00
100101	1	100
10010	2	110
1001	3	011
100	4	011
10	5	111
1	6	101
	7	000 Синдром

Если вектор синдрома нулевой, считается, что принятый вектор является правильным кодовым словом. Если синдром отличен от нуля, значит, обнаружена ошибка и принятый вектор — это искаженное кодовое слово; данная ошибка исправляется путем прибавления к принятому вектору вектора ошибки (указанной синдромом), т.е. аналогично процедуре, описанной в разделе 6.4.8. Этот метод декодирования хорош для простых кодов. Более сложные коды для практического использования требуют применения алгебраических методик.

6.8. Известные блочные коды

6.8.1. Коды Хэмминга

Коды Хэмминга (Hamming codes) — это простой класс блочных кодов, которые имеют следующую структуру:

$$(n, k) = (2^m - 1, 2^m - 1 - m), \tag{6.71}$$

где $m = 2, 3, \ldots$. Минимальное расстояние этих кодов равно 3, поэтому, согласно уравнениям (6.44) и (6.47), они способны исправлять все однобитовые ошибки или определять все ошибочные комбинации из двух или менее ошибок в блоке. Декодирование с помощью синдромов особенно хорошо подходит к кодам Хэмминга. Фактически синдром можно превратить в двоичный указатель местоположения ошибки [5]. Хотя коды Хэмминга не являются слишком мощными, они принадлежат к очень ограниченному классу блочных кодов, называемых *совершенными*; их особенности описывались в разделе 6.5.4.

Если предположить, что используется жесткое декодирование, вероятность появления битовой ошибки можно записать с помощью уравнения (6.46).

$$P_B \approx \frac{1}{n}\sum_{j=2}^{n} j\binom{n}{j} p^j (1-p)^{n-j} \tag{6.72}$$

Здесь p — вероятность ошибочного приема канального символа (вероятность перехода в двоичном симметричном канале). Вместо уравнения (6.72) мы можем использовать другое эквивалентное уравнение (это уравнение (Г.16), которое выводится в приложении Г).

$$P_B \approx p - p(1-p)^{n-1} \tag{6.73}$$

На рис. 6.21 приведен график зависимости вероятности ошибки в декодированном бите от вероятности ошибки в канальном символе, на котором сравниваются разные блочные коды. Для кодов Хэмминга на графике взяты значения $m = 3$, 4 и 5 или $(n, k) = (7, 4)$, (15, 11), (31, 26). Для описания гауссового канала с использованием когерентной демодуляции сигналов BPSK, вероятность ошибки в канальном символе можно выразить через E_b/N_0, как это было сделано в уравнении (4.79).

$$p = Q\left(\sqrt{\frac{2E_c}{N_0}}\right) \tag{6.74}$$

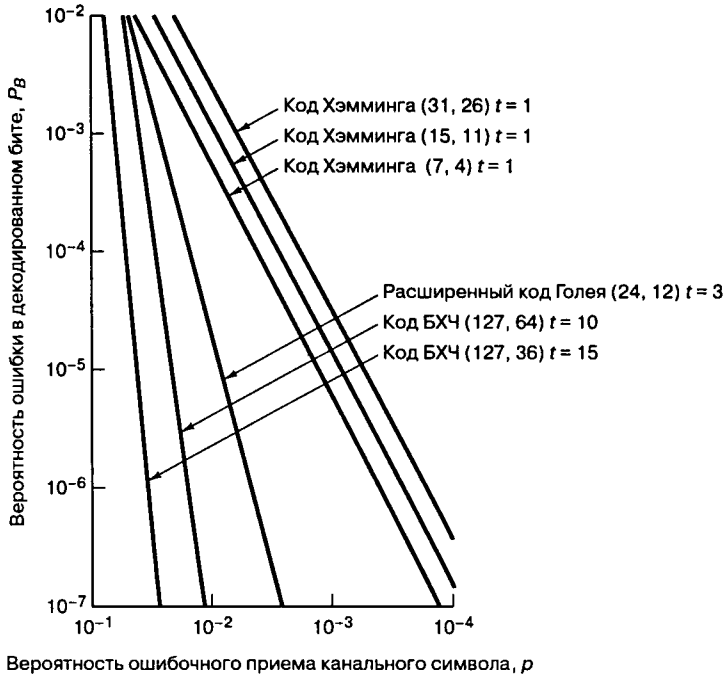


Рис. 6.21. Зависимость вероятности битовой ошибки от вероятности ошибки в канальном символе для нескольких блочных кодов

Здесь E_c/N_0 — отношение энергии кодового символа к спектральной плотности мощности шума, а $Q(X)$ определено в уравнении (3.43). Чтобы связать E_c/N_0 с энергией бита информации на единицу плотности спектрального шума (E_b/N_0), используем следующее выражение.

$$\frac{E_c}{N_0} = \left(\frac{k}{n}\right)\frac{E_b}{N_0} \tag{6.75}$$

Для кодов Хэмминга уравнение (6.75) принимает следующий вид.

$$\frac{E_c}{N_0} = \left(\frac{2^m - 1 - m}{2^m - 1}\right)\frac{E_b}{N_0} \tag{6.76}$$

Объединяя уравнения (6.73), (6.74) и (6.76), P_B при когерентной демодуляции сигналов BPSK в гауссовом канале можно выразить как функцию E_b/N_0. Результаты для различных типов блочных кодов отображены на рис. 6.22. Для кодов Хэмминга взяты следующие значения $(n, k) = (7, 4), (15, 11), (31, 26)$.

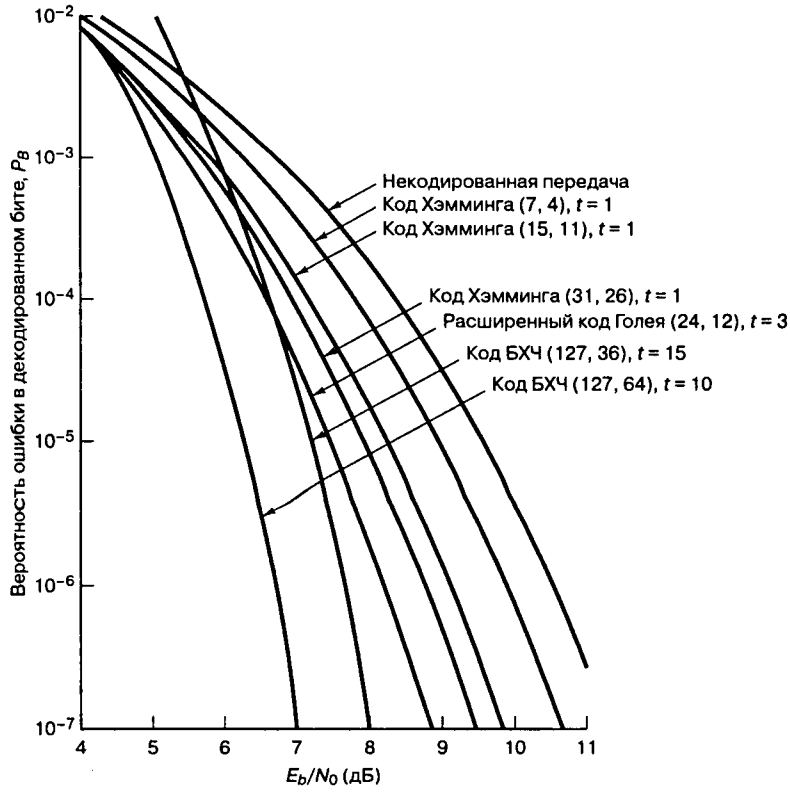


Рис. 6.22. Зависимость P_B от E_b/N_0 при когерентной демодуляции сигналов BPSK в гауссовом канале для нескольких блочных кодов

Пример 6.11. Вероятность ошибки для модулированных и кодированных сигналов

Кодированный сигнал с модуляцией BFSK передается по гауссовому каналу. Сигнал некогерентно обнаруживается и жестко декодируется. Найдите вероятность ошибки в декодированном бите, если кодирование осуществляется блочным кодом Хэмминга (7, 4), а принятое значение E_b/N_0 равно 20.

Решение

Сначала, используя уравнение (6.75), находим E_c/N_0.

$$\frac{E_c}{N_0} = \frac{4}{7}(20) = 11{,}43$$

Затем для кодированного некогерентного сигнала BFSK мы можем связать вероятность ошибки в канальном символе с E_c/N_0, подобно тому, как это было сделано в уравнении (4.96).

$$\begin{aligned} p &= \frac{1}{2}\exp\left(-\frac{E_c}{2N_0}\right) = \\ &= \frac{1}{2}\exp\left(-\frac{11{,}43}{2}\right) = 1{,}6\times 10^{-3} \end{aligned}$$

Подставляя этот результат в уравнение (6.73), получаем следующее значение вероятности ошибки в декодированном бите.

$$P_B \approx p - p(1-p)^{n-1} \approx 1{,}6\times 10^{-5}$$

6.8.2. Расширенный код Голея

Одним из наиболее практичных блочных кодов является двоичный *расширенный код Голея* (extended Golay code) (24, 12), который образован путем прибавления битов четности к совершенному коду (23, 12), известному как *код Голея* (Golay code). Эти дополнительные биты повышают минимальное расстояние $d_{\min}$ с 7 до 8, что дает степень кодирования 1/2, реализовать которую проще (с точки зрения системного тактового генератора), чем степень кодирования кода Голея, равную 12/23. Расширенный код Голея значительно мощнее рассмотренного в предыдущем разделе кода Хэмминга. Цена, которую приходится платить за повышение эффективности, заключается в более сложном декодере и, соответственно, более широкой полосе пропускания.

Для расширенного кода Голея $d_{\min} = 8$, поэтому, исходя из уравнения (6.44), можно сказать, что код гарантирует исправление всех трехбитовых ошибок. Кроме того, декодер можно сконструировать так, чтобы он исправлял *некоторые* комбинации с четырьмя ошибками. Поскольку исправить можно только 16,7% комбинаций с четырьмя ошибками, декодер, для упрощения, обычно реализуется для исправления только трехбитовых ошибочных комбинаций [5]. Если предположить жесткое декодирование, то вероятность битовой ошибки для расширенного кода Голея можно представить как функцию вероятности p ошибки в канальном символе (см. уравнение (6.46)).

$$P_B \approx \frac{1}{24}\sum_{j=4}^{24} j\binom{24}{j} p^j (1-p)^{24-j} \tag{6.77}$$

График зависимости (6.77) показан на рис. 6.21; вероятность появления ошибки для расширенного кода Голея значительно меньше, чем у кодов Хэмминга. Исходя из уравнений (6.77), (6.74) и (6.75), можно связать P_B с E_b/N_0 для сигнала BPSK в гауссовом канале с кодированием расширенным кодом Голея. Результаты показаны на рис. 6.22.

6.8.3. Коды БХЧ

Коды Боуза-Чоудхури-Хоквенгема (Bose-Chadhuri-Hocquenghem — BCH, БХЧ) являются результатом обобщения кодов Хэмминга, которое позволяет исправлять множественные ошибки. Они составляют *мощный класс циклических кодов*, который обеспечивает достаточную свободу выбора длины блока, степени кодирования, размеров алфавита и возможностей коррекции ошибок. В табл. 6.4 приводятся наиболее часто употребляемые при создании кодов БХЧ генераторы $\mathbf{g}(x)$ [8] с разными значениями n, k и t для блоков длиной до 255. Коэффициенты $\mathbf{g}(x)$ представлены восьмеричными числами, оформленными так, что при преобразовании их в двоичные символы крайние правые разряды отвечают коэффициенту нулевой степени в $\mathbf{g}(x)$. С помощью табл. 6.4 можно легко проверить свойство циклического кода — полиномиальный генератор имеет порядок $n - k$. Коды БХЧ очень важны, поскольку при блоках, длина которых равна порядка несколько сотен, коды БХЧ превосходят своими качествами все другие блочные коды с той же длиной блока и степенью кодирования. В наиболее часто применяемых кодах БХЧ используется двоичный алфавит и блок кодового слова длиной $n = 2^m - 1$, где $m = 3, 4, \ldots$.

Из названия табл. 6.4 ясно, что показаны генераторы только для *примитивных* кодов БХЧ. Термин "примитивные" (primitive) — это теоретико-числовое понятие, требующее алгебраического рассмотрения [7, 10-11], которое представлено в разделе 8.1.4. На рис. 6.21 и 6.22 изображены графики вероятности ошибки для двух кодов БХЧ: (127, 64) и (127, 36). На рис. 6.21 показана зависимость P_B от вероятности ошибки в канальном символе при жестком декодировании. На рис. 6.22 показана зависимость P_B от E_b/N_0 для когерентно демодулированного сигнала BPSK в гауссовом канале. Кривые на рис. 6.22 выглядят совсем не так, как можно было бы ожидать. Все они имеют одну и ту же длину блока, но большая избыточность кода (127, 36) не дает той эффективности кодирования, какая имеется у менее избыточного кода (127, 64). Известно, что относительно широкий максимум эффективности кодирования, в зависимости от степени кодирования при фиксированном n, для кодов БХЧ находится примерно между степенью 1/3 и 3/4 [12]. Стоит также отметить, что передача по гауссову каналу сильно ухудшается при переходе от очень высоких до очень низких скоростей [11].

Таблица 6.4. Генераторы примитивных кодов БХЧ

n	k	t	g(x)
7	4	1	13
15	11	1	23
	7	2	721
	5	3	2467
31	26	1	45
	21	2	3551
	16	3	107657
	11	5	5423325
	6	7	313365047
63	57	1	103
	51	2	12471
	45	3	1701317
	39	4	166623567
	36	5	1033500423
	30	6	157464165547
	24	7	17323260404441
	18	10	1363026512351725
	16	11	6331141367235453
	10	13	472622305527250155
	7	15	5231045543503271737
127	120	1	211
	113	2	41567
	106	3	11554743
	99	4	3447023271
	92	5	624730022327
255	171	11	15416214212342356077061630637
	163	12	750041551007560255157472451460l
	155	13	375751300540766501572250646467763 3
	147	14	164213017353716552530416530544101171 1
	139	15	461401732060175561570722730247453567445
	131	18	215713331471510151261250277442142024 1 65471
	123	19	120614505224206600371721032651614122 62 72506267
	115	21	605266655721002472636364046002763525 5 6313472737
	107	22	222057723220662563124173002353474201 7 6574750154441
	99	23	106566672534731742227414162015743322 5 2411076432303431
	91	25	675026503032744417272363172473251107 5 550762720724344561
	87	26	110136763414743236435231634307172046 2 06722545273311721317
	79	27	667000356376575000202703442073661746 2 101532671176654134235 5
	71	29	240247105206443215155541721123311632 0 5444250362557643221706035
	63	30	107544750551635443253152173577070036 6 611172645526761365670254330 1

Окончание табл.6.4

n	k	t	g(x)	n	k	t	g(x)
	85	6	130704476322273	55	31	7315425203501100133015275306032054325
	78	7	2623000216613011 5			4143267550105570444260354736 17
	71	9	6255010703253127753	47	42	25335420170626465630330413774062330 75
	64	10	1206534025570773100045			12333414544604500506602455254317 3
	57	11	335265252505705053517721	45	43	15202056055234161131101346376423701 56
	50	13	54446512523314012421501421			36700244707623730332021570250515 41
	43	14	17721772213651227521220574343	37	45	51363302550670074141774472454375304 20
	36	15	31460746665220750447645747217 35			73570617432343234764435473740304400 3
	29	21	40311446136767060366753014117 6155	29	47	30257155366730714655270640123613771 15
	22	23	12337607040472252243544562663764 7043			34224232420117411406025475741040356 5037
	15	27	22057042445604554770523013762217604353			
	8	31	70472640527510306514762242715677331302 17	21	55	12562152570603326560017731536076121 03
255	247	1	435			22734140565307454252115312161446651
	239	2	267543			3473725
	231	3	156720665	13	59	46417320050525645444265737142500660 04
	223	4	75625541375			33067744547656140317467721357026134
	215	5	23157564726421			460500547
	207	6	16176560567636227	9	63	15726025217472463201031043255355134 61
	199	7	763303127042072234 1			41623672120440745451127661155477055
	191	8	2663470176115333714567			61677516057
	187	9	52755313540001322236351			
	179	10	22624710717340432416300455			

Источник. Перепечатано с разрешения авторов из *Table of Generators for BCH Codes*". IEEE Trans. Inf. Theory, vol. IT10, n. 4, October, 1964, p. 391. © 1964, IEEE.

На рис. 6.23 показаны расчетные характеристики кодов БХЧ для когерентно демодулированного сигнала BPSK с *жестким* и *мягким декодированием*. Мягкое декодирование для блочных кодов не применяется из-за своей сложности, хотя оно и дает увеличение эффективности кодирования порядка 2 дБ по сравнению с жестким декодированием. При данной степени кодирования вероятность ошибки при декодировании уменьшается с ростом длины блока n [4]. Таким образом, при данной степени кодирования интересно рассмотреть необходимую длину блока для сравнения характеристик жесткого и мягкого декодирования. На рис. 6.23 все коды показаны со степенью кодирования, равной приблизительно 1/2. Из рисунка [13] видно, что при фиксированной степени кодирования и жестком декодировании кода БХЧ длиной $8n$ или более наблюдаются лучшие характеристики, чем при мягком декодировании кода БХЧ длиной n. Существует специальный подкласс кодов БХЧ (которые были разработаны раньше кодов БХЧ), который является *недвоичным* набором; это коды *Рида-Соломона* (Reed-Solomon code). Подробнее об этих кодах будет рассказано в разделе 8.1.

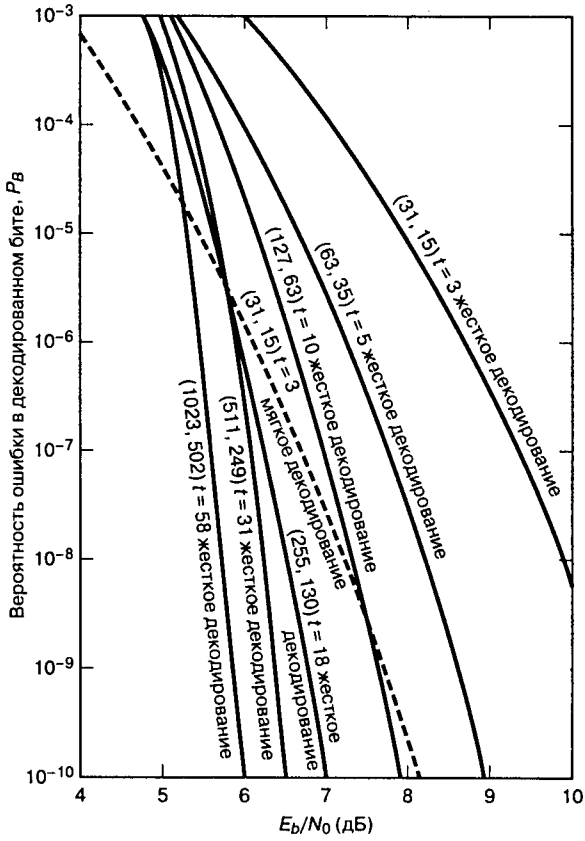


Рис. 6.23. Зависимость P_B от E_b/N_0 для когерентно демодулируемого сигнала BPSK в гауссовом канале с использованием кодов БХЧ. (Перепечатано с разрешения автора из L. J. Weng. "Soft and Hard Decoding Performance Comparison for BCH Codes", Proc. Int. Conf. Commun., 1979, Fig. 3, p. 25.5.5. © 1979, IEEE.)

6.9. Резюме

В этой главе проанализирована главная задача канального кодирования — улучшение рабочих характеристик (вероятности ошибки, E_b/N_0 или пропускной способности) за счет полосы пропускания. Изучение канального кодирования было разбито на две части: кодирование формы сигнала и структурированные последовательности. Кодирование формы сигнала представляет собой преобразование сигналов в усовершенствованные сигналы, которые дают улучшенные пространственные характеристики (по сравнению с исходными сигналами). Структурированные последовательности подразумевают добавление к данным избыточных разрядов, что позволяет обнаруживать и/или исправлять определенные ошибочные комбинации.

Здесь также детально рассмотрены блочные коды. Между кодированием и модуляцией можно провести геометрическую аналогию. Обе процедуры пытаются максимально наполнить пространство сигналов и максимально увеличить расстояние между сигналами в наборе. Из блочных кодов были рассмотрены циклические коды, которые сравнительно легко реализуются с помощью современных технологий интегральных схем. Также было рассмотрено полиномиальное представление кодов и соответствия между полиномиальной структурой, необходимыми алгебраическими операциями и конкретной реализацией таких схем. В заключение были представлены некоторые сведения о самых известных блочных кодах. Другие вопросы, связанные с кодированием, будут рассматриваться в последующих главах. В главе 7 мы обсудим обширный класс сверточных кодов; в главе 8 будут рассмотрены коды Рида-Соломона, каскадные коды и турбокоды; а в главе 9 будет изучено решетчатое кодирование.

Литература

1. Viterbi A. J. *On Coded Phase-Coherent Communications.* IRE Trans. Space Electron. Telem., vol. SET7, March, 1961, pp. 3–14.
2. Lindsey W. C. and Simon M. K. *Telecommunication Systems Engineering.* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973.
3. Proakis J. G. *Digital Communications.* McGraw-Hill Book Company, New York, 1983.
4. Lin S. and Costello D. J. Jr. *Error Control Coding: Fundamentals and Applications.* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1983.
5. Odenwalder J. P. *Error Control Coding Handbook.* Linkabit Corporation, San Diego, Calif., July, 15, 1976.
6. Blahut R. E. *Theory and Practice of Error Control Codes.* Addison-Wesley Publishing Company, Inc., Reading, Mass, 1983.
7. Peterson W. W. and Weldon E. J. *Error Correcting Codes,* 2nd ed. The MIT Press, Cambridge, Mass, 1972.
8. Blahut R. E. *Algebraic Fields, Signal Processing and Error Control.* Proc. IEEE, vol. 73, May, 1985, pp. 874–893.
9. Stenbit J. P. *Table of Generators for Bose-Chadhuri Codes.* IEEE Trans. Inf. Theory, vol. IT10, n. 4, October, 1964, pp. 390–391.
10. Berlekamp E. R. *Algebraic Coding Theory.* McGraw-Hill Book Company, New York, 1968.
11. Clark G. C. Jr. and Cain J. B. *Error-Correction Coding for Digital Communications.* Plenum Press, New York, 1981.
12. Wozencraft J. M. and Jacobs I. M. *Principles of Communication Engineering.* John Wiley & Sons, Inc., New York, 1965.
13. Weng L. J. *Soft and Hard Decoding Performance Comparisons for BCH Codes.* Proc. Int. Conf. Commun., 1979, pp. 25.5.1–25.5.5.

Задачи

6.1. Сконструируйте код (n, k) с проверкой на четность, который будет определять все комбинации, содержащие 1, 3, 5 и 7 ошибочных бит. Найдите значения n и k и определите вероятность невыявленной ошибки в блоке, если вероятность ошибки в канальном символа равна 10^{-2}.

6.2. Определите вероятность ошибки в сообщении для 12-битовой последовательности данных, кодированной линейным блочным кодом (24, 12). Допустим, что код может исправлять одно- и двухбитовые ошибочные комбинации и что ошибочные комбинации с более чем двумя ошибками не подлежат исправлению. Также предположим, что вероятность ошибки в канальном символе равна 10^{-3}.

6.3. Рассмотрим линейный блочный код (127, 92), который может исправлять трехбитовые ошибки.

а) Чему равна вероятность ошибки в сообщении для некодированного блока из 92 бит, если вероятность ошибки в канальном символе равна 10^{-3}?

б) Чему равна вероятность ошибки для сообщения, кодированного блочным кодом (127, 92), если вероятность ошибки в канальном символе равна 10^{-3}?

6.4. Рассчитайте уменьшение вероятности ошибки в сообщении, кодированном линейным блочным кодом (24, 12) с коррекцией двухбитовых ошибок, по сравнению с некодированной передачей. Предположим, что используется когерентная модуляция BPSK и принятое $E_b/N_0 = 10$ дБ.

6.5. Рассмотрим линейный блочный код (24, 12) с возможностью исправления двухбитовых ошибок. Пусть используется модуляция BFSK, а принятое $E_b/N_0 = 14$ дБ.

а) Дает ли код какое-либо уменьшение вероятности ошибки в сообщении? Если да, то насколько? Если нет, то почему?

б) Повторите п. а при $E_b/N_0 = 10$ дБ.

6.6. Телефонная компания применяет кодер типа "лучший из пяти" для некоторых цифровых каналов данных. В такой схеме все биты данных повторяются пять раз, и в приемнике выполняется мажоритарное декодирование сообщения. Если вероятность ошибки в некодированном бите составляет 10^{-3} и используется кодирование "лучший из пяти", чему равна вероятность ошибки в декодированном бите?

6.7. Минимальное расстояние для конкретного линейного блочного кода равно 11. Найдите максимальные возможности кода при исправлении ошибок, максимальные возможности при обнаружении ошибок и максимальные возможности этого кода при коррекции стираний для данной длины блока.

6.8. Дается матрица генератора кода (7, 4) следующего вида.

$$\mathbf{G} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

а) Найдите все кодовые слова кода.

б) Найдите проверочную матрицу **H** этого кода.

в) Рассчитайте синдром для принятого вектора 1101101. Правильно ли принят этот вектор?

г) Каковы возможности кода при исправлении ошибок?

д) Каковы возможности кода при обнаружении ошибок?

6.9. Рассмотрите линейный блочный код, контрольные уравнения которого имеют следующий вид.

$$p_1 = m_1 + m_2 + m_4$$
$$p_2 = m_1 + m_3 + m_4$$
$$p_3 = m_1 + m_2 + m_3$$
$$p_4 = m_2 + m_3 + m_4$$

Здесь m_i — разряды сообщения, а p_i — контрольные разряды.

а) Найдите для этого кода матрицу генератора и проверочную матрицу.

б) Сколько ошибок может исправить этот код?

в) Является ли вектор 10101010 кодовым словом?

г) Является ли вектор 01011100 кодовым словом?

6.10. Рассмотрите линейный блочный код, для которого кодовое слово определяется следующим вектором.

$$\mathbf{U} = m_1 + m_2 + m_4 + m_5,\ m_1 + m_3 + m_4 + m_5,\ m_1 + m_2 + m_3 + m_5,$$
$$m_1 + m_2 + m_3 + m_4,\ m_1,\ m_2,\ m_3,\ m_4,\ m_5$$

а) Найдите матрицу генератора.

б) Найдите проверочную матрицу.

в) Найдите n, k и $d_{\min}$.

6.11. Постройте линейный блочный код $(n, k) = (5, 2)$.

а) Выберите кодовые слова в систематической форме так, чтобы получить максимальное значение d_{min}.

б) Найдите для этого набора кодовых слов матрицу генератора.

в) Рассчитайте проверочную матрицу.

г) Внесите все n-кортежи в нормальную матрицу.

д) Каковы возможности этого кода в обнаружении и исправлении ошибок?

е) Составьте таблицу синдромов для исправимых ошибочных комбинаций.

6.12. Рассмотрим код с повторениями (5, 1), содержащий два кодовых слова 00000 и 11111, соответствующих передаче 0 и 1. Составьте нормальную матрицу для этого кода. Будет ли этот код совершенным?

6.13. Постройте код (3, 1), способный исправлять все однобитовые ошибочные комбинации. Подберите набор кодовых слов и составьте нормальную матрицу.

6.14. Будет ли код (7, 3) совершенным? Будет ли совершенным код (7, 4)? А код (15, 11)? Ответ аргументируйте.

6.15. Линейный блочный код (15, 11) можно определить следующей матрицей четности.

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

а) Найдите для этого кода проверочную матрицу.

б) Укажите образующие элементы классов смежности в нормальной матрице. Является ли этот код совершенным? Обоснуйте свой ответ.

в) Пусть принят вектор $\mathbf{V} = 011111001011011$. Рассчитайте его синдром. Предположите, что сделана однобитовая ошибка, и найдите правильное кодовое слово.

г) Сколько стираний может исправить это код? Ответ аргументируйте.

6.16. Может ли ненулевая ошибочная комбинация дать синдром $\mathbf{S} = \mathbf{0}$? Если да, то сколько таких комбинаций существует для кода (n, k)? Для объяснения ответа воспользуйтесь рис. 6.11.

6.17. Определите, какие (если таковые есть) из следующих полиномов могут генерировать циклический код с кодовым словом длиной $n \leq 7$. Найдите значение (n, k) для каждого из таких кодов.

а) $1 + X^3 + X^4$

б) $1 + X^2 + X^4$

в) $1 + X + X^3 + X^4$

г) $1 + X + X^2 + X^4$

д) $1 + X^3 + X^5$

6.18. Используя полиномиальное деление и генератор $g(X) = 1 + X + X^2 + X^4$, закодируйте в систематической форме сообщение 101.

6.19. Сконструируйте кодер на регистрах сдвига с обратной связью для циклического кода (8, 5) с генератором $g(X) = 1 + X + X^2 + X^3$. С помощью кодера найдите кодовое слово в систематической форме для сообщения 10101.

6.20. На рис. 36.1 сигнал передается в модуляции DPSK, скорость передачи кодовых символов составляет 10 000 кодовых символов в секунду, декодер является декодером (7, 4) с коррекцией однобитовых ошибок. Достаточно ли значения $P_r/N_0 = 48$ дБВт на входе для получения на выходе вероятности ошибки в сообщении 10^{-3}? Обоснуйте свой ответ. Считайте, что блок сообщения содержит 4 бита данных и что можно исправить любую однобитовую ошибочную комбинацию в блоке длиной 7 бит.

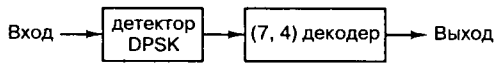


Рис. 36.1

6.21. Циклический код (15, 5) имеет полиномиальный генератор следующего вида.

$$\mathbf{g}(X) = 1 + X + X^2 + X^5 + X^8 + X^{10}$$

а) Нарисуйте схему кодера для этого кода.

б) Найдите полином кода (в систематической форме) для сообщения $\mathbf{m}(X) = 1 + X^2 + X^4$.

в) Будет ли $\mathbf{V}(X) = 1 + X^4 + X^6 + X^8 + X^{14}$ полиномом кода в этой системе? Объясните свой ответ.

6.22. Рассмотрим циклический код (15, 11), который генерируется генератором $g(X) = 1 + X + X^4$.

а) Разработайте для этого кода кодер и декодер на основе регистра с обратной связью.

б) Проиллюстрируйте процедуру кодирования на примере вектора сообщения 11001101011, перечислив все состояния регистра (крайний правый бит является самым первым).

в) Повторите п. б для процедуры декодирования.

6.23. При фиксированной вероятности ошибки в канальном символе вероятность битовой ошибки для кода Хэмминга (15, 11) больше, чем для кода Хэмминга (7, 4). Объясните, почему? В чем тогда заключается преимущество кода (15, 11)? Каков основной компромисс здесь задействован?

6.24. Код БХЧ (63, 36) может исправить пять ошибок. Девять блоков кода (7, 4) могут исправить девять ошибок. Оба кода имеют одинаковую степень кодирования.

а) Код (7, 4) может исправить больше ошибок. Является ли он более мощным? Объясните свой ответ.

б) Сравните оба кода, когда наблюдается пять случайных ошибок в 63 бит.

6.25. Исходная информация разбита на 36-битовые сообщения и передается по каналу AWGN с помощью сигналов в модуляции BFSK.

а) Рассчитайте E_b/N_0, необходимое для получения вероятности ошибки в сообщении 10^{-3}, если применяется кодирование без защиты от ошибок.

б) Пусть при передаче этих сообщений используется линейный блочный код (127, 36). Рассчитайте эффективность кодирования для этого кода при вероятности ошибки в сообщении 10^{-3}. (*Подсказка:* эффективность кодирования определяется как разность между требуемым E_b/N_0 без кодирования и E_b/N_0 с кодированием.)

6.26. а) Пусть последовательность данных кодируется кодом БХЧ (127, 64), а затем модулируется когерентной 16-арной схемой PSK. Если принятое E_b/N_0 равно 10 дБ, чему равны вероятность ошибки в принятом символе, вероятность ошибки в кодовом бите (предполагается, что для присвоения символам битового значения используется код Грея) и вероятность ошибки в информационном бите.

б) Для той же вероятности ошибки в информационном бите, которая была найдена в п. а, определите требуемое значение E_b/N_0, если модуляция в п. а заменена на когерентную ортогональную 16-арную FSK. Объясните отличия.

6.27. В сообщении содержится текст на английском языке (предполагается, что каждое слово в сообщении содержит шесть букв). Каждая буква кодируется 7-битовым символом ASCII. Таким образом, каждое слово текста представляется 42-битовой последовательностью. Сообщение передается по каналу с вероятностью ошибки в символе 10^{-3}.

а) Какова вероятность того, что слово будет передано с ошибкой?

б) Если применяется код с тройным повторением каждой буквы, а приемник осуществляет мажоритарное декодирование, чему равна вероятность появления ошибки в декодированном слове?

в) Если для кодирования каждого 42-битового слова применяется код БХЧ (126, 42) с возможностью исправления ошибок с $t = 14$, то какова будет вероятность появления ошибки в декодированном слове?

г) В реальной системе не совсем явно можно сравнить характеристики кодированной и некодированной вероятностей ошибки в сообщении, используя фиксированную вероятность ошибочной передачи канального символа, поскольку это предполагает фиксированный уровень принятого E_c/N_0 для любого способа кодирования (в том числе и без кодирования). Поэтому повторите пп. а–в при условии, что вероятность ошибочной передачи канального символа определяется уровнем принятого E_b/N_0, равного 12 дБ, где E_b/N_0 — это отношение энергии информационного бита к спектральной плотности шума. Предположим, что скорость передачи информации одинакова для всех типов кодирования и для системы без кодирования. Также допустим, что используется некогерентная ортогональная модуляция FSK, а в канале присутствует шум AWGN.

д) Обсудите относительные возможности надежной работы описанных выше схем кодирования при двух условиях — фиксированная вероятность ошибки в канальном символе и фиксированное отношение E_b/N_0. В каком случае код с повторением может дать повышение достоверности передачи? В каком случае достоверность снизится?

6.28. Последовательность блоков данных из пяти бит с помощью матрицы Адамара преобразуется в ортогонально кодированную последовательность. Когерентное обнаружение осуществляется в течение периода передачи кодового слова, как показано на рис. 6.5. Считая $P_B = 10^{-5}$, рассчитайте эффективность кодирования для побитовой передачи данных с использованием модуляции BPSK.

6.29. Для кода (8, 2), описанного в разделе 6.6.3, проверьте правильность величин матрицы генератора, проверочной матрицы и векторов синдромов для каждого класса смежности 1–10.

6.30. Составьте схему на основе логических элементов исключающего ИЛИ и И, аналогичную схеме на рис. 6.12, исправляющую все однобитовые ошибочные комбинации кода (8, 2), определяемые образующими элементами классов смежности 2–9, показанными на рис. 6.15.

6.31. Подробно объясните возможность составления схемы на основе логических элементов исключающего ИЛИ и И (аналогичной схеме на рис. 6.12), исправляющей все одно- и двухбитовые ошибочные комбинации кода (8, 2) и обнаруживающей трехбитовые ошибочные комбинации (образующие элементы классов смежности или строки 38–64).

6.32. Проверьте, что все коды БХЧ длиной $n = 31$, показанные в табл. 6.4, удовлетворяют условиям пределов Хэмминга и Плоткина.

6.33. При кодировании нулевого блока сообщения в результате получается нулевое кодовое слово. Обычно такую последовательность нулей передавать нежелательно. В одном методе циклического кодирования при такой передаче разряды регистра сдвига предварительно (до кодирования) заполняются единицами, а не нулями, как обычно. Получаемая в результате "псевдочетность" гарантированно содержит некоторое количество единиц. В декодере перед началом декодирования производится обратная операция. Постройте общую схему для инверсной обработки псевдочетных битов в каком-либо циклическом декодере. Воспользуйтесь кодером БХЧ (7, 4), заполненным единицами для кодирования сообщения 1011 (самым первым является крайний правый бит). Затем покажите, что составленная вами инверсная схема позволяет получить правильное декодированное сообщение.

6.34. а) В условиях задачи 6.21 кодируйте в систематической форме последовательность сообщения 11011, воспользовавшись полиномиальным генератором для циклического кода (15, 5). Найдите результирующий полином кодового слова. Какой особенностью характеризуется степень полиномиального генератора?

б) Пусть принятое кодовое слово искажено ошибочной комбинацией $\mathbf{e}(X) = X^8 + X^{10} + X^{13}$. Найдите полином искаженного кодового слова.

в) Исходя из полинома принятого вектора и полиномиального генератора найдите полином синдрома.

г) Исходя из полинома ошибочной комбинации и полиномиального генератора найдите полином синдрома и убедитесь, что это тот же синдром, что и найденный в п. в.

д) Объясните, почему в пп. в и г должен получиться одинаковый результат.

е) Используя свойство нормальной матрицы линейного блочного кода (15, 5), найдите максимальное количество исправлений ошибок, которое может выполнить код с данными параметрами. Является ли код (15, 5) совершенным?

ж) Если мы хотим применить циклический код (15, 5) для одновременного исправления двух стираний и сохранить исправление ошибок, насколько придется пожертвовать возможностью исправления ошибок?

Вопросы

6.1. Опишите четыре типа компромиссов, которые могут быть достигнуты при использовании кода коррекции ошибок (см. раздел 6.3.4).

6.2. В системах связи *реального времени* за получаемую с помощью избыточности эффективность кодирования приходится платить *полосой пропускания*. Чем приходится жертвовать за полученную эффективность кодирования в системах связи *модельного времени* (см. раздел 6.3.4.2)?

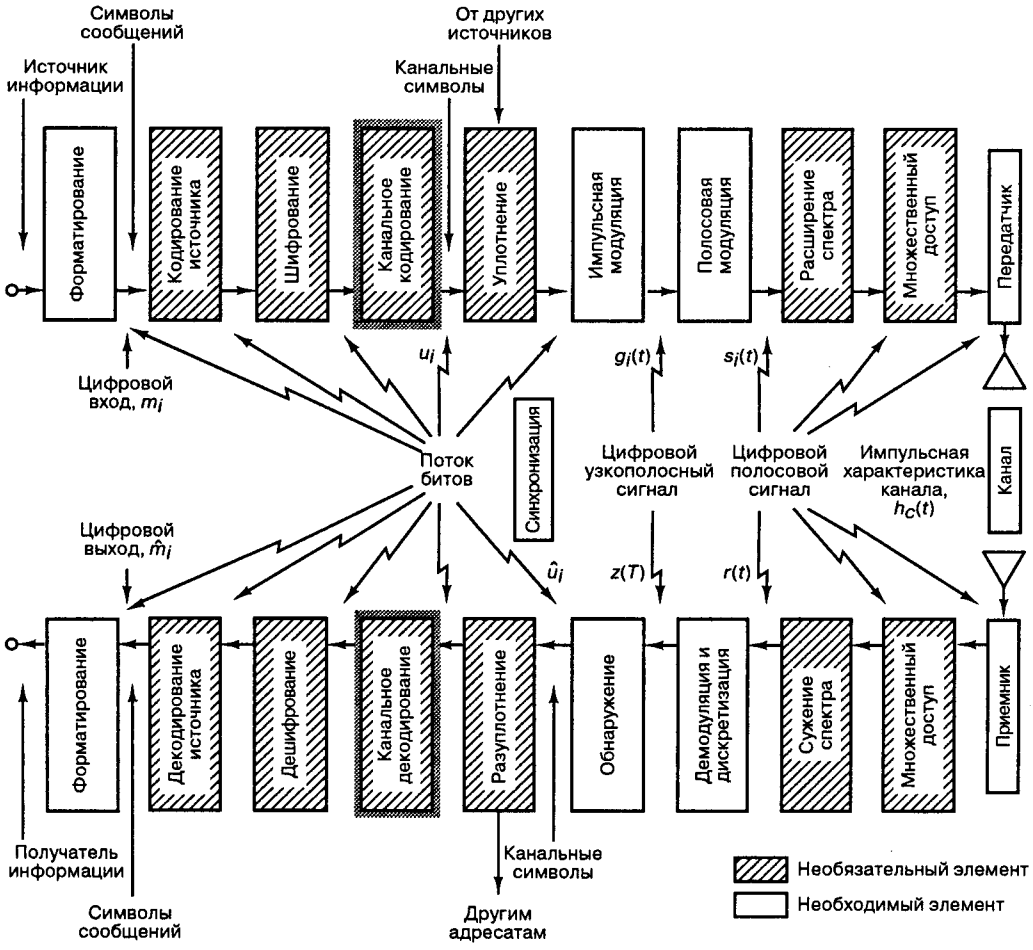
6.3. В системах связи реального времени увеличение избыточности означает повышение скорости передачи сигналов, меньшую энергию на канальный символ и больше ошибок на выходе демодулятора. Объясните, как на фоне такого ухудшения характеристик достигается эффективность кодирования (см. пример 6.2).

6.4. Почему эффективность традиционных кодов коррекции ошибок снижается при низких значениях E_b/N_0 (см. раздел 6.3.4.6)?

6.5. Опишите процесс проверки с использованием синдромов, обнаружения ошибки и ее исправления в контексте примера из области медицины (см. раздел 6.4.8.4).

6.6. Определите место *нормальной матрицы* в понимании блочного кода и оценке его возможностей (см. раздел 6.6.5).

Канальное кодирование: часть 2



В этой главе рассматривается сверточное кодирование. В главе 6 обсуждались основы линейных блочных кодов, которые описываются двумя целыми числами, n и k, и полиномиальным или матричным генератором. Целое число k указывает на число бит данных, которые образуют вход блочного кодера. Целое число n — это суммарное количество разрядов в соответствующем кодовом слове на выходе кодера. Особенностью линейного блочного кода является то, что каждый из n-кортежей кодовых слов однозначно определяется k-кортежем входного сообщения. Отношение k/n, называемое *степенью кодирования* кода (code rate), является мерой добавленной избыточности. Сверточный код описывается тремя целыми числами n, k и K, где отношение k/n имеет такое же значение степени кодирования (информация, приходящаяся на закодированный бит), как и для блочного кода; однако n *не* определяет длину блока или кодового слова, как это было в блочных кодах. Целое число K является параметром, называемым *длиной кодового ограничения* (constraint length); оно указывает число разрядов k-кортежа в кодирующем регистре сдвига. Важная особенность сверточных кодов, в отличие от блочных, состоит в том, что кодер имеет память — n-кортежи, получаемые при сверточном кодировании, являются функцией не только одного входного k-кортежа, но и предыдущих $K-1$ входных k-кортежей. На практике n и k — это небольшие целые числа, а K изменяется с целью контроля мощности и сложности кода.

7.1. Сверточное кодирование

На рис. 1.2 представлена типичная блочная диаграмма системы цифровой связи. Разновидность такой функциональной диаграммы, относящаяся, в первую очередь, к сверточному кодированию/декодированию и модуляции/демодуляции, показана на рис. 7.1. Исходное сообщение на входе обозначается последовательностью $\mathbf{m} = m_1, m_2, ..., m_i, ...$, где m_i — двоичный знак (бит), а i — индекс времени. Если быть точным, то элементы $\mathbf{m}$ следовало бы дополнять индексом члена класса (например, для бинарного кода, 1 или 0) и индексом времени. Однако в этой главе для простоты будет использоваться только индекс, обозначающий время (или расположение элемента внутри последовательности). Мы будем предполагать, что все m_i равновероятно равны единице или нулю и независимы между собой. Будучи независимой, последовательность битов нуждается в некоторой избыточности, т.е. знание о бите m_i не дает никакой информации о бите m_j (при $i \neq j$). Кодер преобразует каждую последовательность $\mathbf{m}$ в уникальную последовательность кодовых слов $\mathbf{U} = G(\mathbf{m})$. Даже несмотря на то что последовательность $\mathbf{m}$ однозначно определяет последовательность $\mathbf{U}$, ключевой особенностью сверточных кодов является то, что данный k-кортеж внутри $\mathbf{m}$ *не* однозначно определяет связанные с ним n-кортежи внутри $\mathbf{U}$, поскольку кодирование каждого из k-кортежей является функцией *не только* k-кортежей, но и предыдущих $K-1$ k-кортежей. Последовательность $\mathbf{U}$ можно разделить на последовательность ответвленных слов: $\mathbf{U} = U_1, U_2, ..., U_i, ...$. Каждое ответвленное слово U_i состоит из двоичных *кодовых символов*, часто называемых *канальными символами*, *канальными битами*, или *битами кода*; в отличие от битов входного сообщения, кодовые символы не являются независимыми.

В типичных системах связи последовательность кодовых слов $\mathbf{U}$ модулируется сигналом $s(t)$. В ходе передачи сигнал искажается шумом, в результате чего, как показано на рис. 7.1, получается сигнал $\hat{s}(t)$ и демодулированная последовательность $\mathbf{Z} = Z_1,$

Z_2, ..., Z_j, ... Задача декодера состоит в получении оценки $\hat{m} = \hat{m}_1, \hat{m}_2, \ldots, \hat{m}_i, \ldots$ исходной последовательности сообщения с помощью полученной последовательности **Z** и априорных знаний о процедуре кодирования.

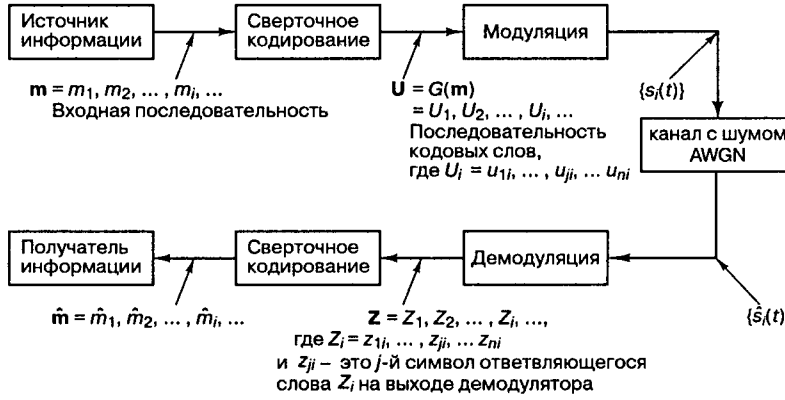


Рис. 7.1. Кодирование/декодирование и модуляция/демодуляция в канале связи

Обычный сверточный кодер, показанный на рис. 7.2, реализуется с kK-разрядным регистром сдвига и n сумматорами по модулю 2, где K — длина кодового ограничения. Длина кодового ограничения — это количество k-разрядных сдвигов, после которых один информационный бит может повлиять на выходной сигнал кодера. В каждый момент времени на место первых k разрядов регистра перемещаются k новых бит; все биты в регистре смещаются на k разрядов вправо, и выходные данные n сумматоров последовательно дискретизируются, давая, в результате, биты кода. Затем эти символы кода используются модулятором для формирования сигналов, которые будут переданы по каналу. Поскольку для каждой входящей группы из k бит сообщения имеется n бит кода, степень кодирования равна k/n бит сообщения на бит кода, где $k < n$.

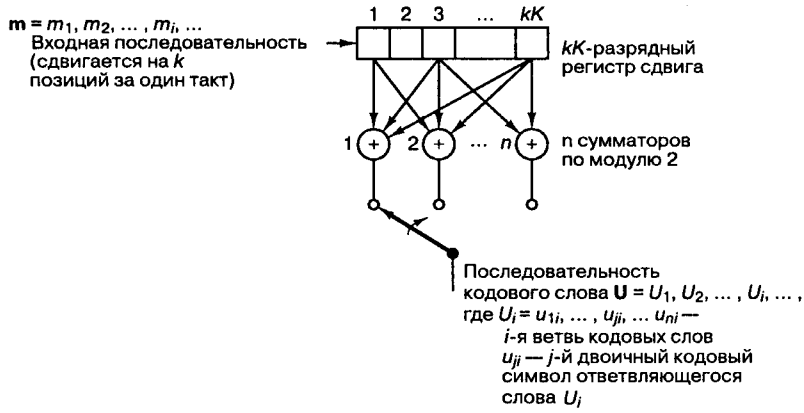


Рис. 7.2. Сверточный кодер с длиной кодового ограничения K и степенью кодирования k/n

Мы будем рассматривать только наиболее часто используемые двоичные сверточные кодеры, для которых $k = 1$, т.е. те кодирующие устройства, в которых биты сообщения сдвигаются по одному биту за раз, хотя обобщение на алфавиты более высоких порядков не вызывает никаких затруднений [1, 2]. Для кодера с $k = 1$, за i-й момент времени бит сообщения m_i будет перемещен на место первого разряда регистра сдвига; все предыдущие биты в регистре будут смещены на один разряд вправо, а выходной сигнал n сумматоров будет последовательно оцифрован и передан. Поскольку для каждого бита сообщения имеется n бит кода, степень кодирования равна $1/n$. Имеющиеся в момент времени t_i n кодовых символов составляют i-е ответвленное слово, $U_i = u_{1i}, u_{2i}, ..., u_{ni}$, где u_{ji} ($j = 1, 2, ..., n$) — это j-й кодовый символ, принадлежащий i-му ответвленному слову. Отметим, что для кодера со степенью кодирования $1/n$, kK-разрядный регистр сдвига для простоты можно называть K-разрядным регистром, а длину кодового ограничения K, которая выражается в единицах разрядов k-кортежей, можно именовать длиной кодового ограничения в битах.

7.2. Представление сверточного кодера

Чтобы иметь возможность описывать сверточный код, необходимо определить кодирующую функцию $G(\mathbf{m})$ так, чтобы по данной входящей последовательности $\mathbf{m}$ можно было быстро вычислить выходную последовательность $\mathbf{U}$. Для реализации сверточного кодирования используется несколько методов; наиболее распространенными из них являются *графическая связь*, *векторы*, *полиномы связи*, *диаграмма состояния*, *древовидная* и *решетчатая диаграммы*. Все они рассматриваются ниже.

7.2.1. Представление связи

При обсуждении сверточных кодеров в качестве модели будем использовать сверточный кодер, показанный на рис. 7.3. На этом рисунке изображен сверточный кодер (2, 1) с длиной кодового ограничения $K = 3$. В нем имеется $n = 2$ сумматора по модулю 2; следовательно, степень кодирования кода k/n равна 1/2. При каждом поступлении бит помещается в крайний левый разряд, а биты регистра смещаются на одну позицию вправо. Затем коммутатор на выходе дискретизирует выходы всех сумматоров по модулю 2 (т.е. сначала верхний сумматор, затем нижний), в результате чего формируются пары кодовых символов, образующих ответвленное слово, связанное с только что поступившим битом. Это выполняется для каждого входного бита. Выбор связи между сумматорами и разрядами регистра влияет на характеристики кода. Всякое изменение в выборе связей приводит в результате к различным кодам. Связь, конечно же, выбирается и изменяется *не* произвольным образом. Задача выбора связей, дающая оптимальные дистанционные свойства, сложна и в общем случае не решается; однако для всех значений длины кодового ограничения, меньших 20, с помощью компьютеров были найдены хорошие коды [3–5].

В отличие от блочных кодов, имеющих фиксированную длину слова n, в сверточных кодах нет определенного размера блока. Однако с помощью *периодического отбрасывания* сверточным кодам часто принудительно придают блочную структуру. Это требует некоторого количества нулевых разрядов, присоединенных к концу входной последовательности данных, которые служат для очистки (или *промывки*) регистра сдвига от бит данных. Поскольку добавленные нули не несут дополнительной инфор-

мации, *эффективная степень кодирования* будет ниже *k/n*. Чтобы степень кодирования оставалась близкой к *k/n*, период отбрасывания чаще всего делают настолько большим, насколько это возможно.

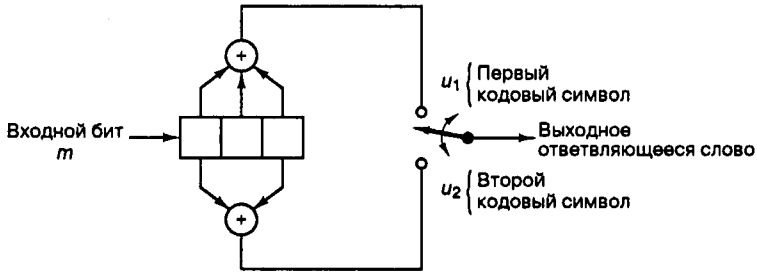


Рис. 7.3. Сверточный кодер (степень кодирования 1/2, K = 3)

Один из способов реализации кодера заключается в определении *n* векторов связи, по одному на каждый из *n* сумматоров по модулю 2. Каждый вектор имеет размерность *K* и описывает связь регистра сдвига кодера с соответствующим сумматором по модулю 2. Единица на *i*-й позиции вектора указывает на то, что соответствующий разряд в регистре сдвига связан с сумматором по модулю 2, а нуль в данной позиции указывает, что связи между разрядом и сумматором по модулю 2 не существует. Для кодера на рис. 7.2 можно записать вектор связи $\mathbf{g}_1$ для верхних связей, а $\mathbf{g}_2$ — для нижних.

$$\mathbf{g}_1 = 1\ 1\ 1$$
$$\mathbf{g}_2 = 1\ 0\ 1$$

Предположим теперь, что вектор сообщения $\mathbf{m} = 1\ 0\ 1$ закодирован с использованием сверточного кода и кодера, показанного на рис. 7.3. Введены три бита сообщения, по одному в момент времени t_1, t_2 и t_3, как показано на рис. 7.4. Затем для очистки регистра в моменты времени t_4 и t_5 введены $(K - 1) = 2$ нуля, что в результате приводит к смещению конечного участка на всю длину регистра. Последовательность на выходе выглядит следующим образом: 1 1 1 0 0 0 1 0 1 1, где крайний левый символ представляет первую передачу. Для декодирования сообщения нужна полная последовательность на выходе (включающая кодовые символы). Для удаления сообщения из кодера требуется на единицу меньше нулей, чем имеется разрядов в регистре, или $K - 1$ очищенных бит. В момент времени t_6 показан нулевой выход, это должно дать читателю возможность убедиться в том, что в момент времени t_5 регистр устанавливается в исходное состояние. Таким образом, в момент времени t_6 уже можно передавать новое сообщение.

7.2.1.1. Реакция кодера на импульсное возмущение

Мы можем описать кодер через его *импульсную характеристику*, т.е. в виде отклика кодера на единичный проходящий бит. Рассмотрим содержимое регистра (рис. 7.3) при прохождении через него двоичной единицы.

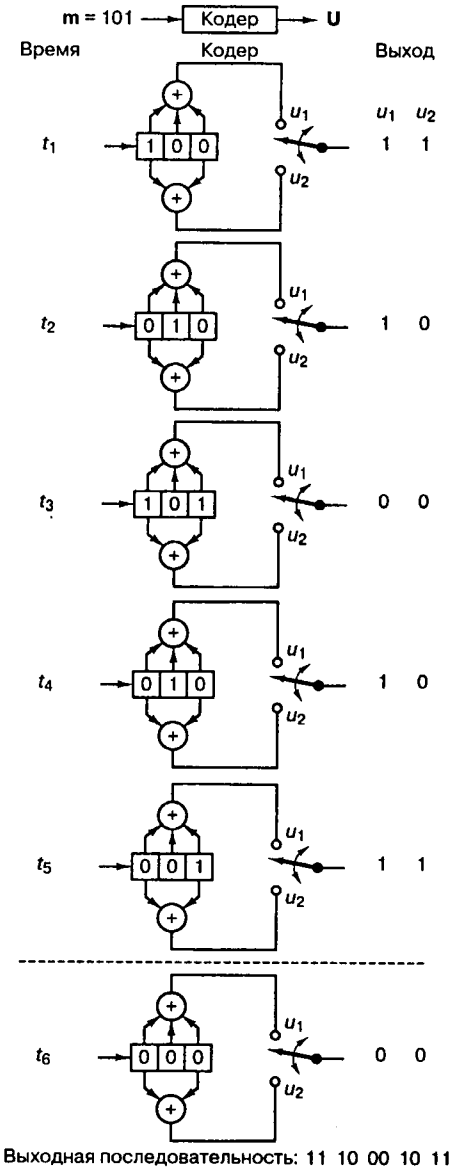


Рис. 7.4. Сверточное кодирование последовательности сообщения со степенью кодирования 1/2 кодером с $K = 3$.

	Содержимое регистра	Ответвляющееся слово u_1	Ответвляющееся слово u_2
	100	1	1
	010	1	0
	001	1	1
Входная последовательность	1 0 0		
Выходная последовательность	11 10 11		

Последовательность на выходе при единице на входе называется откликом кодера на импульсное возмущение, или его импульсной характеристикой. Для входной последовательности $\mathbf{m} = 1\ 0\ 1$ данные на выходе могут быть найдены путем *суперпозиции* или *линейного сложения* смещенных во времени входных "импульсов".

Вход, m	Выход				
1	11	10	11		
0		00	00	00	
1			11	10	11
Сумма по модулю 2	11	10	00	10	11

Обратите внимание на то, что эти данные на выходе такие же, как и на рис. 7.4, что указывает на *линейность сверточных кодов* — точно так же как и в блочных кодах в главе 6. Название *сверточный кодер* (convolutional encoder) возникло именно вследствие этого свойства генерации данных на выходе с помощью линейного сложения (или свертки) смещенных во времени импульсов последовательности на входе с импульсной характеристикой кодера. Такие устройства часто описываются с помощью матричного генератора бесконечного порядка [6].

Отметим, что в рассмотренном выше примере входящей последовательности из 3 бит и последовательности на выходе из 10 бит *эффективная степень кодирования* составляет $k/n = 3/10$, что значительно меньше величины 1/2, которую можно было бы ожидать, зная, что каждый бит данных на входе порождает пару канальных битов на выходе. Причина этого заключается в том, что окончательные биты данных нужно проводить через кодер. Все канальные биты на выходе нуждаются в процессе декодирования. Если бы сообщение было длиннее, скажем 300 бит, последовательность кодовых слов на выходе содержала бы 640 бит и значение для степени кодирования кода 300/640 было бы значительно ближе к 1/2.

7.2.1.2. Полиномиальное представление

Иногда связи кодера описываются с помощью *полиномиального генератора*, аналогичного используемому в главе 6 для описания реализации обратной связи регистра сдвига циклических кодов. Сверточный кодер можно представить в виде набора из n полиномиальных генераторов, по одному для каждого из n сумматоров по модулю 2. Каждый полином имеет порядок $K - 1$ или меньше и описывает связь кодирующего регистра сдвига с соответствующим сумматором по модулю 2, почти так же как и вектор связи. Коэффициенты возле каждого слагаемого полинома порядка $(K - 1)$ равны либо 1, либо 0, в зависимости от того, имеется ли связь между регистром сдвига и сумматором по модулю 2. Для кодера на рис. 7.3

можно записать полиномиальный генератор $\mathbf{g}_1(X)$ для верхних связей и $\mathbf{g}_2(X)$ — для нижних.

$$\mathbf{g}_1(X) = 1 + X + X^2$$
$$\mathbf{g}_2(X) = 1 + X^2$$

Здесь слагаемое самого нижнего порядка в полиноме соответствует входному разряду регистра. Выходная последовательность находится следующим образом.

$$\mathbf{U}(X) = \mathbf{m}(X)\mathbf{g}_1(X) \text{ чередуется с } \mathbf{m}(X)\mathbf{g}_2(X)$$

Прежде всего, выразим вектор сообщения $\mathbf{m} = 1\ 0\ 1$ в виде полинома, т.е. $\mathbf{m}(X) = 1 + X^2$. Для очистки регистра мы снова будем предполагать использование нулей, следующих за битами сообщения. Тогда выходящий полином $\mathbf{U}(X)$, или выходящая последовательность $\mathbf{U}$ кодера (рис. 7.3) для входящего сообщения $\mathbf{m}$ может быть найдена следующим образом.

$$\mathbf{m}(X)\mathbf{g}_1(X) = (1 + X^2)(1 + X + X^2) = 1 + X + X^3 + X^4$$
$$\mathbf{m}(X)\mathbf{g}_2(X) = (1 + X^2)(1 + X^2) = 1 + X^4$$

$\mathbf{m}(X)\mathbf{g}_1(X)$	=	1	+	X	+	$0X^2$	+	X^3	+ X^4
$\mathbf{m}(X)\mathbf{g}_2(X)$	=	1	+	$0X$	+	$0X^2$	+	$0X^3$	+ X^4
$\mathbf{U}(X)$	=	(1,1)	+	$(1,0)X$	+	$(0,0)X^2$	+	$(1,0)X^3$	+ $(1,1)X^4$
$\mathbf{U}$	=	11		10		00		10	11

В этом примере мы начали обсуждение с того, что сверточный кодер можно трактовать как набор *регистров сдвига циклического кода*. Мы представили кодер в виде *полиномиальных генераторов*, с помощью которых описываются циклические коды. Однако мы пришли к той же последовательности на выходе, что и на рис. 7.4, и к той же, что и в предыдущем разделе, полученной при описании реакции на импульсное возмущение. (Чтобы иметь лучшее представление о структуре сверточного кода в контексте линейной последовательной схемы, обратитесь к работе [7].)

7.2.2. Представление состояния и диаграмма состояний

Сверточный кодер принадлежит классу устройств, известных как *конечный автомат* (finite-state machine). Это общее название дано системам, обладающим памятью о прошедших сигналах. Прилагательное *конечный* показывает, что существует ограниченное число состояний, которое может возникнуть в системе. Что имеется в виду под *состоянием* (state) в системах с конечным его числом? В более общем смысле состояние включает наименьшее количество информации, на основе которой вместе с текущими входными данными можно определить данные на выходе системы. Состояние дает некоторое представление о прошлых событиях (сигналах) и об ограниченном наборе возможных выходных данных в будущем. Будущие состояния ограничиваются прошлыми состояниями. Для сверточного кода со степенью кодирования $1/n$ состояние представлено содержимым $K - 1$ крайних правых разрядов (рис. 7.4). Знание состояния плюс знание следующих данных на входе является необходимым и достаточным условием для определения данных на выходе. Итак, пусть состояние кодера в момент времени t_i определяется как $X_i = m_i - 1, m_i - 2, \ldots, m_i - K + 1$. i-я ветвь кодовых слов U_i полностью определяется состоянием X_i и введенными в настоящее время битами m_i; таким обра-

зом, состояние X_i описывает предысторию кодера для определения данных на его выходе. Состояния кодера считаются *Марковскими* в том смысле, что вероятность $P(X_i + 1|X_i, ..., X_0)$ нахождения в состоянии $X_i + 1$, определяемая всеми предыдущими состояниями, зависит только от самого последнего состояния X_i, т.е. она равна $P(X_i + 1|X_i)$.

Одним из способов представления простых кодирующих устройств является *диаграмма состояния* (state diagram); такое представление кодера, изображенного на рис. 7.3, показано на рис. 7.5. Состояния, показанные в рамках диаграммы, представляют собой возможное содержимое $K - 1$ крайних правых разрядов регистра, а пути между состояниями — ответвляющиеся слова на выходе, являющиеся результатом переходов между такими состояниями. Состояния регистра выбраны следующими: $a = 00$, $b = 10$, $c = 01$ и $d = 11$; диаграмма, показанная на рис. 7.5, иллюстрирует все возможные смены состояний для кодера, показанного на рис. 7.3. Существует всего два исходящих из каждого состояния перехода, соответствующие двум возможным входным битам. Далее для каждого пути между состояниями записано ответвляющееся слово на выходе, связанное с переходами между состояниями. При изображении путей, сплошной линией принято обозначать путь, связанный с нулевым входным битом, а пунктирной линией — путь, связанный с единичным входным битом. Отметим, что за один переход *невозможно* перейти из данного состояния в любое произвольное. Так как за единицу времени перемещается только один бит, существует только два возможных перехода между состояниями, в которые регистр может переходить за время прохождения каждого бита. Например, если состояние кодера — 00, при следующем смещении *возможно* возникновение *только* состояний 00 или 10.

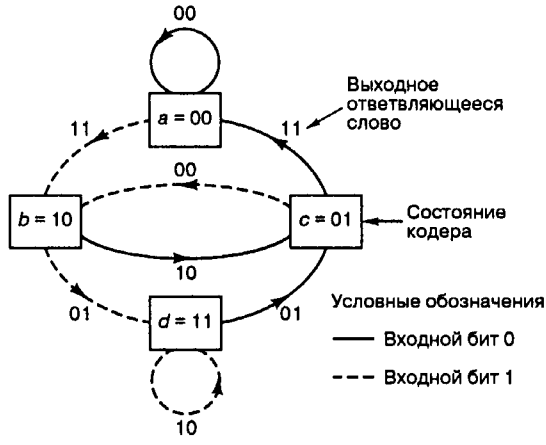


Рис. 7.5. Диаграмма состояний кодера (степень кодирования 1/2, $K = 3$)

Пример 7.1. Сверточное кодирование

Для кодера, показанного на рис. 7.3, найдите изменение состояний и результирующую последовательность кодовых слов **U** для последовательности сообщений $\mathbf{m} = 1\ 1\ 0\ 1\ 1$, за которой следует $K - 1 = 2$ нуля для очистки регистра. Предполагается, что в исходном состоянии регистр содержит одни нули.

Решение

Входные биты, m_i	Содержимое регистра	Состояние в момент времени t_i	Состояние в момент времени t_{i+1}	Ответвляющееся слово в момент времени t_i	
				u_1	u_2
—	0 0 0	0 0	0 0	—	
1	1 0 0	0 0	1 0	1	1
1	1 1 0	1 0	1 1	0	1
0	0 1 1	1 1	0 1	0	1
1	1 0 1	0 1	1 0	0	1
1	1 1 0	1 0	1 1	0	1
0	0 1 1	1 1	0 1	0	1
0	0 0 1	0 1	0 0	1	1
	$\underbrace{}_{t_i}$				
	$\underbrace{}_{t_{i+1}}$				

Последовательность на выходе $\mathbf{U} = 1\,1 \quad 0\,1 \quad 0\,1 \quad 0\,0 \quad 0\,1 \quad 0\,1 \quad 1\,1$

Пример 7.2. Сверточное кодирование

В примере 7.1 исходное содержимое регистра — все нули. Это эквивалентно тому, что данной последовательности на входе предшествовали два нулевых бита (кодирование является функцией настоящих информационных бит и $K - 1$ предыдущих бит). Повторите задание примера 7.1, предполагая, что данной последовательности предшествовали два единичных бита, и убедитесь, что теперь последовательность кодовых слов **U** для входящей последовательности $\mathbf{m} = 1\,1\,0\,1\,1$ отличается от последовательности, найденной в примере 7.1.

Решение

Запись "×" обозначает "неизвестно".

Входные биты, m_i	Содержимое регистра	Состояние в момент времени t_i	Состояние в момент времени t_{i+1}	Ответвляющееся слово в момент времени t_i	
				u_1	u_2
—	1 1 ×	1 ×	1 1	—	
1	1 1 1	1 1	1 1	1	0
1	1 1 1	1 1	1 1	1	0
0	0 1 1	1 1	0 1	0	1
1	1 0 1	0 1	1 0	0	0
1	1 1 0	1 0	1 1	0	1
0	0 1 1	1 1	0 1	0	1
0	0 0 1	0 1	0 0	1	1
	$\underbrace{}_{t_i}$				
	$\underbrace{}_{t_{i+1}}$				

Последовательность на выходе $\mathbf{U} = 1\,0 \quad 1\,0 \quad 0\,1 \quad 0\,0 \quad 0\,1 \quad 0\,1 \quad 1\,1$

Сравнивая эти результаты с результатами из примера 7.1, можно видеть, что каждое ответвленное слово выходной последовательности **U** является функцией *не только* входного бита, но и предыдущих $K - 1$ бит.

7.2.3. Древовидные диаграммы

Несмотря на то что диаграммы состояний полностью описывают кодер, по сути, их нельзя использовать для легкого отслеживания переходов кодера в зависимости от времени, поскольку диаграмма не представляет динамики изменений. Древовидная диаграмма (tree diagram) прибавляет к диаграмме состояния *временное измерение*. Древовидная диаграмма сверточного кодера, показанного на рис. 7.3, изображена на рис. 7.6. В каждый последующий момент прохождения входящего бита процедура кодирования может быть описана с помощью перемещения по диаграмме слева направо, причем каждая ветвь дерева описывает ответвленное слово на выходе. Правило ветвления для нахождения последовательности кодовых слов следующее: если входящим битом является нуль, то он связывается со словом, которое находится путем перемещения в следующую (по направлению вверх) крайнюю правую ветвь; если входящий бит — это единица, то ответвленное слово находится путем перемещения в следующую (по направлению вниз) крайнюю правую ветвь. Предполагается, что первоначально кодер содержал одни нули. Диаграмма показывает, что если первым входным битом был нуль, то ответвляющимся словом на выходе будет 00, а если первым входным битом была единица, то ответвляющимся словом на выходе будет 11. Аналогично, если первым входным битом была единица, а вторым — нуль, на выходе вторым ответвляющимся словом будет 10. Если первым входным битом была единица и вторым входным битом была единица, вторым ответвляющимся словом на выходе будет 01. Следуя этой процедуре, видим, что входящая последовательность 1 1 0 1 1 представляется жирной линией, нарисованной на древовидной диаграмме (рис. 7.6). Этот путь соответствует выходной последовательности кодовых слов 1 1 0 1 0 1 0 0 0 1.

Дополнительное временное измерение в древовидной диаграмме (по сравнению с диаграммой состояния) допускает динамическое описание кодера как функции конкретной входной последовательности. Однако заметили ли вы, что при попытке описания с помощью древовидной диаграммы последовательности произвольной длины возникает проблема? Число ответвлений растет как 2^L, где L — это количество ответвляющихся слов в последовательности. При большом L вы бы очень быстро исписали бумагу и исчерпали терпение.

7.2.4. Решетчатая диаграмма

Исследование древовидной диаграммы на рис. 7.6 показывает, что в этом примере после третьего ветвления в момент времени t_4 структура повторяется (древовидная структура *повторяется после K ответвлений*, где K — длина кодового ограничения). Пометим каждый узел в дереве (рис. 7.6), ставя в соответствие четыре возможных состояния в регистре сдвига: $a = 00$, $b = 10$, $c = 01$ и $d = 11$. Первое ветвление древовидной структуры в момент времени t_1 дает пару узлов, помеченных как a и b. При каждом последующем ветвлении количество узлов удваивается. Второе ветвление в момент времени t_2 дает в результате четыре узла, помеченных как a, b, c и d. После *третьего* ветвления всего имеется восемь узлов: два — a, два — b, два — c и два — d.

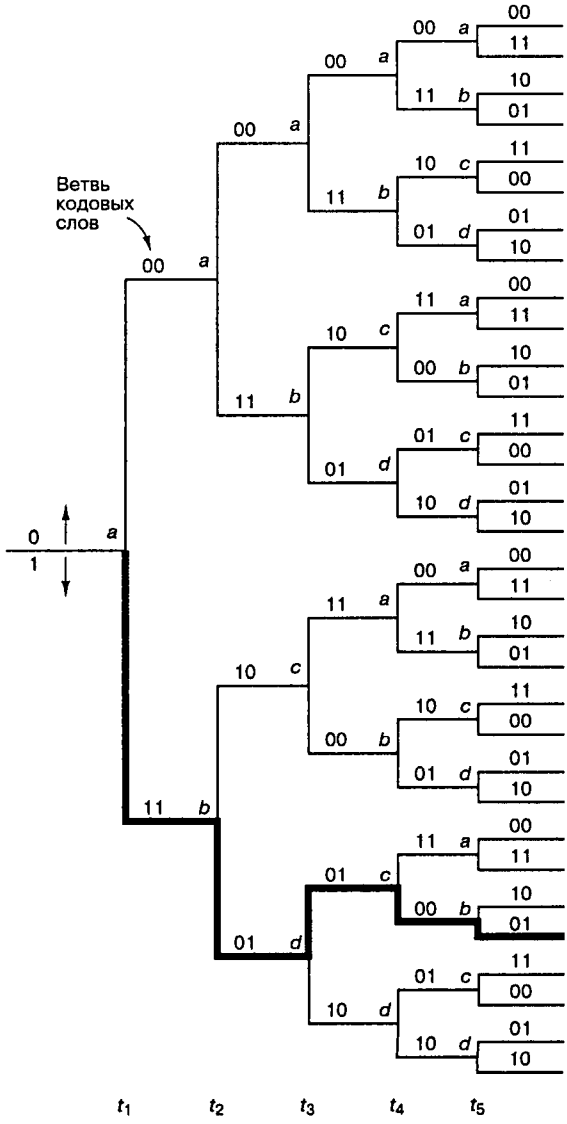


Рис. 7.6. Древовидное представление кодера (степень кодирования 1/2, $K = 3$)

Можно видеть, что все ветви выходят из двух узлов одного и того же состояния, образуя идентичные ветви последовательностей кодовых слов. В этот момент дерево делится на идентичные верхнюю и нижнюю части. Смысл этого становится яснее после рассмотрения кодера, изображенного на рис. 7.3. Когда четвертый входной бит входит в кодер слева, первый входной бит справа выбрасывается и больше не влияет на ответвленные слова на выходе. Следовательно, входящие последовательности 1 0 0 x y ... и 0 0 0 x y ..., где крайний левый бит является самым ранним, после ($K = 3$)-го ветвления генерируют одинаковые ответвляющиеся слова. Это означает, что любые состояния, имеющие одинаковую метку в один и тот же момент t_i, можно соединить, поскольку все последующие

пути будут неразличимы. Если мы проделаем это для древовидной структуры, показанной на рис. 7.6, получим иную диаграмму, называемую решетчатой. *Решетчатая диаграмма*, которая использует повторяющуюся структуру, дает более удобное описание кодера, по сравнению с древовидной диаграммой. Решетчатая диаграмма для сверточного кодера, изображенного на рис. 7.3, показана на рис. 7.7

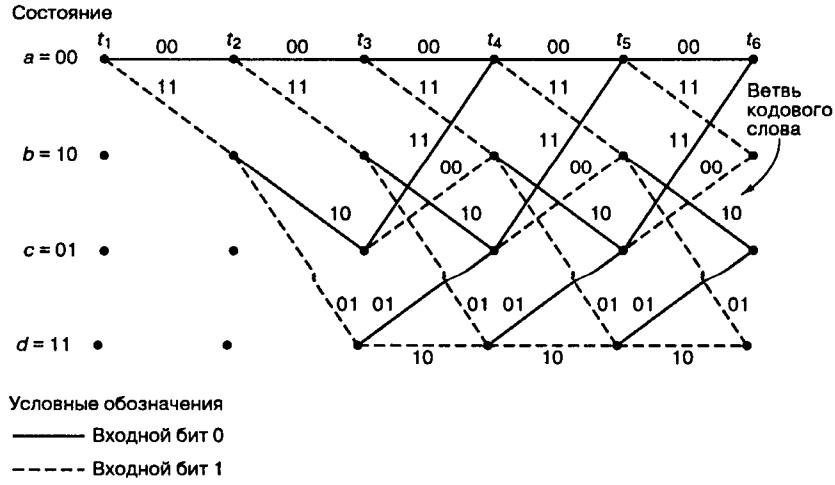


Рис. 7.7. Решетчатая диаграмма кодера (степень кодирования 1/2, $K = 3$)

При изображении решетчатой диаграммы мы воспользовались теми же условными обозначениями, что и для диаграммы состояния: сплошная линия обозначает выходные данные, генерируемые входным нулевым битом, а пунктирная — выходные данные, генерируемые входным единичным битом. Узлы решетки представляют состояния кодера; первый ряд узлов соответствует состоянию $a = 00$, второй и последующие — состояниям $b = 10$, $c = 01$ и $d = 11$. В каждый момент времени для представления 2^{K-1} возможных состояний кодера решетка требует 2^{K-1} узлов. В нашем примере после достижения глубины решетки, равной трем (в момент времени t_4), замечаем, что решетка имеет фиксированную периодическую структуру. В общем случае фиксированная структура реализуется после достижения глубины K. Следовательно, с этого момента в каждое состояние можно войти из любого из двух предыдущих состояний. Также из каждого состояния можно перейти в одно из двух состояний. Из двух исходящих ветвей одна соответствует нулевому входному биту, а другая — единичному входному биту. На рис. 7.7 ответвляющиеся слова на выходе соответствуют переходам между состояниями, показанными как метки на ветвях решетки.

Один столбец временного интервала сформировавшейся решетчатой структуры кодирования полностью определяет код. Несколько столбцов показаны исключительно для визуализации последовательности кодовых символов как функции времени. Состояние сверточного кодера представлено содержанием крайних правых $K - 1$ разрядов в регистре кодера. Некоторые авторы описывают состояние с помощью крайних левых $K - 1$ разрядов. Какое описание правильно? Они оба верны. Каждый переход имеет начальное и конечное состояние. Крайние правые $K - 1$ разрядов описывают начальное состояние для текущих входных данных, которые находятся в крайнем левом разряде (степень кодирования предполагается равной $1/n$). Крайние левые $K - 1$

разрядов являются конечным состоянием для такого перехода. Последовательность кодовых символов характеризуется N ветвями (что представляет N бит данных), занимающими N интервалов времени. Она связана с конкретным состоянием в каждый из $N+1$ интервалов времени (от начала до конца). Таким образом, мы запускаем биты в моменты времени $t_1, t_2, ..., t_N$ и интересуемся метрикой состояния в моменты времени $t_1, t_2, ..., t_{N+1}$. Здесь использовано следующее условие: текущий бит располагается в крайнем левом разряде, а крайние правые $K-1$ разрядов стартуют из состояния со всеми нулями. Этот момент времени обозначим как *начальное время*, t_1. Время завершения последнего перехода обозначим как *время прекращения работы*, t_{N+1}.

7.3. Формулировка задачи сверточного кодирования

7.3.1. Декодирование по методу максимального правдоподобия

Если все входные последовательности сообщений равновероятны, минимальная вероятность ошибки получается при использовании декодера, который сравнивает условные вероятности и выбирает максимальную. Условные вероятности также называют *функциями правдоподобия* $P(\mathbf{Z}|\mathbf{U}^{(m)})$, где $\mathbf{Z}$ — это принятая последовательность, а $\mathbf{U}^{(m)}$ — одна из возможных переданных последовательностей. Декодер выбирает $\mathbf{U}^{(m')}$, если

$$P(\mathbf{Z}|\mathbf{U}^{(m')}) = \max_{\text{по всем } \mathbf{U}^{(m)}} P(\mathbf{Z}|\mathbf{U}^{(m)}) \tag{7.1}$$

Принцип *максимального правдоподобия*, определяемый уравнением (7.1), является фундаментальным достижением теории принятия решений (см. приложение Б); это формализация способа принятия решений, основанного на "здравом смысле", когда имеются статистические данные о вероятностях. При рассмотрении двоичной демодуляции в главах 3 и 4, предполагалась передача только двух равновероятных сигналов $s_1(t)$ и $s_2(t)$. Следовательно, принятие двоичного решения на основе принципа максимального правдоподобия, касающееся данного полученного сигнала, означает, что в качестве переданного сигнала выбирается $s_1(t)$, если

$$p(z|s_1) > p(z|s_2).$$

В противном случае считается, что передан был сигнал $s_2(t)$. Параметр z представляет собой величину $z(T)$, значение принятого сигнала до детектирования в конце каждого периода передачи символа $t = T$. Однако при использовании принципа максимального правдоподобия в задаче сверточного декодирования, в сверточном коде обнаруживается наличие памяти (полученная последовательность является суперпозицией текущих и предыдущих двоичных разрядов). Таким образом, применение принципа максимального правдоподобия при декодировании бит данных, закодированных сверточным кодом, осуществляется в контексте выбора *наиболее вероятной последовательности*, как показано в уравнении (7.1). Обычно имеется *множество* возможных переданных последовательностей кодовых слов. Что касается двоичного кода, то последовательность из L ответвленных слов является членом набора из 2^L возможных последовательностей. Следовательно, в контексте максимального правдоподобия можно сказать, что в качестве переданной последовательности декодер выбирает $\mathbf{U}^{(m')}$, если вероятность $P(\mathbf{Z}|\mathbf{U}^{(m')})$ больше вероятности всех остальных возможно переданных последовательностей. Такой оп-

тимальный декодер, минимизирующий вероятность ошибки (когда все переданные последовательности равновероятны), известен как *декодер, работающий по принципу максимального правдоподобия* (maximum likelihood detector). Функция правдоподобия задается или вычисляется, исходя из спецификации канала.

Предположим, что мы имеем дело с аддитивным белым гауссовым шумом с нулевым средним в канале без памяти, т.е. шум влияет на каждый символ кода *независимо* от остальных символов. При степени кодирования сверточного кода, равной $1/n$, правдоподобие можно выразить следующим образом.

$$P(\mathbf{Z}|\mathbf{U}^{(m)}) = \prod_{i=1}^{\infty} P(Z_i|U_i^{(m)}) = \prod_{i=1}^{\infty}\prod_{j=1}^{n} P(z_{ji}|u_{ji}^{(m)}) \tag{7.2}$$

Здесь Z_i — это i-я ветвь полученной последовательности $\mathbf{Z}$, $U_i^{(m)}$ — это ветвь отдельной последовательности кодовых слов $\mathbf{U}^{(m)}$, z_{ji} — это j-й кодовый символ Z_i, $u_{ji}^{(m)}$ — это j-й кодовый символ $U_i^{(m)}$, а каждая ветвь состоит из n кодовых символов. Задача декодирования заключается в выборе пути сквозь решетку, показанную на рис. 7.7 (каждый возможный путь определяет последовательность кодовых слов), таким образом, чтобы произведение

$$\prod_{i=1}^{\infty}\prod_{j=1}^{n} P(z_{ji}|u_{ji}^{(m)}) \text{ было максимальным}. \tag{7.3}$$

Как правило, при вычислениях удобнее пользоваться логарифмом функции правдоподобия, поскольку это позволяет произведение заменить суммированием. Мы можем воспользоваться таким преобразованием, поскольку логарифм является монотонно возрастающей функцией и, следовательно, не внесет изменений в выбор окончательного кодового слова. Логарифмическую функцию правдоподобия можно определить следующим образом.

$$\gamma_{\mathbf{U}}(m) = \lg P(\mathbf{Z}|\mathbf{U}^{(m)}) = \sum_{i=1}^{\infty} \lg P(Z_i|U_i^{(m)}) = \sum_{i=1}^{\infty}\sum_{j=1}^{n} \lg P(z_{ji}|u_{ji}^{(m)}) \tag{7.4}$$

Теперь задача декодирования заключается в выборе пути вдоль дерева на рис. 7.6 или решетки на рис. 7.7 таким образом, чтобы $\gamma_{\mathbf{U}}(m)$ было максимальным. При декодировании сверточных кодов можно использовать как древовидную, так и решетчатую структуру. При древовидном представлении кода игнорируется то, что пути снова объединяются. Для двоичного кода количество возможных последовательностей, состоящих из L ответвленных слов, равно 2^L. Поэтому декодирование полученных последовательностей, основанное на принципе максимального правдоподобия с использованием древовидной диаграммы, требует метода "грубой силы" или исчерпывающего сопоставления 2^L накопленных логарифмических метрик правдоподобия, описывающих все варианты возможных последовательностей кодовых слов. Поэтому рассматривать декодирование на основе принципа максимального правдоподобия с помощью древовидной структуры практически невозможно. В предыдущем разделе было показано, что при решетчатом представлении кода декодер можно построить так, чтобы можно было отказываться от путей, которые не могут быть кандидатами на роль максимально правдоподобной последовательности. Путь декодирования выбирается из некоего сокращенного набора *выживших путей*. Такой декодер тем не менее

является оптимальным; в том смысле, что путь декодирования такой же, как и путь, полученный с помощью декодера критерия максимального правдоподобия, действующего "грубой силой", однако предварительный отказ от неудачных путей снижает сложность декодирования.

В качестве великолепного пособия для изучения структуры сверточных кодов, декодирования на основе критерия максимального правдоподобия и реализации кода можно порекомендовать работу [8]. Существует несколько алгоритмов, которые дают *приблизительные* решения задачи декодирования на основе критерия максимального правдоподобия, включая последовательный [9, 10] и пороговый [11]. Каждый из этих алгоритмов является подходящим для узкоспециальных задач; однако все они близки к оптимальному. *Алгоритм декодирования Витерби*, напротив, осуществляет декодирование на основе критерия максимального правдоподобия и, следовательно, является оптимальным. Это не означает, что алгоритм Витерби в любой реализации является наилучшим; при его использовании существуют жесткие условия, налагаемые на аппаратное обеспечение. Алгоритм Витерби обсуждается в разделах 7.3.3. и 7.3.4.

7.3.2. Модели каналов: мягкое или жесткое принятие решений

Перед тем как начать разговор об алгоритме, который задает схему принятия максимально правдоподобного решения, давайте сначала опишем канал. Последовательность кодовых слов $\cdot\mathbf{U}^{(m)}$, определяемую ответвленными словами, каждое из которых состоит из n кодовых символов, можно рассматривать как бесконечный поток, в отличие от блочного кода, где исходные данные и их кодовые слова делятся на блоки строго определенного размера. *Последовательность кодовых слов, показанная на рис. 7.1*, выдается сверточным кодером и подается на модулятор, где кодовые символы преобразуются в сигналы. Модуляция может быть узкополосной (например, модуляция импульсными сигналами) или полосовой (например, модуляция PSK или FSK). Вообще, за такт в сигнал $s_i(t)$ преобразуется l символов, где l — целое, причем $i = 1, 2, ...,$ а $M = 2^l$. Если $l = 1$, модулятор преобразует каждый кодовый символ в двоичный сигнал. Предполагается, что канал, по которому передается сигнал, искажает сигнал гауссовым шумом. После того как искаженный сигнал принят, он сначала обрабатывается демодулятором, а затем подается на декодер.

Рассмотрим ситуацию, когда двоичный сигнал передается за отрезок времени $(0, T)$, причем двоичная единица представляется сигналом $s_1(t)$, а двоичный нуль — сигналом $s_2(t)$. Принятый сигнал имеет вид $r(t) = s_i(t) + n(t)$, где $n(t)$ представляет собой вклад гауссового шума с нулевым средним. В главе 3 мы описывали обнаружение $r(t)$ в два основных этапа. На первом этапе принятый сигнал переводится в число $z(T) = a_i + n_0$, где a_i — это компонент сигнала $z(T)$, а n_0 — компонент шума. Компонент шума n_0 — это *случайная переменная*, значения которой имеют *гауссово* распределение с нулевым средним. Следовательно, $z(T)$ также будет *случайной гауссовой величиной* со средним a_1 или a_2, в зависимости от того, какая величина была отправлена — двоичная единица или двоичный нуль. На втором этапе процесса обнаружения принимается решение о том, какой сигнал был передан. Это решение принимается на основе сравнения $z(T)$ с порогом. Условные вероятности $z(T)$, $p(z|s_1)$ и $p(z|s_2)$, показанные на рис. 7.8, обозначены как правдоподобие s_1 и s_2. Демодулятор, представленный на рис. 7.1, преобразует упорядоченный по времени набор случайных переменных $\{z(T)\}$ в кодовую последовательность $\mathbf{Z}$ и подает ее на декодер. Выход демодулятора можно

настроить по-разному. Можно реализовать его в виде *жесткой схемы принятия решений* относительно того, представляет ли $z(T)$ единицу или нуль. В этом случае выход демодулятора квантуется на два уровня, нулевой и единичный, и соединяется с декодером (это абсолютно та же схема пороговых решений, о которой шла речь в главах 3 и 4). Поскольку декодер работает в режиме жесткой схемы принятия решений, принятых демодулятором, такое декодирование называется *жестким*.

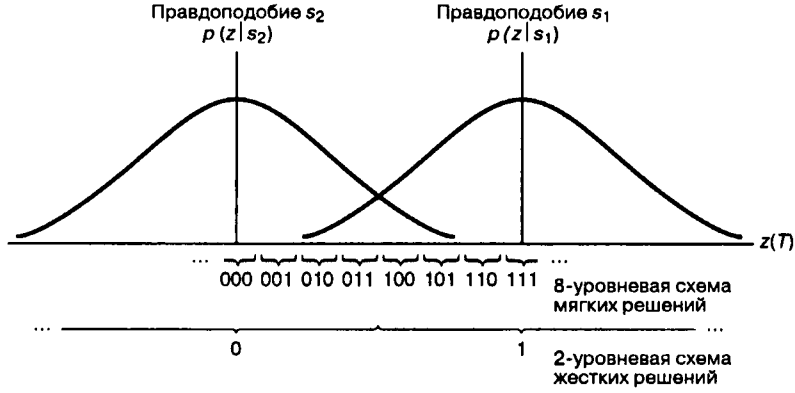


Рис. 7.8. Жесткая и мягкая схемы декодирования

Аналогично демодулятор можно настроить так, чтобы он подавал на декодер *значение* $z(T)$, *квантованное более чем на два уровня*. Такая схема обеспечивает декодер большим количеством информации, чем жесткая схема решений. Если выход демодулятора имеет более двух уровней квантования, то декодирование называется *мягким*. На рис. 7.8 на оси абсцисс изображено восемь (3-битовых) уровней квантования. Если в демодуляторе реализована жесткая схема принятия двоичных решений, он отправляет на декодер только один двоичный символ. Если в демодуляторе реализована мягкая двоичная схема принятия решений, квантованная на восемь уровней, он отправляет на декодер 3-битовое слово, описывающее интервал, соответствующий $z(T)$. По сути, поступление такого 3-битового слова, вместо одного двоичного символа, эквивалентно передаче декодеру *меры достоверности* вместе с решением относительно кодового символа. Согласно рис. 7.8, если с демодулятора поступила на декодер последовательность 1 1 1, это равносильно утверждению, что с очень высокой степенью достоверности кодовым символом была 1, в то время как переданная последовательность 1 0 0 равносильна утверждению, что с очень низкой степенью достоверности кодовым символом была 1. Совершенно ясно, что в конечном счете каждое решение, принятое декодером и касающееся сообщения, должно быть жестким; в противном случае на распечатках компьютера можно было бы увидеть нечто, подобное следующему: "думаю, это 1", "думаю, это 0" и т.д. То, что после демодулятора *не принимается жесткое решение* и на декодер поступает больше данных (мягкое принятие решений), можно понимать как промежуточный этап, необходимый для того, чтобы на декодер поступило больше информации, с помощью которой он затем сможет восстановить последовательность сообщения (с более высокой достоверностью передачи сообщения по сравнению с декодированием в рамках жесткой схемы принятия решений). Показанная на рис. 7.8, 8-уровневая метрика мягкой схемы принятия решений часто обозначается как –7, –5, –3, –1, 1, 3, 5, 7. Такие обозначения вводятся для простоты ин-

терпретации мягкой схемы принятия решения. Знак метрики характеризует решение (например, выбирается s_1, если величина положительна, и s_2, если отрицательна), а величина метрики описывает степень достоверности этого решения. Преимуществом метрики, показанной на рис. 7.8, является только то, что в ней не используются отрицательные числа.

Для гауссова канала восьмиуровневое квантование, по сравнению с двухуровневым, приводит в результате к улучшению на 2 дБ требуемого отношения сигнал/шум. Это означает, что восьмиуровневое квантование с мягкой схемой принятия решений может дать ту же вероятность появления ошибочного бита, что и декодирование с жесткой схемой принятия решений, однако требует на 2 дБ *меньшего* значения E_b/N_0 при прочих равных характеристиках. Аналоговое квантование (или квантование с бесконечным числом уровней) дает в результате улучшение на 2,2 дБ, по сравнению с двухуровневым; следовательно, при восьмиуровневом квантовании, по сравнению с квантованием с бесконечным числом уровней, теряется приблизительно 0,2 дБ. По этой причине квантование более чем на восемь уровней может дать только небольшое улучшение производительности [12]. Какова цена, которую следует заплатить за такое улучшение параметров декодирования с мягкой схемой принятия решений? В случае декодирования с жесткой схемой принятия решений, для описания каждого кодового символа используется один бит, в то время как при восьмиуровневой мягкой схеме принятия решения для описания каждого символа применяется 3 бит; следовательно, в течение процесса декодирования нужно успеть обработать в три раза больше данных. Поэтому за мягкое декодирование приходится платить увеличением требуемых объемов памяти (и, возможно, возникнут проблемы со скоростью обработки).

В настоящее время существуют блочные и сверточные алгоритмы декодирования, функционирующие на основе жесткой *или* мягкой схемы принятия решений. Однако при блочном декодировании мягкая схема принятия решений, как правило, не используется, поскольку ее значительно сложнее реализовать, чем схему жесткого принятия решений. Чаще всего мягкая схема принятия решений применяется в *алгоритме сверточного декодирования Витерби*, поскольку при декодировании Витерби мягкое принятие решений лишь незначительно усложняет вычисления.

7.3.2.1. Двоичный симметричный канал

Двоичный симметричный канал (binary symmetric channel — BSC) — это дискретный канал без памяти (см. раздел 6.3.1), имеющий на входе и выходе двоичный алфавит и симметричные вероятности перехода. Как показано на рис. 7.9, его можно описать с помощью условных вероятностей.

$$P(0|1) = P(1|0) = p$$

$$P(1|1) = P(0|0) = 1 - p \tag{7.5}$$

Вероятность того, что выходной символ будет отличаться от входного, равна p, а вероятность того, что выходной символ будет идентичен входному, равна $(1-p)$. Канал BSC является примером *канала с жесткой схемой принятия решений*; это, в свою очередь, означает, что даже если демодулятор получил сигнал с непрерывным значением, BSC позволяет принять только какое-то одно определенное решение, так что каждый символ z_{ji} на выходе демодулятора, как показано на рис. 7.1, содержит одно из двух двоичных значений. Индексы величины z_{ji} указывают на j-й кодовый символ i-го ответвленного слова Z_i. Далее демодулятор передает последовательность $\mathbf{Z} = \{Z_i\}$ на декодер.

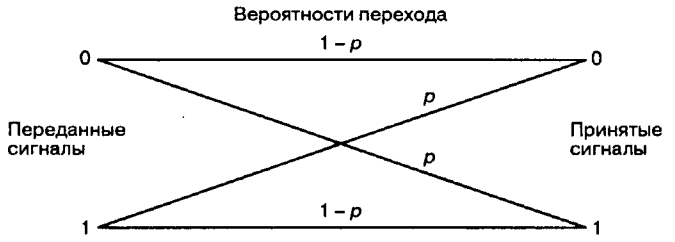


Рис. 7.9. Двоичный симметричный канал (канал с жесткой схемой принятия решений)

Пусть $\mathbf{U}^{(m)}$ — это переданное по каналу BSC кодовое слово с вероятностью появления ошибочного символа p, $\mathbf{Z}$ — соответствующая последовательность, полученная декодером. Как отмечалось ранее, декодер, работающий по принципу максимального правдоподобия, выбирает кодовое слово $\mathbf{U}^{(m')}$, имеющее максимальное правдоподобие $P(\mathbf{Z}|\mathbf{U}^{(m)})$ или его логарифм. Для BSC это эквивалентно выбору кодового слова $\mathbf{U}^{(m')}$, находящегося на наименьшем *расстоянии Хэмминга* от $\mathbf{Z}$ [8]. Расстояние Хэмминга — это удобная метрика для описания расстояния или степени сходства между $\mathbf{U}^{(m)}$ и $\mathbf{Z}$. Из всех возможных переданных последовательностей $\mathbf{U}^{(m)}$ декодер выбирает такую последовательность $\mathbf{U}^{(m')}$, для которой расстояние до $\mathbf{Z}$ минимально. Предположим, что каждая из последовательностей $\mathbf{U}^{(m)}$ и $\mathbf{Z}$ имеет длину L бит и отличается на d_m позиций (т.е. расстояние Хэмминга между $\mathbf{U}^{(m)}$ и $\mathbf{Z}$ равно d_m). Тогда, поскольку предполагалось, что канал не имеет памяти, вероятность того, что $\mathbf{U}^{(m)}$ преобразовалось в $\mathbf{Z}$, находящееся на расстоянии d_m от $\mathbf{U}^{(m)}$, может быть записана в следующем виде.

$$P(\mathbf{Z}|\mathbf{U}^{(m)}) = p^{d_m}(1-p)^{L-d_m} \tag{7.6}$$

Логарифмическая функция правдоподобия будет иметь следующий вид.

$$\lg P(\mathbf{Z}|\mathbf{U}^{(m)}) = -d_m \lg\left(\frac{1-p}{p}\right) + L\lg(1-p) \tag{7.7}$$

Если вычислить эту величину для каждой возможно переданной последовательности, последнее слагаемое в уравнении будет постоянным для всех случаев. Если предположить, что $p < 0{,}5$, уравнение (7.7) можно записать в следующей форме.

$$\lg P(\mathbf{Z}|\mathbf{U}^{(m)}) = -Ad_m - B \tag{7.8}$$

Здесь A и B — положительные константы. Следовательно, такой выбор кодового слова $\mathbf{U}^{(m')}$, чтобы расстояние Хэмминга до полученной последовательности $\mathbf{Z}$ было минимальным, соответствует *максимизации метрики правдоподобия или логарифма правдоподобия*. Следовательно, в канале BSC метрика логарифма правдоподобия легко заменяется расстоянием Хэмминга, а декодер, работающий по принципу максимального правдоподобия, будет выбирать на древовидной или решетчатой диаграмме путь, соответствующий *минимальному расстоянию Хэмминга* между последовательностью $\mathbf{U}^{(m')}$ и полученной последовательностью $\mathbf{Z}$.

7.3.2.2. Гауссов канал

Для гауссова канала каждый выходной символ демодулятора z_{ji}, как показано на рис. 7.1, принимает значения из непрерывного алфавита. Символ z_{ji} нельзя пометить

для обнаружения как правильное или неправильное решение. Передачу на декодер таких мягких решений можно рассматривать как поступление семейства условных вероятностей различных символов (см. раздел 6.3.1). Можно показать [8], что максимизация $P(\mathbf{Z}|\mathbf{U}^{(m)})$ эквивалентна максимизации скалярного произведения последовательности кодовых слов $\mathbf{U}^{(m)}$ (состоящей из двоичных символов, представленных как биполярные значения) и аналогового значения полученной последовательности $\mathbf{Z}$. Таким образом, декодер выбирает кодовое слово $\mathbf{U}^{(m')}$, если выражение

$$\sum_{i=1}^{\infty}\sum_{j=1}^{n} z_{ji}u_{ji}^{(m)} \tag{7.9}$$

имеет максимальное значение. Это эквивалентно выбору кодового слова $\mathbf{U}^{(m')}$, находящегося на ближайшем *евклидовом кодовом расстоянии* от $\mathbf{Z}$. Даже несмотря на то что каналы с жестким и мягким принятием решений требуют различных метрик, концепция выбора кодового слова $\mathbf{U}^{(m')}$, ближайшего к полученной последовательности $\mathbf{Z}$, одинакова для обоих случаев. Чтобы в уравнении (7.9) точно выполнить максимизацию, декодер должен осуществлять арифметические операции с аналоговыми величинами. Это непрактично, поскольку обычно декодеры являются цифровыми. Таким образом, необходимо дискретизировать полученные символы z_{ji}. Не напоминает ли вам уравнение (7.9) демодуляционную обработку, рассмотренную в главах 3 и 4? Уравнение (7.9) является дискретным вариантом корреляции входного полученного сигнала $r(t)$ с опорным сигналом $s_i(t)$, которая выражается уравнением (4.15). Квантованный гауссов канал, обычно называемый *каналом с мягкой схемой решений*, — это модель канала, в которой предполагается, что декодирование осуществляется на основе описанной ранее мягкой схемы принятия решения.

7.3.3. Алгоритм сверточного декодирования Витерби

Алгоритм декодирования Витерби был открыт и проанализирован Витерби (Viterbi) [13] в 1967 году. В алгоритме Витерби, по сути, реализуется декодирование, основанное на принципе максимального правдоподобия; однако в нем уменьшается вычислительная нагрузка за счет использования особенностей структуры конкретной решетки кода. Преимущество декодирования Витерби, по сравнению с декодированием по методу "грубой силы", заключается в том, что сложность декодера Витерби не является функцией количества символов в последовательности кодовых слов. Алгоритм включает в себя вычисление *меры подобия* (или *расстояния*), между сигналом, полученным в момент времени t_1, и всеми путями решетки, входящими в каждое состояние в момент времени t_1. В алгоритме Витерби не рассматриваются те пути решетки, которые, согласно принципу максимального правдоподобия, заведомо не могут быть оптимальными. Если в одно и то же состояние входят два пути, выбирается тот, который имеет лучшую метрику; такой путь называется *выживающим*. Отбор выживающих путей выполняется для каждого состояния. Таким образом, декодер углубляется в решетку, принимая решения путем исключения менее вероятных путей. Предварительный отказ от маловероятных путей упрощает процесс декодирования. В 1969 году Омура (Omura) [14] показал, что алгоритм Витерби — это, фактически, максимальное правдоподобие. Отметим, что задачу отбора оптимальных путей можно выразить как выбор кодового слова с *максимальной метрикой правдоподобия* или *минимальной метрикой расстояния*.

7.3.4. Пример сверточного декодирования Витерби

Для простоты предположим, что мы имеем дело с каналом BSC; в таком случае приемлемой мерой расстояния будет расстояние Хэмминга. Кодер для этого примера показан на рис. 7.3, а решетчатая диаграмма — на рис. 7.7. Для представления декодера, как показано на рис. 7.10, можно воспользоваться подобной решеткой. Мы начинаем в момент времени t_1 в состоянии 00 (вследствие очистки кодера между сообщениями декодер находится в начальном состоянии). Поскольку в этом примере возможны только два перехода, начинающиеся в некотором состоянии, для начала не нужно показывать все ветви. Полная решетчатая структура образуется после момента времени t_3. Принцип работы происходящего после процедуры декодирования можно понять, изучив решетку кодера на рис. 7.7 и решетку декодера, показанную на рис. 7.10. Для решетки декодера каждую ветвь за каждый временной интервал удобно пометить *расстоянием Хэмминга* между полученным кодовым символом и ответвляющимся словом, соответствующим той же ветви из решетки кодера. На рис. 7.10 показана последовательность сообщений $\mathbf{m}$, соответствующая последовательности кодовых слов $\mathbf{U}$, и искаженная шумом последовательность $\mathbf{Z} = 11\ 01\ 01\ 10\ 01\ \ldots$. Как показано на рис. 7.3, кодер характеризуется кодовыми словами, находящимися на ветвях решетки кодера и заведомо известными как кодеру, так и декодеру. Эти ответвляющиеся слова являются кодовыми символами, которые можно было бы ожидать на выходе кодера в результате каждого перехода между состояниями. Пометки на ветвях *решетки декодера* накапливаются декодером *в процессе*. Другими словами, когда получен кодовый символ, каждая ветвь решетки декодера помечена метрикой подобия (расстоянием Хэмминга) между полученным кодовым символом и каждым ответвляющимся словом за этот временной интервал. Из полученной последовательности $\mathbf{Z}$, показанной на рис. 7.10, можно видеть, что кодовые символы, полученные в (следующий) момент времени t_1, — это 11. Чтобы пометить ветви декодера подходящей метрикой расстояния Хэмминга в (прошедший) момент времени t_1, рассмотрим решетку кодера на рис. 7.7. Видим, что переход между состояниями $00 \rightarrow 00$ порождает на выходе ответвляющееся слово 00. Однако получено 11. Следовательно, на решетке декодера помечаем переход между состояниями $00 \rightarrow 00$ расстоянием Хэмминга между ними, а именно 2. Глядя вновь на решетку кодера, видим, что переход между состояниями $00 \rightarrow 10$ порождает на выходе ответвляющееся слово 11, точно соответствующее полученному в момент t_1 кодовому символу. Следовательно, переход на решетке декодера между состояниями $00 \rightarrow 10$ помечаем расстоянием Хэмминга 0. В итоге, метрика входящих в решетку декодера ветвей описывает разницу (расстояние) между тем, что было получено, и тем, что "могло бы быть" получено, имея ответвленные слова, связанные с теми ветвями, с которых они были переданы. По сути, эти метрики описывают величину, подобную корреляциям между полученным ответвляющимся словом и каждым из кандидатов на роль ответвляющегося слова. Таким же образом продолжаем помечать ветви решетки декодера по мере получения символов в каждый момент времени t_i. В алгоритме декодирования эти метрики расстояния Хэмминга используются для нахождения *наиболее вероятного* (с минимальным расстоянием) пути через решетку.

Смысл декодирования Витерби заключается в следующем. Если любые два пути сливаются в одном состоянии, то при поиске оптимального пути один из них всегда можно исключить. Например, на рис. 7.11 показано два пути, сливающихся в момент времени t_5 в состоянии 00.

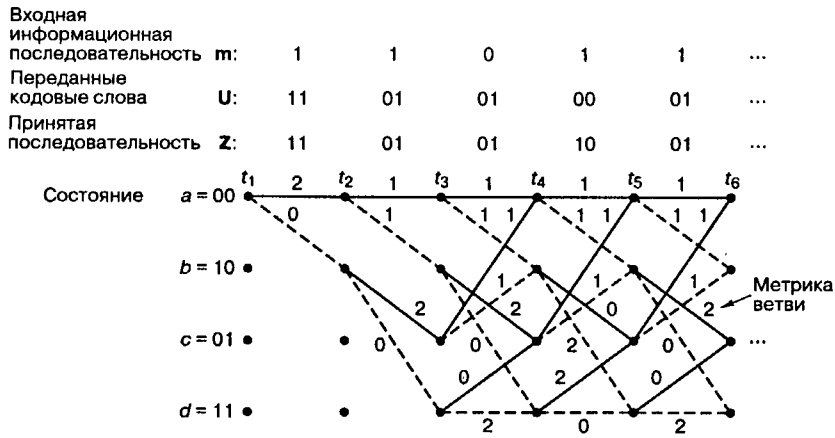


Рис. 7.10. Решетчатая диаграмма декодера (степень кодирования 1/2, $K = 3$)

Давайте определим *суммарную метрику пути по Хэммингу* для данного пути в момент времени t_i как сумму метрик расстояний Хэмминга ветвей, по которым проходит путь до момента t_i. На рис. 7.11 верхний путь имеет метрику 4, нижний — метрику 1. Верхний путь нельзя выделить как оптимальный, поскольку нижний путь, входящий в то же состояние, имеет меньшую метрику. Это наблюдение поддерживается Марковской природой состояний кодера. Настоящее состояние завершает историю кодера в том смысле, что предыдущие состояния не могут повлиять на будущие состояния или будущие ветви на выходе.

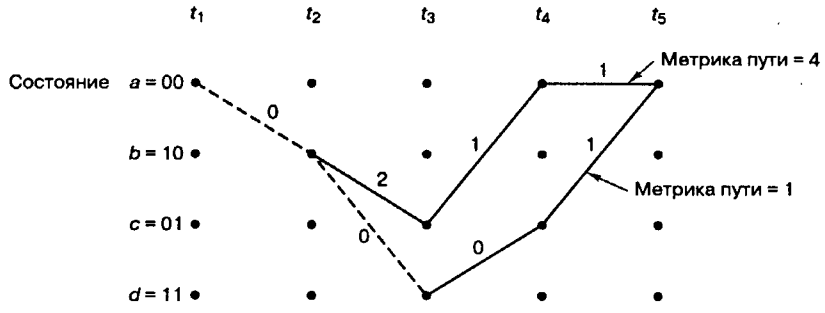


Рис. 7.11. Метрики пути для двух сливающихся путей

В каждый момент времени t_i в решетке существует 2^{K-1} состояний, где K — это длина кодового ограничения, и в каждое состояние *может войти два пути.* Декодирование Витерби состоит в вычислении метрики двух путей, входящих в каждое состояние, и *исключении одного из них.* Такие вычисления проводятся для каждого из 2^{K-1} состояний или узлов в момент времени t_i; затем декодер переходит к моменту времени t_{i+1}, и процесс повторяется. В данный момент времени метрика выжившего пути для каждого состояния обозначается как *метрика* для этого состояния в этот момент времени. Первые несколько шагов в нашем примере декодирования будут следующими (рис. 7.12). Предположим, что последовательность входящих данных **m**, кодовое слово **U** и полученная последовательность **Z** аналогичны показанным на рис. 7.10. Допустим, что декодер знает верное ис-

ходное состояние решетки. (Это предположение не является необходимым, однако упрощает объяснения.) В момент времени t_1 получены кодовые символы 11. Из состояния 00 можно перейти только в состояние 00 или 10, как показано на рис. 7.12, *а*. Переход между состояниями 00 → 10 имеет метрику ветви 0; переход между состояниями 00 → 00 — метрику ветви 2. В момент времени t_2 из каждого состояния также может выходить только две ветви, как показано на рис. 7.12, *б*. Суммарная метрика этих ветвей обозначена как метрика состояний Γ_a, Γ_b, Γ_c и Γ_d, соответствующих конечным состояниям. В момент времени t_3 на рис. 7.12, *в* опять есть две ветви, выходящие из каждого состояния. В результате имеется два пути, входящих в каждое состояние, в момент времени t_4. Один из путей, входящих в каждое состояние, может быть исключен, а точнее — это путь, имеющий большую суммарную метрику пути. Если бы метрики двух входящих путей имели одинаковое значение, то путь, который будет исключаться, выбирался бы произвольно. Выживший путь в каждом состоянии показан на рис. 7.12, *г*. В этой точке процесса декодирования имеется только один выживший путь, который называется *полной ветвью*, между моментами времени t_1 и t_2. Следовательно, декодер теперь может решить, что между моментами t_1 и t_2 произошел переход 00 → 10. Поскольку переход вызывается единичным входным битом, на выходе декодера первым битом будет единица. Здесь легко можно проследить процесс декодирования выживших ветвей, поскольку ветви решетки показаны пунктирными линиями для входных нулей и сплошной линией для входных единиц. Заметим, что первый бит не декодируется, пока вычисление метрики пути не пройдет далее вглубь решетки. Для обычного декодера такая задержка декодирования может оказаться раз в пять больше длины кодового ограничения в битах.

На каждом следующем шаге процесса декодирования всегда будет два пути для каждого состояния; после сравнения метрик путей один из них будет исключен. Этот шаг в процессе декодирования показан на рис. 7.12, *д*. В момент t_5 снова имеется по два входных пути для каждого состояния, и один путь из каждой пары подлежит исключению. Выжившие пути на момент t_5 показаны на рис. 7.12, *е*. Заметим, что в нашем примере мы еще не можем принять решения относительно второго входного информационного бита, поскольку еще остается два пути, исходящих в момент t_2 из состояния в узле 10. В момент времени t_6 на рис. 7.12, *ж* снова можем видеть структуру сливающихся путей, а на рис. 7.12, *з* — выжившие пути на момент t_6. Здесь же, на рис. 7.12, *з*, на выходе декодера в качестве второго декодированного бита показана единица как итог единственного оставшегося пути между точками t_2 и t_3. Аналогичным образом декодер продолжает углубляться в решетку и принимать решения, касающиеся информационных битов, устраняя все пути, кроме одного.

Отсекание (сходящихся путей) в решетке гарантирует, что у нас никогда не будет путей больше, чем состояний. В этом примере можно проверить, что после каждого отсекания (рис. 7.12, *б*–*д*) остается только 4 пути. Сравните это с попыткой применить "грубую силу" (без привлечения алгоритма Витерби) при использовании для получения последовательности принципа максимального правдоподобия. В этом случае число возможных путей (соответствующее возможным вариантам последовательности) является степенной функцией длины последовательности. Для двоичной последовательности кодовых слов с длиной ответвленных слов L имеется 2^L возможные последовательности.

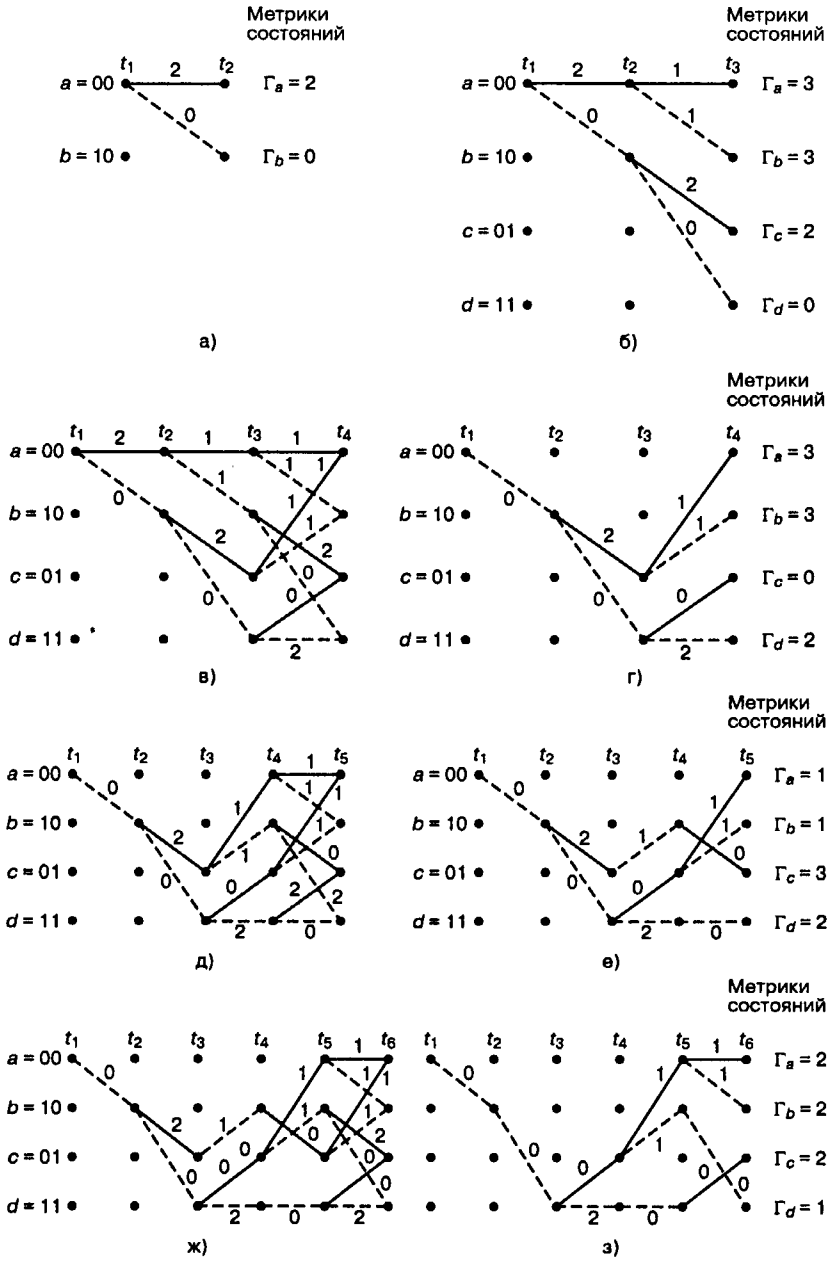


Рис. 7.12. Выбор выживших путей: а) выжившие на момент t_2; б) выжившие на момент t_3; в) сравнение метрик в момент t_4; г) выжившие на момент t_4; д) сравнение метрик в момент t_5; е) выжившие на момент t_5; ж) сравнение метрик в момент t_6; з) выжившие на момент t_6

7.3.5. Реализация декодера

В контексте решетчатой диаграммы, показанной на рис. 7.10, переходы за один промежуток времени можно сгруппировать в $2^{\nu-1}$ непересекающиеся ячейки; каждая ячейка будет изображать четыре возможных перехода, причем $\nu = K - 1$ называется *памятью кодера* (encoder memory). Если $K = 3$, то $\nu = 2$, и, следовательно, мы имеем $2^{\nu-1} = 2$ ячейки. Эти ячейки показаны на рис. 7.13, где буквы a, b, c и d обозначают состояния в момент t_i, а a', b', c' и d' — состояния в момент времени t_{i+1}. Для каждого перехода изображена *метрика ветви* δ_{xy}, индексы которой означают переход из состояния x в состояние y. Эти ячейки и соответствующие логические элементы, которые корректируют *метрики состояний* $\{\Gamma_x\}$, где x означает конкретное состояние, представляют основные составляющие элементы декодера.

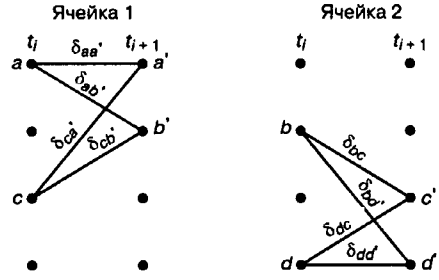


Рис. 7.13. Примеры ячеек декодера

7.3.5.1. Процедура сложения, сравнения и выбора

Вернемся к примеру двух ячеек с $K = 3$. На рис. 7.14 показан логический блок, соответствующий ячейке 1. Логическая схема осуществляет специальную операцию, которая называется *сложение, сравнение и выбор* (add-compare-select — ACS). Метрика состояния $\Gamma_{a'}$ вычисляется путем прибавления метрики предыдущего состояния a, Γ_a, к метрике ветви $\delta_{aa'}$ и метрики предыдущего состояния c, Γ_c, к метрике ветви $\delta_{ca'}$. Это даст в результате две метрики путей в качестве кандидатов для новой метрики состояния $\Gamma_{a'}$. Оба кандидата сравниваются в логическом блоке, показанном на рис. 7.14. Наиболее правдоподобная из двух метрик путей (с наименьшим расстоянием) запоминается как новая метрика состояния $\Gamma_{a'}$ для состояния a'. Также сохраняется новая история путей $\hat{m}_{a'}$ для состояния a, где $\hat{m}_{a'}$ — история пути информации для данного состояния, дополненная сведениями о выжившем пути.

На рис. 7.14 также показана логическая схема ACS для ячейки 1, которая дает новую метрику состояния $\Gamma_{b'}$ и новую историю состояния $\hat{m}_{b'}$. Операция ACS аналогичным образом осуществляется и для путей в других ячейках. Выход декодера составляют последние биты на путях с наименьшими метриками состояний.

7.3.5.2. Вид процедуры сложения, сравнения и выбора на решетке

Рассмотрим тот же пример, которым мы воспользовались в разделе 7.3.4 для описания декодирования на основе алгоритма Витерби. Последовательность сообщений имела вид $\mathbf{m} = 1\ 1\ 0\ 1\ 1$, последовательность кодовых слов — $\mathbf{U} = 11\ 01\ 01\ 00\ 01$, а принятая последовательность — $\mathbf{Z} = 11\ 01\ 01\ 10\ 01$.

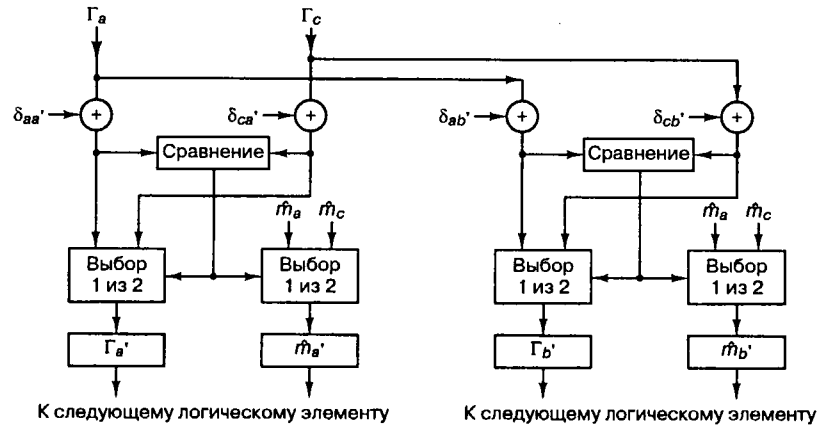


Рис. 7.14. Логический блок, предназначенный для осуществления операции сложения, сравнения и выбора

Решетчатая диаграмма декодирования, аналогичная показанной на рис. 7.10, изображена на рис. 7.15. Метрика ветви, которая описывает каждую ветвь, — это расстояние Хэмминга между принятым кодовым символом и соответствующим ответвленным словом из решетки кодера. Еще на решетке (рис. 7.15) показаны значения каждого состояния x в каждый момент t_2–t_6, метрика состояния которых обозначена Γ_x. Операция ACS выполняется после появления двух переходов, входящих в состояние, т.е. для момента t_4 и более поздних. Например, в момент времени t_4 значение метрики состояния для состояния a вычисляется суммированием метрики состояния $\Gamma_a = 3$ в момент t_3 и метрики ветви $\delta_{ad'} = 1$, что в итоге дает значение 4. В то же время к метрике состояния $\Gamma_c = 2$ в момент времени t_3 прибавляется метрика ветви $\delta_{cd'} = 1$, что даст значение 3. В ходе процедуры ACS происходит отбор наиболее правдоподобной метрики (с минимальным расстоянием), т.е. новой метрики состояния; поэтому для состояния a в момент t_4 новой метрикой состояния будет $\Gamma_{d'} = 3$. Отобранный путь изображен жирной линией, а путь, который был отброшен, показан светлой линией. На рис. 7.15 на решетке слева направо показаны все метрики состояний. Убедимся, что в любой момент времени значение каждой метрики состояния получается суммированием метрики состояния, соединенного с предыдущим состоянием вдоль отобранного пути (жирная линия), и метрики ветви, соединяющей эти состояния. В определенной точке решетки (после временного интервала, равного 4 или 5 длинам кодового ограничения) будут декодированы самые ранние биты. Чтобы показать это, посмотрим на рис. 7.15 в момент t_6. Видим, что значение метрики состояния, соответствующей минимальному расстоянию, равно 1. Отобранный путь можно проследить из состояния d обратно, к моменту t_1, и убедиться, что декодированное сообщение совпадает с исходным. Напомним, что пунктирные и сплошные линии соответствуют двоичным единице и нулю.

7.3.6. Память путей и синхронизация

Требования к памяти декодера, работающего согласно алгоритму Витерби, растут с увеличением длины кодового ограничения как степенная функция. Для кода со степенью кодирования $1/n$ после каждого шага декодирования декодер держит в памяти набор из 2^{K-1} путей.

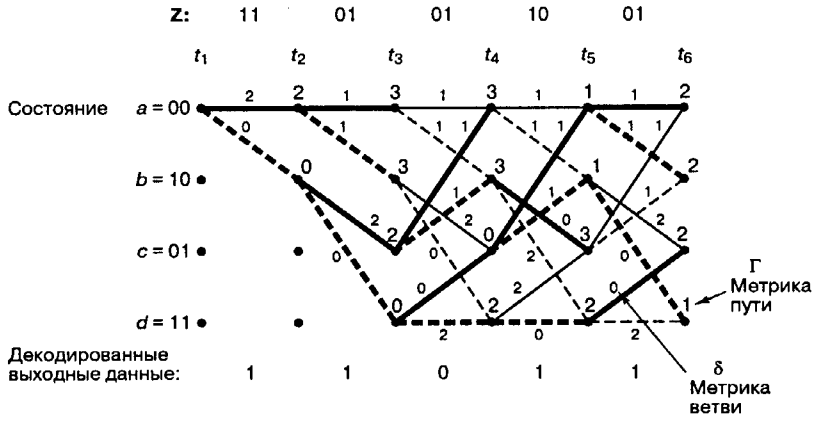


Рис. 7.15. Операция сложения, сравнения и выбора при декодировании по алгоритму Витерби

С высокой степенью вероятности при превышении существующей глубины декодирования эти пути не будут взаимно непересекающимися [12]. Все 2^{K-1} пути ведут к полной ветви, которая в конце концов разветвляется на разные состояния. Поэтому, если декодер сохраняет историю 2^{K-1} путей, самые первые биты на всех путях будут одинаковы. Следовательно, простой декодер имеет *фиксированный объем истории путей* и выдает самые ранние биты произвольного пути каждый раз, когда продвигается на один уровень вглубь решетки. Требуемый объем сохраняемых путей будет равен следующему [12].

$$u = h2^{K-1} \tag{7.10}$$

Здесь h — длина истории пути информационного бита на состояние. При уточнении, которое проводится для минимизации h, вместо самых ранних битов произвольных путей на выходе декодера используются самые ранние биты наиболее вероятных путей. Было показано [12], что значения h, равного 4 или 5 длинам кодового ограничения, достаточно, чтобы характеристики декодера были близки к оптимальным. Необходимый объем памяти u является основным ограничением при разработке декодеров, работающих согласно алгоритму Витерби. В серийно выпускаемых декодерах длина кодового ограничения равна величине порядка $K = 10$. Попытка повысить эффективность кодирования за счет увеличения длины кодового ограничения вызывает степенной рост требований к памяти (и сложности), как это следует из уравнения (7.10).

Синхронизация ответвляющихся слов — это процесс определения начала ответвляющегося слова в принятой последовательности. Такую синхронизацию можно осуществить, не прибавляя новую информацию к потоку передаваемых символов, поскольку можно видеть, что, пока принятые данные не синхронизированы, у них непомерно высокая частота появления ошибок. Следовательно, синхронизацию можно осуществить просто: нужно проводить сопутствующее наблюдение за уровнем частоты появления ошибок, т.е. нас должна интересовать частота, при которой увеличиваются метрики состояний, или частота, при которой сливаются выжившие пути на решетке. Параметр, за которым следят, сравнивается с пороговым значением, после чего соответствующим образом осуществляется синхронизация.

7.4. Свойства сверточных кодов

7.4.1. Пространственные характеристики сверточных кодов

Рассмотрим пространственные характеристики сверточных кодов в контексте простого кодера (рис. 7.3) и его решетчатой диаграммы (рис. 7.7). Мы хотим узнать расстояния между всеми возможными парами последовательностей кодовых слов. Как и в случае блочных кодов (см. раздел 6.5.2), нас интересует *минимальное расстояние* между всеми такими парами последовательностей кодовых слов в коде, поскольку минимальное расстояние связано с возможностями кода в коррекции ошибок. Поскольку сверточный код является групповым или *линейным* [6], можно без потери общности просто найти минимальное расстояние между последовательностью кодовых слов и нулевой последовательностью. Другими словами, для линейного кода данное контрольное сообщение окажется точно таким же "хорошим", как и любое другое. Так почему бы не взять то сообщение, которое легко проследить, а именно нулевую последовательность? Допустим, что на вход передана нулевая последовательность; следовательно, нас интересует такой путь, который начинается и заканчивается в состоянии 00 и не возвращается к состоянию 00 нигде внутри пути. Всякий раз, когда расстояние любых других путей, которые сливаются с состоянием $a = 00$ в момент t_i, окажется меньше расстояния нулевого пути, вплоть до момента t_l, будет появляться ошибка, вызывая в процессе декодирования отбрасывание нулевого пути. Иными словами, при нулевой передаче ошибка возникает всегда, когда *не выживает нулевой путь*. Следовательно, ошибка, о которой идет речь, связана с выживающим путем, который расходится, а затем снова сливается с нулевым путем. Может возникнуть вопрос, зачем нужно, чтобы пути сливались? Не будет ли для обнаружения ошибки достаточно лишь того, чтобы пути расходились? В принципе, достаточно, но если ошибка характеризуется *только* расхождением, то декодер, начиная с этой точки, будет выдавать вместо оставшегося сообщения сплошной "мусор". Мы хотим выразить возможности декодера через число обычно появляющихся ошибок, т.е. хотим узнать "самый легкий" для декодера способ сделать ошибку. Минимальное расстояние для такой ошибки можно найти, полностью изучив все пути из состояния 00 в состояние 00. Итак, давайте сначала заново начертим решетчатую диаграмму, как показано на рис. 7.16, и обозначим каждую ветвь не символом ответвляющегося слова, а ее расстоянием Хэмминга от нулевого кодового слова. Расстояние Хэмминга между двумя последовательностями разной длины можно получить путем их сравнивания, т.е. прибавив к началу более короткой последовательности нужное количество нулей. Рассмотрим все пути, которые расходятся из нулевого пути и затем в какой-то момент снова сливаются в произвольном узле. Из диаграммы на рис. 7.16 можно получить расстояние этих путей до нулевого пути. Итак, на расстоянии 5 от нулевого пути имеется один путь; этот путь отходит от нулевого в момент t_1 и сливается с ним в момент t_4. Точно так же имеется два пути с расстоянием 6, один отходит в момент t_1 и сливается в момент t_5, а другой отходит в момент t_1 и сливается в момент t_6 и т.д. Также можно видеть (по пунктирным и сплошным линиям на диаграмме), что входными битами для расстояния 5 будут 1 0 0; от нулевой входной последовательности эта последовательность отличается только одним битом. Точно так же входные биты для путей с расстоянием 6 будут 1 1 0 0 и 1 0 1 0 0; каждая из этих последовательностей отличается от нулевого пути в двух местах. Минимальная длина пути из числа расходящихся, а затем сливающихся путей на-

зывается *минимальным просветом* (minimum free distance), или просто *просветом* (free distance). Его можно видеть на рис. 7.16, где он показан жирной линией. Для оценки возможностей кода коррекции ошибок, мы повторно приведем уравнение (6.44) с заменой минимального расстояния d_{min} на просвет d_f.

$$t = \left\lfloor \frac{d_f}{2} \right\rfloor \tag{7.11}$$

Здесь $\lfloor x \rfloor$ означает наибольшее целое, не большее x. Положив $d_f = 5$, можно видеть, что код, описываемый кодером на рис. 7.3, может исправить две любые ошибки канала (см. раздел 7.4.1.1).

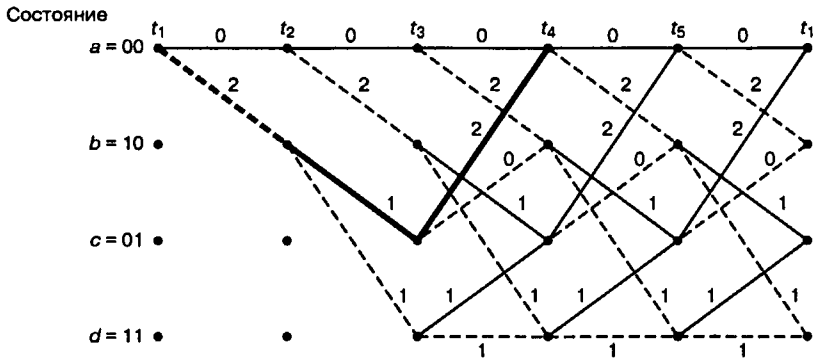


Рис. 7.16. Решетчатая диаграмма с обозначенными расстояниями от нулевого пути

Решетчатая диаграмма представляет собой "правила игры". Она является как бы символическим описанием всех возможных переходов и соответствующих начальных и конечных состояний, ассоциируемых с конкретным конечным автоматом. Эта диаграмма позволяет взглянуть глубже на выгоды (эффективность кодирования), которые дает применение кодирования с коррекцией ошибок. Взглянем на рис. 7.16 и на возможные ошибочные расхождения и слияния путей. Из рисунка видно, что декодер не может сделать ошибку *произвольным образом*. Ошибочный путь должен следовать одному из возможных переходов. Решетка позволяет нам определить все такие доступные пути. Получив по этому пути кодированные данные, мы можем наложить ограничения на переданный сигнал. Если декодер знает об этих ограничениях, то это позволяет ему более просто (используя меньшее E_b/N_0) удовлетворять требованиям надежной безошибочной работы.

Хотя на рис. 7.16 представлен способ прямого вычисления просвета, для него можно получить более строгое аналитическое выражение, воспользовавшись для этого диаграммой состояний, изображенной на рис. 7.5. Для начала обозначим ветви диаграммы состояний как $D^0 = 1$, D^1 или D^2, как это показано на рис. 7.17, где показатель D означает расстояние Хэмминга между ответвленным словом этой ветви и нулевой ветвью. Петлю в узле a можно убрать, поскольку она не дает никакого вклада в пространственные характеристики последовательности кодовых слов относительно нулевой последовательности. Более того, узел a можно разбить на два узла (обозначим их a и e), один из них представляет вход, а другой — выход диаграммы состояний. Все

пути, начинающиеся из состояния $a = 00$ и заканчивающиеся в $e = 00$, можно проследить на модифицированной диаграмме состояний, показанной на рис. 7.17. Передаточную функцию пути *a b c e (который начинается и заканчивается в состоянии 00)* можно рассчитать через неопределенный "заполнитель" D как $D^2DD^2 = D^5$. Степень D — общее число единиц на пути, а значит, расстояние Хэмминга до нулевого пути. Точно так же пути *a b d c e* и *a b c b c e* имеют передаточную функцию D^6 и, соответственно, расстояние Хэмминга, равное 6, до нулевого пути. Теперь уравнения состояния можем записать следующим образом.

$$\begin{aligned} X_b &= D^2X_a + X_c \\ X_c &= DX_b + DX_d \\ X_d &= DX_b + DX_d \\ X_e &= D^2X_c \end{aligned} \tag{7.12}$$

Здесь X_a, ..., X_e являются фиктивными переменными неполных путей между промежуточными узлами. *Передаточную функцию*, $T(D)$, которую иногда называют *производящей функцией кода*, можно записать как $T(D) = X_e/X_a$. Решение уравнений состояния (7.12) имеет следующий вид [15, 16].

$$\begin{aligned} T(D) &= \frac{D^5}{1-2D} = \\ &= D^5 + 2D^6 + 4D^7 + \ldots + 2^l D^{l+5} + \ldots \end{aligned} \tag{7.13}$$

Передаточная функция этого кода показывает, что имеется один путь с расстоянием 5 до нулевого вектора, два пути — с расстоянием 6, четыре — с расстоянием 7. Вообще, существуют 2^l пути с расстоянием $l + 5$ до нулевого вектора, причем $l = 0, 1, 2, \ldots$. Просвет d_f кода является весовым коэффициентом Хэмминга члена наименьшего порядка разложения $T(D)$. В данном случае $d_f = 5$. Для оценки пространственных характеристик при большой длине кодового ограничения передаточную функцию $T(D)$ использовать нельзя, поскольку сложность $T(D)$ растет с увеличением длины кодового ограничения как степенная функция.

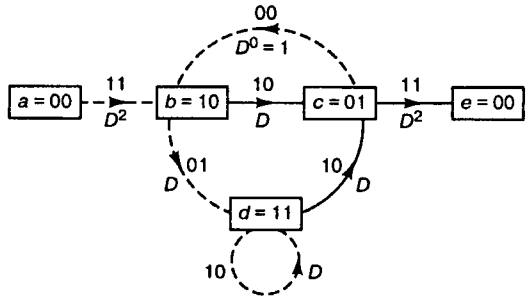


Рис. 7.17. Диаграмма состояний с обозначенными расстояниями до нулевого пути

С помощью передаточной функции можно получить более подробную информацию, чем при использовании лишь расстояния между различными путями. В каждую ветвь диаграммы состояний введем множитель L так, чтобы показатель L мог служить счетчиком ветвей в любом пути из состояния $a = 00$ в состояние $e = 00$. Более того, мы

можем ввести множитель N во все ветви переходов, порожденных входной двоичной единицей. Таким образом, после прохождения ветви суммарный множитель N возрастает на единицу, только если этот переход ветви вызван входной битовой единицей. Для сверточного кода, описанного на рис. 7.3, на перестроенной диаграмме состояний (рис. 7.18) показаны дополнительные множители L и N. Уравнения (7.12) теперь можно переписать следующим образом.

$$\begin{aligned} X_b &= D^2LNX_a + LNX_c \\ X_c &= DLX_b + DLX_d \\ X_d &= DLNX_b + DLNX_d \\ X_e &= D^2LX_c \end{aligned} \tag{7.14}$$

Передаточная функция такой доработанной диаграммы состояний будет следующей.

$$\begin{aligned} T(D, L, N) &= \frac{D^5L^3N}{1 - DL(1+L)N} = \\ &= D^5L^3N + D^6L^4(1+L)N^2 + D^7L^5(1+L^2)N^3 + \\ &\quad + \ldots + D^{l+5}L^{l+3}N^{l+1} + \ldots \end{aligned} \tag{7.15}$$

Таким образом, мы можем проверить некоторые свойства путей, показанные на рис. 7.16. Существует один путь с расстоянием 5 и длиной 3, который отличается от нулевого пути одним входным битом. Имеется два пути с расстоянием 6, один из них имеет длину 4, другой — длину 5, и оба отличаются от нулевого пути двумя входными битами. Также есть пути с расстоянием 7, из которых один имеет длину 5, два — длину 6 и один — длину 7; все четыре пути соответствуют входной последовательности, которая отличается от нулевого пути тремя входными битами. Следовательно, если нулевой путь является правильным и шум приводит к тому, что мы выбираем один из неправильных путей, то в итоге получится три битовые ошибки.

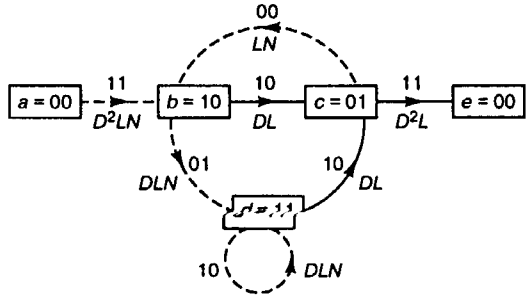


Рис. 7.18. Диаграмма состояний с обозначением расстояния, длины и числа входных единиц

7.4.1.1. Возможности сверточного кода в коррекции ошибок

В главе 6 при изучении блочных кодов говорилось, что способность кода к коррекции ошибок, t, представляет собой количество ошибочных кодовых символов, которые можно исправить в каждом блоке кода путем декодирования по методу максимального правдоподобия. В то же время при декодировании сверточных кодов способность кода к коррекции ошибок нельзя сформулировать так лаконично. Из уравнения (7.11) можно сказать, что при декодировании по принципу максимального правдоподобия код способен исправить t ошибок в пределах нескольких длин кодо-

вого ограничения, причем "несколько" — это где-то от 3 до 5. Точное значение длины зависит от распределения ошибок. Для конкретного кода и ошибочной комбинации длину можно ограничить с использованием методов передаточной функции. Такое ограничение будет описано позднее.

7.4.2. Систематические и несистематические сверточные коды

Систематический сверточный код — это код, в котором входной k-кортеж фигурирует как часть выходного n-кортежа ответвляющегося слова, соответствующего этому k-кортежу. На рис. 7.19 показан двоичный систематический кодер со степенью кодирования 1/2 и $K = 3$. Для линейных блочных кодов любой несистематический код можно преобразовать в систематический с такими же пространственными характеристиками блоков. При использовании сверточных кодов это не так. Это означает, что сверточные коды сильно зависят от *просвета*; при построении сверточного кода в систематической форме при данной длине кодового ограничения и степени кодирования максимально возможное значение просвета *снижается*.

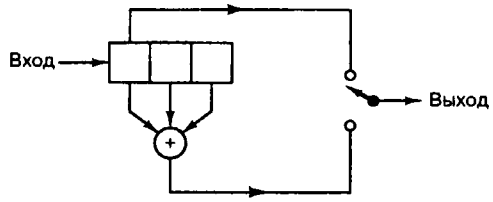


Рис. 7.19. Систематический сверточный кодер (степень кодирования 1/2, $K = 3$)

В табл. 7.1 показан максимальный просвет при степени кодирования 1/2 для систематического и несистематического кодов с K от 2 до 8. При большой длине кодового ограничения результаты отличаются еще сильнее [17].

Таблица 7.1. Сравнение систематического и несистематического просветов, степень кодирования 1/2

Длина кодового ограничения	Просвет систематического кода	Просвет несистематического кода
2	3	3
3	4	5
4	4	6
5	5	7
6	6	8
7	6	10
8	7	10

Источник: A. J. Viterbi and J. K. Omura. *Principles of Digital Communication and Coding*, McGraw-Hill Book Company, New-York, 1979, p. 251.

7.4.3. Накопление катастрофических ошибок в сверточных кодах

Катастрофическая ошибка возникает, когда конечное число ошибок в кодовых символах вызывает бесконечное число битовых ошибок в декодированных данных. Мэсси

(Massey) и Сейн (Sain) указали необходимые и достаточные условия для сверточного кода, при которых возможно накопление катастрофических ошибок. Условием накопления катастрофических ошибок для кода со степенью кодирования 1/2, реализованного на полиномиальных генераторах, описанных в разделе 7.2.1, будет наличие у генераторов общего полиномиального делителя (степени не менее единицы). Например, на рис. 7.20, *а* показан кодер с $K = 3$, степенью кодирования 1/2, со старшим полиномом $\mathbf{g}_1(X)$ и младшим $\mathbf{g}_2(X)$.

$$\begin{aligned} \mathbf{g}_1(X) &= 1 + X \\ \mathbf{g}_2(X) &= 1 + X^2 \end{aligned} \tag{7.16}$$

Генераторы $\mathbf{g}_1(X)$ и $\mathbf{g}_2(X)$ имеют общий полиномиальный делитель $1 + X$, поскольку

$$1 + X^2 = (1 + X)(1 + X).$$

Следовательно, в кодере, показанном на рис. 7.20, *а*, может происходить *накопление катастрофической ошибки.*

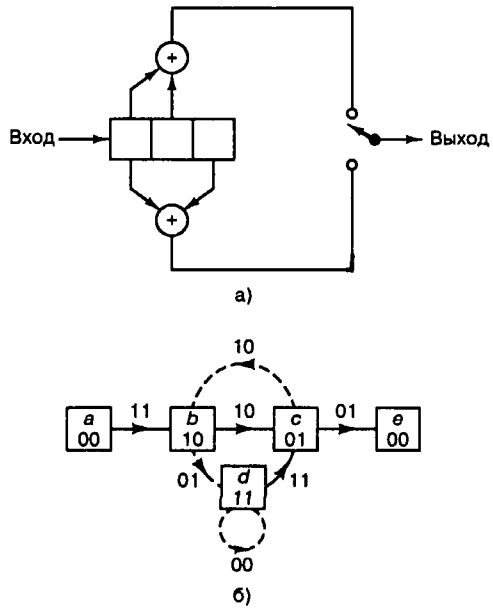


Рис. 7.20. Кодер, в котором возможно накопление катастрофической ошибки: а) кодер; б) диаграмма состояний

Если говорить о диаграмме состояний кода произвольной степени кодирования, то катастрофическая ошибка может появиться тогда и только тогда, когда любая петля пути на диаграмме имеет нулевой весовой коэффициент (нулевое расстояние до нулевого пути). Чтобы проиллюстрировать это, рассмотрим пример, приведенный на рис. 7.20. На диаграмме (рис. 7.20, *б*) узел состояния $a = 00$ разбит на два узла, *а* и *е*, как и ранее. Допустим, что нулевой путь является правильным, тогда неправильный путь *a b d d ... d c e* имеет точно 6 единиц, независимо от того, сколько раз мы обойдем вокруг петли в узле *d*. Поэтому, например, для канала BSC к выбору этого неправильного пути могут привести три канальные ошибки. На таком пути может появить-

ся сколь угодно большое число ошибок (две плюс количество раз обхода петли). Для кодов со степенью кодирования $1/n$ можно видеть, что если каждый сумматор в кодере имеет четное количество соединений, петли, которые соответствуют информационным состояниям со всеми единицами, будут иметь нулевой вес, и, следовательно, *код будет катастрофическим.*

Единственное преимущество описанного ранее систематического кода заключается в том, что он никогда не будет катастрофическим, поскольку каждая петля должна содержать по крайней мере одну ветвь, порождаемую ненулевым входным битом; следовательно, каждая петля должна содержать ненулевой кодовый символ. Впрочем, можно показать [19], что только небольшая часть несистематических кодов (исключая тот, в котором все сумматоры имеют четное количество соединений) является катастрофической.

7.4.4. Границы рабочих характеристик сверточных кодов

Можно показать [8], что вероятность битовой ошибки P_B в бинарном сверточном коде, использующем при декодировании жесткую схему принятия решений, может быть ограничена сверху следующим образом.

$$P_B \le \left.\frac{dT(D, N)}{dN}\right|_{N=1,\, D=2\sqrt{p(1-p)}} \tag{7.17}$$

где p — вероятность ошибки в канальном символе. Для примера, приведенного на рис. 7.3, $T(D, N)$ получено из $T(D, L, N)$ путем задания $L = 1$ в уравнении (7.15).

$$T(D, N) = \frac{D^5 N}{1 - 2DN} \tag{7.18}$$

и

$$\left.\frac{dT(D, N)}{dN}\right|_{N=1} = \frac{D^5}{(1-2D)^2} \tag{7.19}$$

Объединяя уравнения (7.17) и (7.19), можем записать следующее.

$$P_B \le \frac{\{2[p(1-p)]^{1/2}\}^5}{\{1-4[p(1-p)]^{1/2}\}^2} \tag{7.20}$$

Можно показать, что при когерентной модуляции BPSK в канале с аддитивным белым гауссовым шумом (additive white Gaussian noise — AWGN) вероятность битовой ошибки ограничивается следующей величиной.

$$P_B \le Q\left(\sqrt{2d_f \frac{E_c}{N_0}}\right) \exp\left(d_f \frac{E_c}{N_0}\right) \left.\frac{dT(D, N)}{dN}\right|_{N=1,\, D=\exp(-E_c/N_0)} \tag{7.21}$$

где

$E_c/N_0 = rE_b/N_0$,

E_b/N_0 — отношение энергии информационного бита к спектральной плотности мощности шума,

E_c/N_0 — отношение энергии канального символа к спектральной плотности мощности шума,

$r = k/n$ — степень кодирования,

а $Q(x)$ определяется уравнениями (3.43) и (3.44) и приведено в табл. Б.1. Следовательно, для кода со степенью кодирования 1/2 и просветом $d_f = 5$, при использовании когерентной схемы BPSK и жесткой схемы принятия решений при декодировании, можем записать следующее.

$$\begin{aligned} P_B &\leq Q\left(\sqrt{\frac{5E_b}{N_0}}\right)\exp\left(\frac{5E_b}{2N_0}\right)\frac{\exp(-5E_b/2N_0)}{[1-2\exp\{-E_b/2N_0\}]^2} \leq \\ &\leq \frac{Q(\sqrt{5E_b/N_0})}{[1-2\exp(-E_b/2N_0)]^2} \end{aligned} \tag{7.22}$$

7.4.5. Эффективность кодирования

Эффективность кодирования, представленная уравнением (6.19), определяется как уменьшение (обычно выраженное в децибелах) отношения E_b/N_0, требуемого для достижения определенной вероятности появления ошибок в кодированной системе, по сравнению с некодированной системой с той же модуляцией и характеристиками канала. В табл. 7.2 перечислены верхние границы эффективности кодирования. Они сравниваются с некодированным сигналом с когерентной модуляцией BPSK для нескольких значений минимальных просветов сверточного кода. Длина кодового ограничения в гауссовом канале с жесткой схемой принятия решений при декодировании изменяется от 3 до 9. В таблице отражен тот факт, что даже при использовании простого сверточного кода можно достичь значительной эффективности кодирования. Реальная эффективность кодирования будет изменяться в зависимости от требуемой вероятности появления битовых ошибок [20].

Таблица 7.2. Верхние границы эффективности кодирования для некоторых сверточных кодов

Коды со степенью кодирования *1/2*			Коды со степенью кодирования *1/2*		
K	d_f	Верхняя граница (дБ)	*K*	d_f	Верхняя граница (дБ)
3	5	3,97	3	8	4,26
4	6	4,76	4	10	5,23
5	7	5,43	5	12	6,02
6	8	6,00	6	13	6,37
7	10	6,99	7	15	6,99
8	10	6,99	8	16	7,27
9	12	7,78	9	18	7,78

Источник: V. K. Bhargava, D. Haccoun, R. Matyas and P. Nuspl. *Digital Communications by Satellite.* John Wiley & Sons, Inc., New York, 1981.

В табл. 7.3 приводятся оценки эффективности кодов, сравниваемые с некодированным сигналом с когерентной модуляцией BPSK, реализованной аппаратным путем или путем моделирования на компьютере, в гауссовом канале с мягкой схемой принятия решений при декодировании [21]. Некодированное значение E_b/N_0 дано в крайнем левом столбце. Из табл. 7.3 можно видеть, что эффективность кодирования возрастает при уменьшении вероятности появления битовой ошибки. Однако эффективность кодирования не может возрастать бесконечно. Как показано в таблице, она имеет верхнюю границу. Эту границу (в децибелах) можно выразить следующим образом.

$$\text{эффективность кодирования} \leq 10 \lg(rd_f) \tag{7.23}$$

Здесь r — степень кодирования, а d_f — просвет. При изучении табл. 7.3 обнаруживается также, что (при $p_B = 10^{-7}$) для кодов со степенью кодирования 1/2 и 2/3 более слабые коды имеют тенденцию находиться ближе к верхней границе, чем более мощные коды.

Таблица 7.3. Основные значения эффективности кодирования (в дБ) при использовании мягкой схемы принятия решений в ходе декодирования по алгоритму Витерби

Некодированное E_b/N_0	Степень кодирования		1/3		1/2			2/3		3/4	
(дБ)	P_B	K	7	8	5	6	7	6	8	6	9
6,8	10^{-3}		4,2	4,4	3,3	3,5	3,8	2,9	3,1	2,6	2,6
9,6	10^{-5}		5,7	5,9	4,3	4,6	5,1	4,2	4,6	3,6	4,2
11,3	10^{-7}		6,2	6,5	4,9	5,3	5,8	4,7	5,2	3,9	4,8
Верхняя граница			7,0	7,3	5,4	6,0	7,0	5,2	6,7	4,8	5,7

Источник: I. M. Jacobs. *Practical Applications of Coding.* IEEE Trans. Inf. Theory, vol. IT20, May 1974, pp. 305–310.

Как правило, декодирование по алгоритму Витерби используется в двоичном входном канале с жестким или мягким 3-битовым квантованным выходом. Длина кодового ограничения варьируется от 3 до 9, причем степень кодирования кода редко оказывается меньше 1/3, и память путей составляет несколько длин кодового ограничения [12]. Памятью путей называется глубина входных битов, которая сохраняется в декодере. После рассмотрения в разделе 7.3.4 декодирования по алгоритму Витерби может возникнуть вопрос об ограничении объема памяти путей. Из этого примера может показаться, что декодирование ответвленного слова в любом узле может происходить сразу, как только останется один выживший путь в этом узле. Это действительно так; хотя для создания реального декодера таким способом потребуется большое количество постоянных проверок после декодирования ответвленного слова. На практике вместо всего этого *обеспечивается фиксированная задержка*, после которой ответвляющееся слово декодируется. Было показано [12, 22], что информации о происхождении состояния с наименьшей метрикой состояния (с использованием фиксированного объема путей, порядка 4 или 5 длин кодового ограничения) достаточно для получения характеристик декодера, которые для гауссова канала и канала BSC на величину порядка 0,1 дБ меньше характеристик оптимального канала. На рис. 7.21 показаны характерные результаты моделирования достоверности передачи при декодировании по алгоритму Витерби с жесткой схемой квантования [12]. Заметьте, что каждое увеличение длины кодового ограничения приводит к улучшению требуемого значения E_b/N_0 на величину, равную приблизительно 0,5 дБ, при $P_B = 10^{-5}$.

7.4.6. Наиболее известные сверточные коды

Векторы связи или полиномиальные генераторы сверточного кода обычно выбираются исходя из свойств просветов кода. Главным критерием при выборе кода является требование, чтоб код не допускал катастрофического накопления ошибок и имел максимальный просвет при данной степени кодирования и длине кодового ограничения.

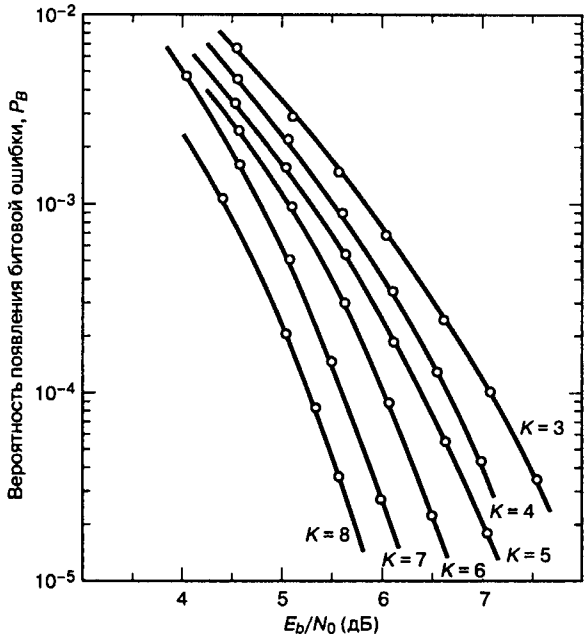


Рис. 7.21. Зависимость вероятности появления битовой ошибки от E_b/N_0 при степени кодирования кодов 1/2; используется когерентная модуляция BPSK в канале BSC, декодирование согласно алгоритму Витерби и 32-битовая память путей. (Перепечатано с разрешения авторов из J. A. Heller and I. M. Jacobs. "Viterbi Decoding for Satellite and Space Communication". IEEE Trans. Commun. Technol., vol. COM19, n. 5, October, 1971, Fig. 7, p. 84 © 1971, IEEE.)

Затем при данном просвете d_f минимизируется число путей или число ошибочных битов данных, которые представляют путь. Процедуру выбора можно усовершенствовать, рассматривая количество путей или ошибочных битов при $d_f + 1$, $d_f + 2$ и т.д., пока не останется только один код или класс кодов. Список наиболее известных кодов со степенью кодирования 1/2 при K, равном от 3 до 9, и со степенью кодирования 1/3 при K, равном от 3 до 8, соответствующих этому критерию, был составлен Оденуальдером (Odenwalder) [3, 23] и приводится в табл. 7.4. Векторы связи в этой таблице представляют наличие или отсутствие (1 или 0) соединения между соответствующими регистрами сверточного кодера, причем крайний левый элемент соответствует крайнему левому разряду регистра кодера. Интересно, что эти соединения можно обратить (заменить в указанной выше схеме крайние левые на крайние правые). При декодировании по алгоритму Витерби обратные соединения приведут к кодам с точно такими же пространственными характеристиками, а значит, и с такими же рабочими характеристиками, как показаны в табл. 7.4.

Таблица 7.4. Оптимальные коды с малой длиной кодового ограничения (степень кодирования 1/2 и 1/3)

Степень кодирования	Длина кодового ограничения	Просвет	Вектор кода
1/2	3	5	111 101

Окончание табл. 7.4

Степень кодирования	Длина кодового ограничения	Просвет	Вектор кода
1/2	4	6	1111
			1011
1/2	5	7	10111
			11001
1/2	6	8	101111
			110101
1/2	7	10	1001111
			1101101
1/2	8	10	10011111
			11100101
1/2	9	12	110101111
			100011101
			111
1/3	3	8	111
			101
			1111
1/3	4	10	1011
			1101
			11111
1/3	5	12	11011
			10101
			10111
1/3	6	13	110101
			111001
			1001111
1/3	7	15	1010111
			1101101
			11101111
1/3	8	16	10011011
			10101001

Источник: J. P. Odenwalder. *Error Control Coding Handbook.* Linkabit Corp., San Diego, Calif., July, 15, 1976.

7.4.7. Компромиссы сверточного кодирования

7.4.7.1. Производительность при когерентной передаче сигналов с модуляцией PSK

Возможности схемы кодирования в коррекции ошибок возрастают при увеличении числа канальных символов n, приходящихся на число информационных бит k, или при снижении степени кодирования k/n. В то же время при этом увеличивается ширина полосы пропускания канала и сложность декодера. Выгода низких степеней кодирования при использовании сверточного кода совместно с когерентной модуляцией PSK проявляется в снижении требуемого значения E_b/N_0 (для широкого диапазона

степеней кодирования), что позволяет при заданном значении мощности осуществить передачу на более высоких скоростях или снизить мощность при заданной скорости передачи информации. Компьютерное моделирование показало [16, 22], что при фиксированной длине кодового ограничения снижение степени кодирования с 1/2 до 1/3 в итоге приводит к уменьшению требуемого значения E_b/N_0 примерно на 0,4 дБ (сложность декодера при этом возрастает примерно на 17%). Для меньших значений степени кодирования улучшение рабочих характеристик по отношению к росту сложности декодирования быстро убывает [22]. В конечном счете, существует точка, по достижении которой дальнейшее снижение степени кодирования приводит к падению эффективности кодирования (см. раздел 9.7.7.2).

7.4.7.2. Производительность при некогерентной ортогональной передаче сигналов

В отличие от модуляции PSK, при некогерентной ортогональной передаче сигналов существует оптимальное значение степени кодирования, приблизительно равное 1/2. Надежность передачи при степени кодирования 1/3, 2/3 и 3/4 хуже, чем при степени кодирования 1/2. При фиксированной длине кодового ограничения и степени кодирования 1/3, 2/3 или 3/4 качество кодирования, как правило, падает на 0,25, 0,5 и 0,3 дБ, соответственно, по сравнению с достоверностью передачи при степени кодирования 1/2 [16].

7.4.8. Мягкое декодирование по алгоритму Витерби

Для двоичной кодовой системы со степенью кодирования 1/2, демодулятор подаст на декодер два кодовых символа за раз. Для жесткого (двухуровневого) декодирования каждую пару принятых кодовых символов можно изобразить на плоскости в виде одного из углов квадрата, как показано на рис. 7.22, *а*. Углы помечены двоичными числами (0, 0), (0, 1), (1, 0) и (1, 1), представляющими четыре возможных значения, которые могут принимать два кодовых символа в жесткой схеме принятия решений. Аналогично для 8-уровневого мягкого декодирования каждую пару кодовых символов можно отобразить на плоскости в виде равностороннего прямоугольника размером 8×8, состоящего из 64 точек, как показано на рис. 7.22, *б*. В этом случае демодулятор больше не выдает жестких решений; он выдает квантованные сигналы с шумом (мягкая схема принятия решений).

Основное различие между мягким и жестким декодированием по алгоритму Витерби состоит в том, что в мягкой схеме не используется метрика расстояния Хэмминга, поскольку она имеет ограниченное разрешение. Метрика расстояний, которая имеет нужное разрешение, называется евклидовым кодовым расстоянием, поэтому далее, чтобы облегчить ее применение, соответствующим образом преобразуем двоичные числа из единиц и нулей в восьмеричные числа от 0 до 7. Это можно увидеть на рис. 7.22, *в*, где соответствующим образом обозначены углы квадрата; теперь для описания любой из 64 точек мы будем пользоваться парами целых чисел от 0 до 7. На рис. 7.22, *в* также изображена точка 5,4, представляющая пример пары значений кодовых символов с шумом. Представим себе, что квадрат на рис. 7.22, *в* изображен в координатах (x, y). Каким будет евклидово кодовое расстояние между точкой с шумом 5,4 и точкой без шума 0,0? Оно равно $\sqrt{(5-0)^2+(4-0)^2} = \sqrt{41}$. А если мы захотим узнать евклидово кодовое расстояние между точкой с шумом 5,4 и точкой без шума 7,7? Аналогично $\sqrt{(5-7)^2+(4-7)^2} = \sqrt{13}$.

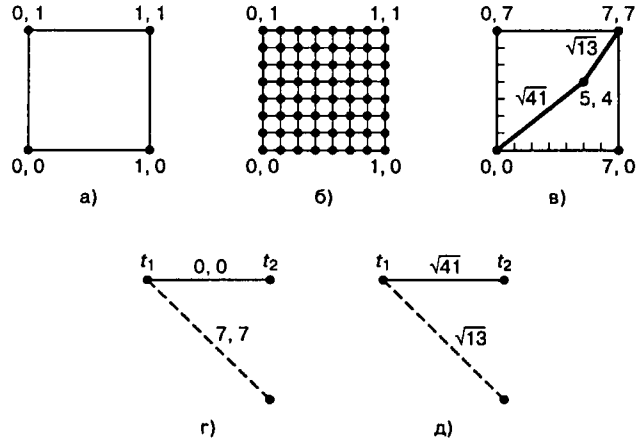


Рис. 7.22. Декодирование Витерби: а) плоскость жесткой схемы принятия решений; б) 8-уровневая плоскость мягкой схемы принятия решений; в) пример мягких кодовых символов; г) секция решетки кодирования; д) секция решетки декодирования

Мягкое декодирование по алгоритму Витерби, по большей части, осуществляется так же, как и жесткое декодирование (как описывалось в разделах 7.3.4 и 7.3.5). Единственное отличие состоит в том, что здесь не используется расстояние Хэмминга. Поэтому рассмотрим мягкое декодирование, осуществляемое с евклидовым кодовым расстоянием. На рис. 7.22, *г* показана первая секция решетки кодирования, которая вначале имела вид, приведенный на рис. 7.7. При этом кодовые слова преобразованы из двоичных в восьмеричные. Допустим, что пара кодовых символов, поступившая на декодер во время первого перехода, согласно мягкой схеме декодирования имеет значения 5,4. На рис. 7.22, *д* показана первая секция решетки декодирования. Метрика $(\sqrt{41})$, представляющая евклидово кодовое расстояние между прибывшим ответвленным словом 5,4 и ответвленным словом 0,0, обозначена сплошной линией. Аналогично метрика $(\sqrt{13})$ представляет собой евклидово кодовое расстояние между поступившим кодовым символом 5,4 и кодовым символом 7,7; это расстояние показано пунктирной линией. Оставшаяся часть задачи декодирования, которая сводится к отсечению решетки и поиску полной ветви, осуществляется аналогично схеме жесткого декодирования. Заметим, что в реальных микросхемах, предназначенных для сверточного декодирования, евклидово кодовое расстояние в действительности не применяется, вместо него используется монотонная метрика, которая обладает сходными свойствами, но значительно проще в реализации. Примером такой метрики является квадрат евклидова кодового расстояния, в котором исключается рассмотренная выше операция взятия квадратного корня. Более того, если двоичные кодовые символы представлены биполярными величинами, тогда можно использовать метрику скалярного произведения, определяемую уравнением (7.9). При такой метрике вместо минимального расстояния мы должны будем рассматривать максимальные корреляции.

7.5. Другие алгоритмы сверточного декодирования

7.5.1. Последовательное декодирование

Ранее, до того как Витерби открыл оптимальный алгоритм декодирования сверточных кодов, существовали и другие алгоритмы. Самым первым был *алгоритм последовательного декодирования*, предложенный Уозенкрафтом (Wozencraft) [24, 25] и модифицированный Фано (Fano) [2]. В ходе работы последовательного декодера генерируется гипотеза о переданной последовательности кодовых слов и рассчитывается метрика между этой гипотезой и принятым сигналом. Эта процедура продолжается до тех пор, пока метрика показывает, что выбор гипотезы правдоподобен, в противном случае гипотеза последовательно заменяется, пока не будет найдена наиболее правдоподобная. Поиск при этом происходит методом проб и ошибок. Для мягкого или жесткого декодирования можно разработать последовательный декодер, но обычно мягкого декодирования стараются избегать из-за сложных расчетов и большой требовательности к памяти.

Рассмотрим ситуацию, когда используется кодер, изображенный на рис. 7.3, и последовательность $\mathbf{m} = 1\,1\,0\,1\,1$ кодирована в последовательность кодовых слов $\mathbf{U} = 1\,1\,0\,1\,0\,1\,0\,0\,0\,1$, как было в примере 7.1. Допустим, что принятая последовательность **Z** является, фактически, *правильной* передачей **U**. У декодера имеется копия кодового дерева, показанная на рис. 7.6, и он может воспользоваться принятой последовательностью **Z** для прохождения дерева. Декодер начинает с узла дерева в момент t_1 и генерирует оба пути, исходящие из этого узла. Декодер следует пути, который согласуется с полученными n кодовыми символами. На следующем уровне дерева декодер снова генерирует два пути, выходящие из узла, и следует пути, согласующемуся со второй группой n символов. Продолжая аналогичным образом, декодер быстро перебирает все дерево.

Допустим теперь, что принятая последовательность **Z** является *искаженным* кодовым словом **U**. Декодер начинает с узла дерева в момент t_1 и генерирует оба пути, выходящие из этого узла. Если принятые n кодовых символов совпадают с одним из сгенерированных путей, декодер следует этому пути. Если согласования нет, то декодер следует *наиболее вероятному пути*, но при этом ведет общий подсчет несовпадений между принятыми символами и ответвляющимися словами на пути следования. Если две ветви оказываются равновероятными, то приемник делает произвольный выбор, как и в случае с нулевым входным путем. На каждом уровне дерева декодер генерирует новые ветви и сравнивает их со следующим набором n принятых кодовых символов. Поиск продолжается до тех пор, пока все дерево не будет пройдено по наиболее вероятному пути, и при этом составляется счет несовпадений.

Если счет несовпадений превышает некоторое число (оно может увеличиваться после прохождения дерева), декодер решает, что он находится на неправильном пути, отбрасывает этот путь и повторяет все снова. Декодер помнит список отброшенных путей, чтобы иметь возможность избежать их при следующем прохождении дерева. Допустим, кодер, представленный на рис. 7.3, кодирует информационную последовательность $\mathbf{m} = 1\,1\,0\,1\,1$ в последовательность кодовых слов **U**, как показано на рис. 7.1. Предположим, что четвертый и седьмой биты переданной последовательности **U** приняты с ошибкой.

Время		t_1	t_2	t_3	t_4	t_5
Информационная последовательность:	$\mathbf{m} =$	1	1	0	1	1
Переданная последовательность:	$\mathbf{U} =$	11	01	01	00	01
Принятая последовательность:	$\mathbf{Z} =$	11	00	01	10	01

Давайте проследим за траекторией пути декодирования на рис. 7.23. Допустим, что критерием возврата и повторного прохождения путей будет общий счет несогласующихся путей, равный 3. На рис. 7.23 числа у путей прохождения представляют собой текущее значение счетчика несовпадений. Итак, прохождение дерева будет иметь следующий вид.

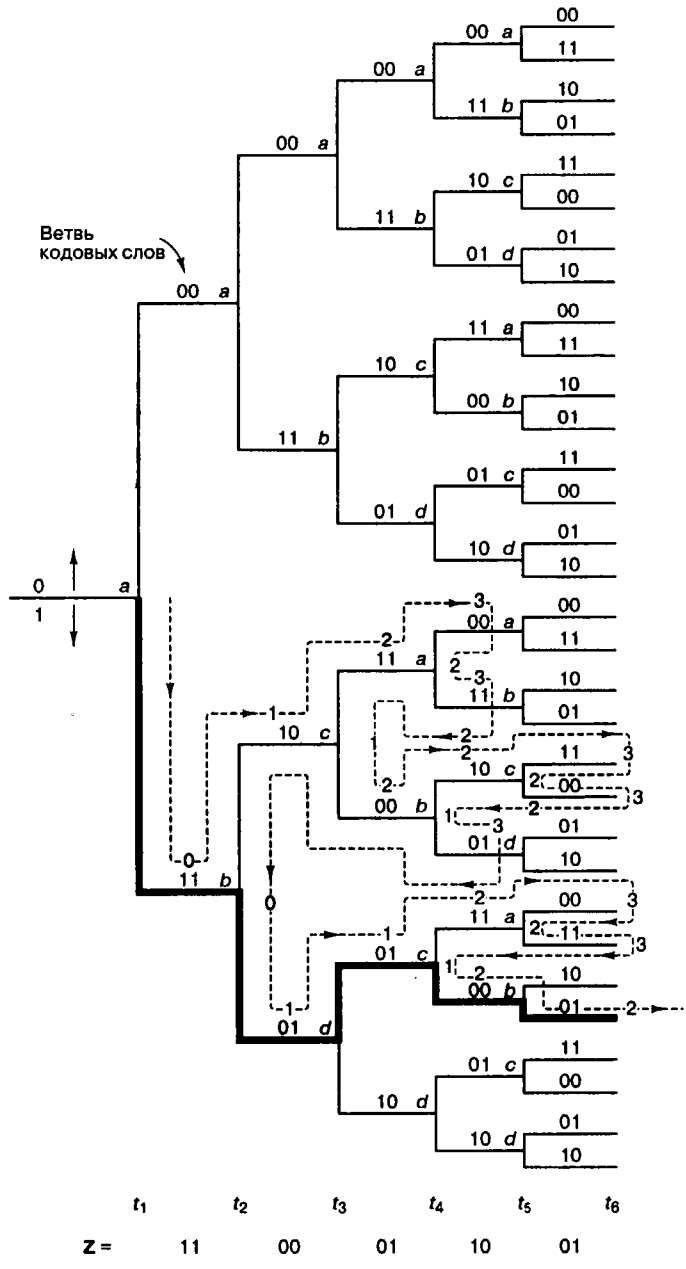


Рис. 7.23. Схема последовательного декодирования

1. В момент времени t_1 мы принимаем символы 11 и сравниваем их с ответвляющимся словами, исходящими из первого узла.
2. Наиболее вероятна та ветвь, у которой ответвляющееся слово 11 (соответствующее входной битовой единице или ответвлению вниз), поэтому декодер решает, что входная битовая единица правильно декодирована, и переходит на следующий уровень.
3. В момент t_2 на этом втором уровне декодер принимает символы 00 и сравнивает их с возможными ответвленными словами 10 и 01.
4. Здесь нет "хорошего" пути, поэтому декодер произвольно выбирает путь, соответствующий входному битовому нулю (или ответвляющемуся слову 10), и счетчик несовпадений регистрирует 1.
5. В момент времени t_3 декодер принимает символы 01 и сравнивает их на третьем уровне с ответвляющимися словами 11 и 00.
6. Здесь снова ни один из путей не имеет преимуществ. Декодер произвольно выбирает нулевой входной путь (или ответвленное слово 11), и счетчик несовпадений возрастает до 2.
7. В момент t_4 декодер принимает символы 10 и сравнивает их на четвертом уровне с ответвленными словами 00 и 11.
8. Здесь снова ни один из путей не имеет преимуществ, и декодер произвольно выбирает нулевой входной путь (или ответвленное слово 00); счетчик несовпадений возрастает до 3.
9. Поскольку счет несовпадений, равный 3, соответствует точке возврата, декодер "делает откат" и пробует альтернативный путь. Счетчик переустанавливается на 2 несовпадения.
10. Альтернативный путь на четвертом уровне соответствует пути входной битовой единицы (или ответвленному слову 11). Декодер принимает этот путь, но после сравнения его с принятыми символами 10 несовпадение остается равным 1, и счетчик устанавливается равным 3.
11. Счет 3 является критерием точки возврата, поэтому декодер делает откат назад с этого пути, и счетчик снова устанавливается на 2. На данном уровне t_4 все альтернативные пути использованы, поэтому декодер возвращается на узел в момент t_3 и переустанавливает счетчик на 1.
12. В узле t_3 декодер сравнивает символы, принятые в момент времени t_3, а именно 01, с неиспользованным путем 00. В данном случае несовпадение равно 1, и счетчик устанавливается на 2.
13. В узле t_4 декодер следует за ответвляющимся словом 10, которое совпадает с принятым в момент t_4 кодовым символом 10. Счетчик остается равным 2.
14. В узле t_5 ни один из путей не имеет преимуществ, и декодер, как и определяется правилами, следует верхней ветви. Счетчик устанавливается на 3 несовпадения.
15. При таком счете декодер делает откат, переустанавливает счетчик на 2 и пробует альтернативный путь в узле t_5. Поскольку другим ответвляющимся словом является 00, снова получаем одно несовпадение с принятым в момент t_5 кодовым символом 01, и счетчик устанавливается равным 3.

16. Декодер уходит с этого пути, и счетчик переустанавливается на 2. На этом уровне t_5 все альтернативные пути использованы, поэтому декодер возвращается на узел в момент t_4 и переустанавливает счетчик на 1.
17. Декодер пробует альтернативный путь в узле t_4, метрика которого возрастает до 3, поскольку в ответвляющемся слове имеется несовпадение в двух позициях. В этот момент декодер должен сделать откат всех путей до момента t_2, поскольку все пути более высоких уровней уже использованы. Счетчик снова переустановлен на нуль.
18. В узле t_2 декодер следует ответвляющемуся слову 01. Поскольку имеется несовпадение в одной позиции с принятыми в момент t_2 кодовыми символами 00, то счетчик устанавливается на 1.

Далее декодер продолжает свои поиски таким же образом. Как видно из рис. 7.23, финальный путь, счетчик которого не нарушает критерия точки возврата, дает правильно декодированную информационную последовательность 1 1 0 1 1. Последовательное декодирование можно понимать как тактику проб и ошибок для поиска правильного пути на кодовом дереве. Поиск осуществляется последовательно; всегда рассматривается только один путь за раз. Если принимается неправильное решение, последующие пути будут ошибочными. Декодер может со временем распознать ошибку, отслеживая метрики пути. Алгоритм напоминает путешественника, отыскивающего путь на карте дорог. До тех пор, пока путешественник видит, что дорожные ориентиры соответствуют таковым на карте, он продолжает путь. Когда он замечает странные ориентиры (увеличение его своеобразной метрики), в конце концов приходит к выводу, что он находится на неправильном пути, и возвращается к точке, где он может узнать ориентиры (его метрика возвращается в приемлемые рамки). Тогда он пробует альтернативный путь.

7.5.2. Сравнение декодирования по алгоритму Витерби с последовательным декодированием и их ограничения

Главный недостаток декодирования по алгоритму Витерби заключается в том, что, когда вероятность появления ошибки экспоненциально убывает с ростом длины кодового ограничения, число кодовых состояний, а значит сложность декодера, *экспоненциально растет с увеличением длины кодового ограничения*. С другой стороны, вычислительная сложность алгоритма Витерби является независимой характеристикой канала (в отличие от жесткого и мягкого декодирования, которые требуют обычного увеличения объемов вычислений). Последовательное декодирование асимптотически достигает той же вероятности появления ошибки, что и декодирование по принципу максимального правдоподобия, но без поиска всех возможных состояний. Фактически при последовательном декодировании число перебираемых состояний существенно *независимо от длины кодового ограничения*, и это позволяет использовать очень большие ($K = 41$) длины кодового ограничения. Это является важным фактором при обеспечении таких низких вероятностей появления ошибок. Основным недостатком последовательного декодирования является то, что количество перебираемых метрик состояний является случайной величиной. Для последовательного декодирования ожидаемое число неудачных гипотез и повторных переборов является функцией канального отношения сигнал/шум (signal to noise ratio — SNR). При низком SNR приходится перебирать больше гипотез, чем при высоком SNR. Из-за такой изменчивости вычислительной нагрузки, поступившие последовательности необходимо сохранять в буфере памяти. При низком SNR последовательности поступают в буфер до тех пор, пока декодер не сможет найти вероятную ги-

потезу. Если средняя скорость передачи символов превышает среднюю скорость декодирования, буфер будет переполняться, вне зависимости от его емкости, и данные будут теряться. Обычно, пока идет переполнение, буфер убирает данные без ошибок, в то время как декодер пытается выполнить процедуру восстановления. Порог переполнения буфера существенно зависит от SNR. Поэтому важным техническим требованием к последовательному декодеру является *вероятность переполнения буфера.*

На рис. 7.24 показаны типичные кривые, отображающие зависимость P_B от E_b/N_0 для двух распространенных схем — декодирования по алгоритму Витерби и последовательного декодирования. Здесь сравниваются их характеристики при использовании когерентной модуляции BPSK в канале AWGN. Сравниваются кривые для декодирования по алгоритму Витерби (степень кодирования 1/2 и 1/3, $K = 7$, жесткое декодирование), декодирования по алгоритму Витерби (степень кодирования 1/2 и 1/3, $K = 7$, мягкое декодирование) и последовательного декодирования (степень кодирования 1/2 и 1/3, $K = 41$, жесткое декодирование). Из рис. 7.24 можно видеть, что при последовательном декодировании можно достичь эффективности кодирования порядка 8 дБ при $P_B = 10^{-6}$. Поскольку в работе Шеннона (Shannon) [26] предсказывается потенциальная эффективность кодирования около 11 дБ, по сравнению с некодированной передачей с модуляцией BPSK, похоже, что, в основном, теоретически достижимые возможности уже получены.

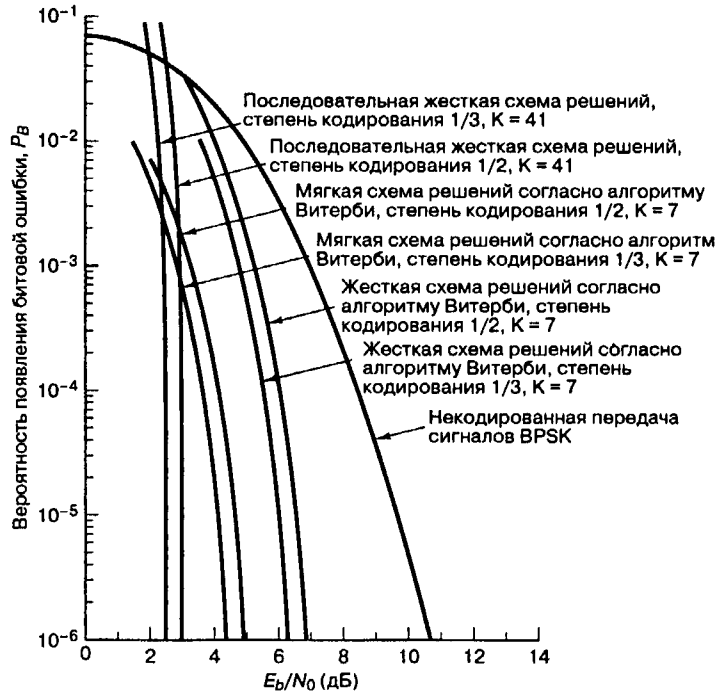


Рис. 7.24. Вероятности появления битовых ошибок для различных схем декодирования по алгоритму Витерби и последовательного декодирования при когерентной модуляции BPSK в канале AWGN. (Перепечатано с разрешения авторов из J. K. Omura and B. K. Levitt. "Coded Error Probability Evaluation for Antijam Communication Systems". IEEE Trans. Commun., vol. COM30, n. 5, May, 1982, Fig. 4, p. 900. © 1982, IEEE.)

7.5.3. Декодирование с обратной связью

Декодер с обратной связью реализует жесткую схему принятия решений относительно информационного бита в разряде j, исходя при этом из метрик, полученных из разрядов $j, j+1, ..., j+m$, где m — заранее установленное положительное целое число. *Длина упреждения* (look-ahead length) L определяется как $L = m + 1$, количество принятых кодовых символов, выраженных через соответствующее число входных битов, задействованных для декодирования информационного бита. Решение о том, является ли информационный бит нулем или единицей, принимается в зависимости от того, на какой ветви путь минимального расстояния Хэмминга переходит в *окне упреждения* (look-ahead window) из разряда j в разряд $j+m$. Поясним это на конкретном примере. Рассмотрим декодер с обратной связью, предназначенный для сверточного кода со степенью кодирования 1/2, который показан на рис. 7.3. На рис. 7.25 приведена древовидная диаграмма и работа декодера с обратной связью при $L = 3$. Иными словами, при декодировании бита из ветви j декодер содержит пути из ветвей $j, j+1$ и $j+2$.

Начиная из первой ветви, декодер вычисляет 2^L (восемь) совокупных метрик путей расстояния Хэмминга и решает, что бит для первой ветви является нулевым, если путь минимального расстояния содержится в верхней части дерева, и единичным, если путь минимального расстояния находится в нижней части дерева. Пусть принята последовательность $\mathbf{Z} = 1\,1\,0\,0\,0\,1\,0\,0\,0\,1$. Рассмотрим восемь путей от момента t_1 до момента t_3 в блоке, обозначенном на рис. 7.24 буквой A, и рассчитаем метрики, сравнивая эти восемь путей для первых шести принятых кодовых символов (три ветви вглубь умножить на два символа для ветви). Выписав метрики Хэмминга общих путей (начиная с верхнего пути), видим, что они имеют следующие значения.

Метрики верхней части	3, 3, 6, 4
Метрики нижней части	2, 2, 1, 3

Видим, что наименьшая метрика содержится в нижней части дерева. Следовательно, первый декодированный бит является единицей (и определяется сдвигом вниз на дереве). Следующий шаг будет состоять в расширении нижней части дерева (выживающий путь) на один разряд глубже, и здесь снова вычисляется восемь метрик, теперь уже для моментов t_2–t_4. Получив, таким образом, два декодированных символа, мы теперь можем сдвинуться на два символа вправо и снова начать расчет метрик путей, но уже для шести кодовых символов. Эта процедура видна в блоке, обозначенном на рис. 7.25 буквой B. И снова, проследив метрики верхних и нижних путей, находим следующее.

Метрики верхней части	2, 4, 3, 3
Метрики нижней части	3, 1, 4, 4

Минимальная метрика для ожидаемой принятой последовательности находится в нижней части блока B. Следовательно, второй декодируемый бит также является единицей.

Таким образом, процедура продолжается до тех пор, пока не будет декодировано все сообщение целиком. Декодер называется *декодером с обратной связью*, поскольку найденное решение подается *обратно* в декодер, чтобы потом использовать его в определении подмножества кодовых путей, которые будут рассматриваться следующими. В канале BSC декодер с обратной связью может оказаться почти таким же эффективным, как и декодер, работающий по алгоритму Витерби [17]. Кроме того, он может исправлять все наиболее вероятные ошибочные комбинации, а именно — те, которые имеют весовой коэффициент $(d_f - 1)/2$ или менее, где d_f — просвет кода. Важным па-

раметром разработки сверточного декодера с обратной связью является L, длина упреждения. Увеличение L приводит к повышению эффективности кодирования, но при этом растет сложность конструкции декодера.

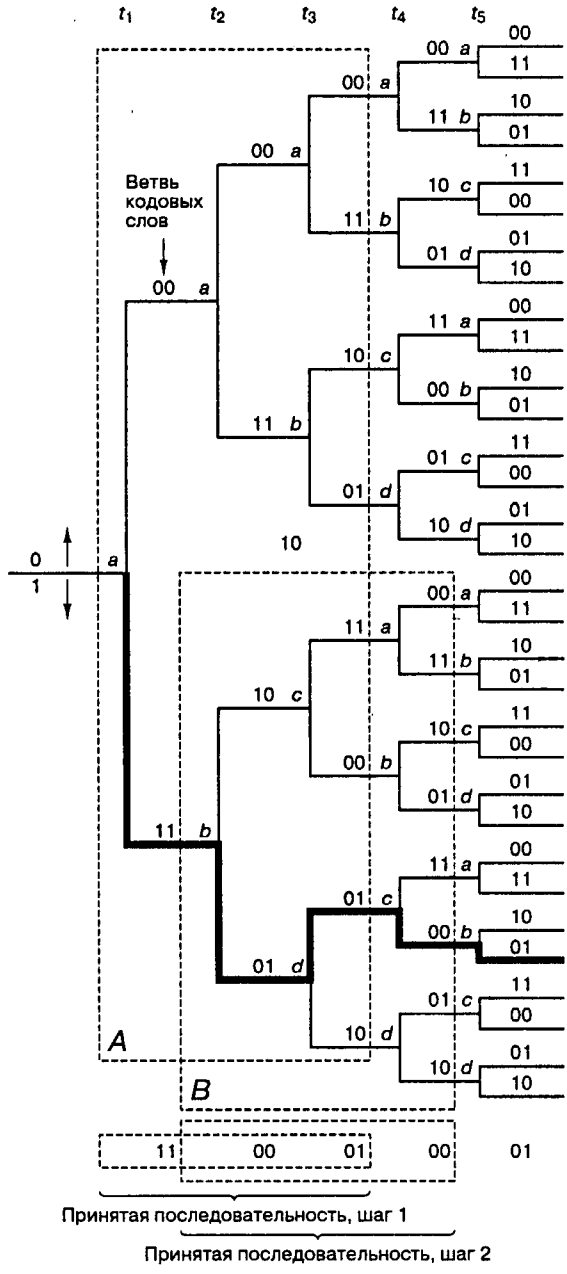


Рис. 7.25. Пример декодирования с обратной связью

7.6. Резюме

В течение последних десяти лет наиболее популярной схемой кодирования являлась сверточная, поскольку почти во всех приложениях сверточные коды лучше блочных при той же конструктивной сложности кодера и декодера. Для каналов спутниковой связи схемы прямого исправления ошибок позволяют легко понизить на 5–6 дБ требуемое значение SNR для заданной достоверности передачи. Из этой эффективности кодирования непосредственно вытекает снижение эффективной изотропной излучаемой мощности спутника (effective isotropic radiated power — EIRP), что, соответственно, приводит к снижению веса и стоимости спутника.

В этой главе мы описали значительную структурную разницу между блочными и сверточными кодами — сверточные коды со степенью кодирования $1/n$ сохраняют в памяти предыдущие $K-1$ бит, где K означает длину кодового ограничения. С такой памятью кодирование каждого входного бита данных зависит не только от значения этого бита, но и от предшествующих ему $K-1$ бит. Задача описывалась в контексте алгоритма максимального правдоподобия. При его использовании изучаются все возможные последовательности кодовых слов, которые могли быть созданы кодером, и выбирается та, которая выглядит статистически наиболее вероятной. Решение опирается на метрику расстояния принятых кодовых символов. Анализ безошибочной работы сверточных кодов является более сложным, чем простое биномиальное разложение, описывающее работу без ошибок многих блочных кодов. Здесь также введено понятие просвета и указана связь между просветом и границами надежной работы. Кроме того, в этой главе описаны основные идеи, касающиеся последовательного декодирования и декодирования с обратной связью, а также приведены некоторые сравнительные характеристики и таблицы различных схем кодирования.

Литература

1. Gallager R. G. *Information Theory and Reliable Communication.* John Wiley & Sons, Inc., New York, 1968.
2. Fano R. M. *A Heuristic Discussion of Probabilistic Decoding.* IRE Trans. Inf. Theory, vol. IT9. n. 2, 1963, pp. 64–74.
3. Odenwalder J. P. *Optimal Decoding of Convolutional Codes.* Ph. D. dissertation, University of California, Los Angeles, 1970.
4. Curry S. J. *Selection of Convolutional Codes Having Large Free Distance.* Ph. D. dissertation, University of California, Los Angeles, 1971.
5. Larsen K. J. *Short Conolutional Codes with Maximal Free Distance for Rates 1/2, 1/3, and 1/4.* IEEE Trans. Inf. Theory, vol. IT19, n. 3, 1973, pp. 371–372.
6. Lin S. and Costello D. J. Jr. *Error Control Coding: Fundamentals and Applications.* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1983.
7. Forney G. D. Jr. *Convolutional Codes: I. Algebraic Structure.* IEEE Trans. Inf. Theory, vol. IT16, n. 6, November, 1970, pp. 720–738.
8. Viterbi A. *Convolutional Codes and Their Performance in Communication Systems.* IEEE Trans. Commun. Technol., vol. COM19, n. 5, October, 1971, pp. 751–772.
9. Forney G. D. Jr. and Bower E. K. *A High Speed Sequential Decoder: Prototype Design and Test.* IEEE Trans. Commun. Technol., vol. COM19, n. 5, October, 1971, pp. 821–835.
10. Jelinek F. *Fast Sequential Decoding Algorithm Using a Stack.* IBM J. Res. Dev., vol.13, November, 1969, pp. 675–685.
11. Massey J. L. *Threshold Decoding.* The MIT Press, Cambridge, Mass., 1963.

12. Heller J. A. and Jacobs I. W. *Viterbi Decoding for Satellite and Space Communication.* IEEE Trans. Commun. Technol., vol. COM19, n. 5, October, 1971, pp. 835–848.
13. Viterbi A. J. *Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm.* IEEE Trans. Inf. Theory, vol. IT13, April, 1967, pp. 260–269.
14. Omura J. K. *On the Viterbi Decoding Algorithm* (correspondence). IEEE Trans. Inf. Theory, vol. IT15, January, 1969, pp. 177–179.
15. Mason S. J. and Zimmerman H. J. *Electronic Circuits, Signals, and Systems.* John Wiley & Sons, Inc. New York, 1960.
16. Clark G. C. and Cain J. B. *Error-Correction Coding for Digital Communications.* Plenum Press, New York, 1981.
17. Viterbi A. J. and Omura J. K. *Principles of Digital Communication and Coding.* McGraw-Hill Book Company, New York, 1979.
18. Massey J. L. and Sain M. K. *Inverse of Linear Sequential Circuits.* IEEE Trans. Comput., vol. C17, April, 1968, pp. 330–337.
19. Rosenberg W. J. *Structural Properties of Convolutional Codes.* Ph. D. dissertation, University of California, Los Angeles, 1971.
20. Bhargava V. K., Haccoun D., Matyas R. and Nuspl P. *Digital Communications by Satellite.* John Wiley & Sons, Inc., New York, 1981.
21. Jacobs I. M. *Practical Applications of Coding.* IEEE Trans. Inf. Theory, vol. IT20, May, 1974, pp. 305–310.
22. Linkabit Corporation. *Coding Systems Study for High Data Rate Telemetry Links.* NASA Ames Res. Center, Final Rep. CR-114278, Contract NAS-2-6-24, Moffett Field, Calif., 1970.
23. Odenwalder J. P. *Error Control Coding Handbook.* Linkabit Corporation, San Diego, Calif., July, 15, 1976.
24. Wozencraft J. M. *Sequential Decoding for Reliable Commumication.* IRE Natl. Conv. Rec., vol. 5, pt. 3, 1957, pp. 11–25.
25. Wozencraft J. M. and Reiffen B. *Sequential Decoding.* The MIT Press, Cambridge, Mass., 1961.
26. Shannon C. E. *A Mathematical Theory of Communication.* Bell Syst. Tech. J., vol. 27, 1948, pp. 379–423, 623–656.

Задачи

7.1. Нарисуйте диаграмму состояний, древовидную и решетчатую диаграммы для кода со степенью кодирования 1/3 при $K = 3$, который имеет следующие генераторы.

$$\mathbf{g}_1(X) = X + X^2$$
$$\mathbf{g}_2(X) = 1 + X$$
$$\mathbf{g}_3(X) = 1 + X + X^2$$

7.2. Дан двоичный сверточный код со степенью кодирования 1/2 и $K = 3$ с частично заполненной диаграммой состояний, изображенной на рис. 37.1. Найдите полную диаграмму состояний и опишите ее для кодера.

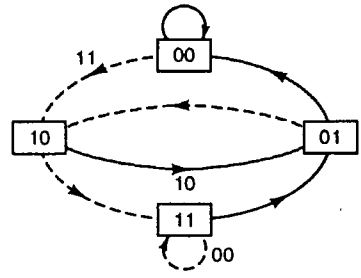


Рис. 37.1

7.3. Нарисуйте диаграмму состояний, древовидную и решетчатую диаграммы для сверточного кодера, который описывается блочной диаграммой, показанной на рис. 37.2.

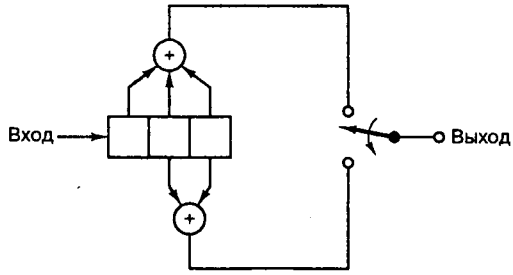


Рис. 37.2

7.4. Допустим, что вы пытаетесь найти самый быстрый путь из Лондона в Вену поездом или на судне. Диаграмма на рис. 37.3 построена с учетом различных расписаний. Обозначения возле каждого пути являются временем путешествия. Используя алгоритм Витерби, найдите наиболее быстрый маршрут из Лондона в Вену. Объясните, как работает этот алгоритм, какие вычисления необходимо проделать и какие данные нужно сохранить в памяти для включения их в алгоритм.

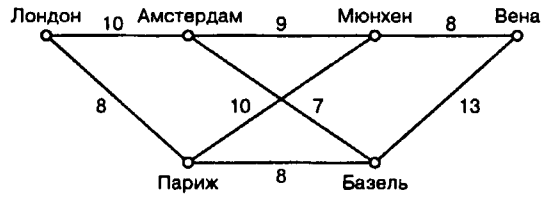


Рис. 37.3

7.5. Рассмотрим сверточный кодер, показанный на рис. 37.4.

а) Запишите векторы и полиномы связи для этого кодера.

б) Нарисуйте диаграмму состояний, древовидную и решетчатую диаграммы.

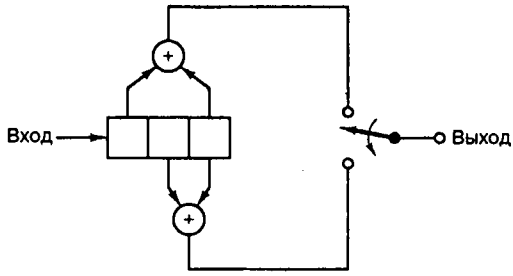


Рис. 37.4

7.6. Какой будет импульсная характеристика в задаче 7.5? Используя эту характеристику, определите выходную последовательность, если на вход подается 1 0 1. Проверьте ответ с помощью полиномиальных генераторов.

7.7. Будет ли кодер, описанный в задаче 7.5, давать возможность для накопления катастрофической ошибки? Приведите пример в защиту своего ответа.

7.8. Найдите просвет для кодера из задачи 7.3, используя передаточную функцию.

7.9. Пусть кодовые слова в схеме кодирования имеют следующий вид.

$$a = 0\,0\,0\,0\,0\,0$$
$$b = 1\,0\,1\,0\,1\,0$$
$$c = 0\,1\,0\,1\,0\,1$$
$$d = 1\,1\,1\,1\,1\,1$$

Если по двоичному симметричному каналу принимается последовательность 1 1 1 0 1 0 и при этом осуществляется декодирование по принципу максимального правдоподобия, то каким будет декодированный символ?

7.10. Пусть на двоичном симметричном канале (binary symmetric channel — BSC) используется кодер со степенью кодирования 1/2 и $K = 3$, как показано на рис. 7.3. Допустим, что начальным состоянием кодера будет 00. На выходе канала BSC принимается последовательность $\mathbf{Z} = (1\,1\,0\,0\,0\,0\,1\,0\,1\,1$ и остальное все "0").

а) Найдите на решетчатой диаграмме максимально правдоподобный путь и определите первые 5 декодированных информационных битов. При наличии двух сливающихся путей выбирайте верхнюю ветвь.

б) Определите канальные биты в $\mathbf{Z}$, которые подверглись искажению в ходе передачи.

7.11. Выясните, какие из следующих ниже кодов со степенью кодирования 1/2 будут катастрофическими.

а) $\mathbf{g}_1(X) = X^2$, $\mathbf{g}_2(X) = 1 + X + X^3$

б) $\mathbf{g}_1(X) = 1 + X^2$, $\mathbf{g}_2(X) = 1 + X^3$

в) $\mathbf{g}_1(X) = 1 + X + X^2$, $\mathbf{g}_2(X) = 1 + X + X^3 + X^4$

г) $\mathbf{g}_1(X) = 1 + X + X^3 + X^4$, $\mathbf{g}_2(X) = 1 + X^2 + X^4$

д) $\mathbf{g}_1(X) = 1 + X^4 + X^6 + X^7$, $\mathbf{g}_2(X) = 1 + X^3 + X^4$

е) $\mathbf{g}_1(X) = 1 + X^3 + X^4$, $\mathbf{g}_2(X) = 1 + X + X^2 + X^4$

7.12. а) Рассмотрим сигнал BPSK с когерентным обнаружением, кодируемый с помощью кодера, показанного на рис. 7.3. Найдите верхнюю границу вероятности появления битовой ошибки, P_B, если номинальное значение E_b/N_0 равно 6 дБ. Предполагается жесткое декодирование.

б) Сравните значение P_B с некодированным случаем и определите выигрыш в отношении сигнал/шум.

7.13. С помощью последовательного декодирования изобразите путь вдоль древовидной диаграммы, показанной на рис. 7.22, если принята последовательность 0 1 1 1 0 0 0 1 1 1. Критерием отката будет три несовпадения.

7.14. Повторите пример декодирования из задачи 7.13, воспользовавшись декодированием с обратной связью при длине упреждения 3. В случае появления связи выбирайте верхнюю часть дерева.

7.15. На рис. 37.5 показан сверточный кодер с длиной кодового ограничения, равной 2.

а) Нарисуйте диаграмму состояний, древовидную и решетчатую диаграммы.

б) Допустим, что от этого кодера поступило сообщение 1 1 0 0 1 0. Декодируйте это сообщение, воспользовавшись алгоритмом декодирования с обратной связью и считая длину упреждения равной 2.

7.16. С помощью данных об ответвляющемся слове решетки кодера на рис. 7.7, декодируйте последовательность $\mathbf{Z} = (01\ 11\ 00\ 01\ 11$, остальные все "0"), считая, что используется жесткая схема принятия решений и алгоритм декодирования Витерби.

7.17. Рассмотрим сверточный кодер со степенью кодирования 2/3, показанный на рис. 37.6. За раз в кодер подается $k = 2$ бит; $n = 3$ бит подается на выход кодера. Имеется $kK = 4$ разряда регистра, и длина кодового ограничения равна $K = 2$ в единицах 2-битовых байтов. Со-

стояние кодера определяется как содержимое $K-1$ крайних правых разрядов k-кортежа. Нарисуйте диаграмму состояний, древовидную и решетчатую диаграммы.

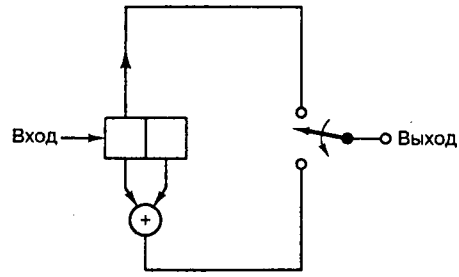


Рис. 37.5

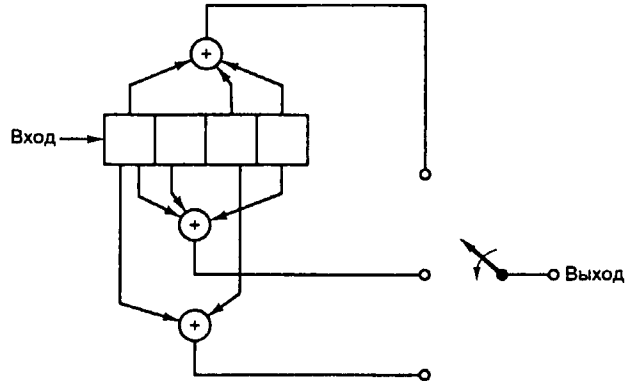


Рис. 37.6

7.18. Найдите додетекторное значение спектральной плотности отношения сигнал/шум P_r/N_0, требуемое для получения скорости передачи декодированных данных в 1 Мбит/с, при вероятности появления ошибки 10^{-5}. Предположите, что применяется двоичная некогерентная модуляция FSK. Также считайте, что осуществляется сверточное кодирование и

$$P_B = 2000 p_c^4,$$

где p_c и P_B — это вероятности появления ошибок внутри и вне декодера,.

7.19. Исходя из табл. 7.4, разработайте двоичный сверточный кодер со степенью кодирования 1/2 и $K = 4$.

а) Нарисуйте его блок-схему.

б) Нарисуйте решетку кодирования и обозначьте на ней состояния и ответвляющиеся слова.

в) Подберите ячейки, которые должны быть реализованы в алгоритме ACS.

7.20. Для следующей демодулированной последовательности выполните мягкое декодирование, используя код со степенью кодирования 1/2 и $K = 3$, который описывается схемой кодера, изображенной на рис. 7.3. Сигналы — это квантованные на 8 уровней целые числа от 0 до 7. Уровень 0 представляет собой идеальный двоичный 0, а уровень 7 — идеальную двоичную 1. Вход декодера: 6, 7, 5, 3, 1, 0, 1, 1, 2, 0, где крайнее левое число является самым первым. Декодируйте первые два бита данных, используя решетчатую диаграмму декодирования. Предположите, что кодер начинает из состояния 00 и процесс декодирования полностью синхронизирован.

Вопросы для самопроверки

7.1. Зачем нужна периодическая *очистка* регистра при сверточном кодировании (см. разделы 7.2.1 и 7.3.4)?

7.2. Дайте определение *состоянию* системы (см. раздел 7.2.2).

7.3. Что такое *конечный автомат* (см. раздел 7.2.2)?

7.4. Что такое *мягкая схема принятия решений* и насколько *более сложным* является мягкое декодирование по алгоритму Витерби в сравнении с жестким декодированием (см. разделы 7.3.2 и 7.4.8)?

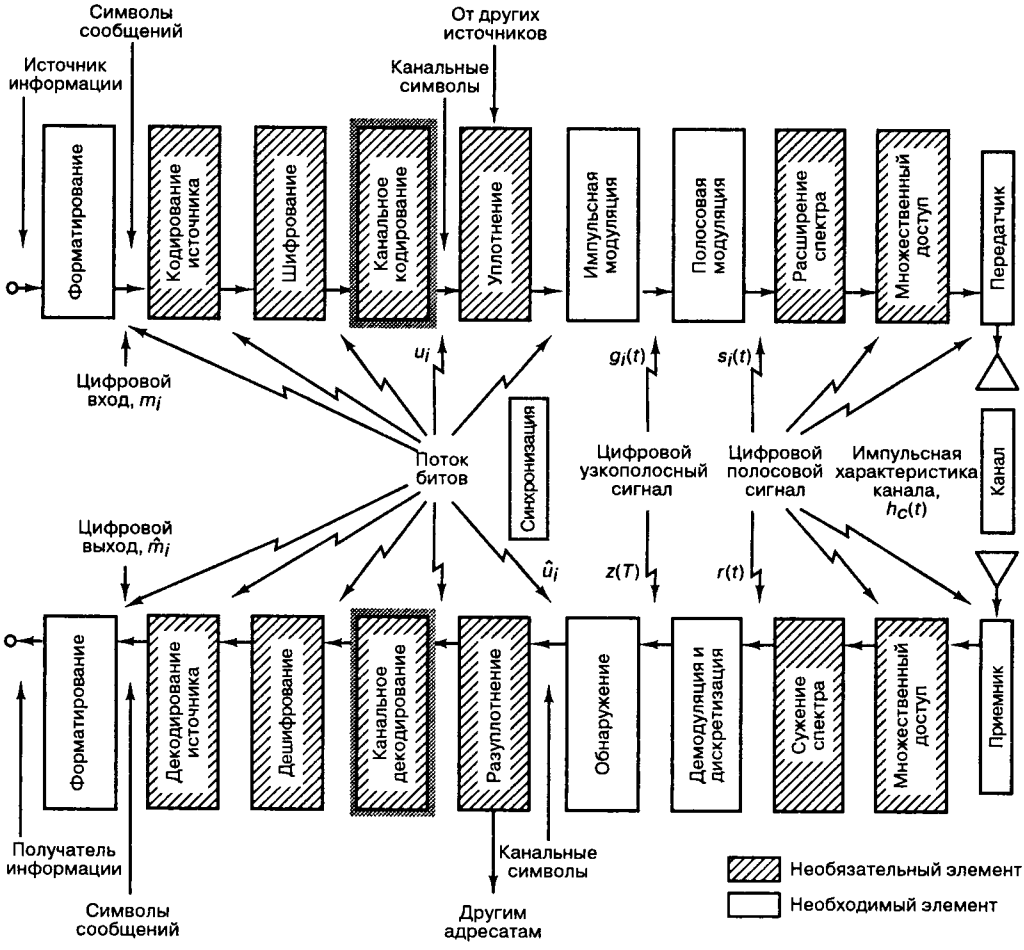
7.5. Каково иное (описательное) название двоичного симметричного канала (binary symmetric channel — BSC) (см. раздел 7.3.2.1)?

7.6. Опишите расчеты *процедуры сложения, сравнения и выбора* (add-compare-select — ASC), которые осуществляются в ходе декодирования по алгоритму Витерби (см. раздел 7.3.5).

7.7. На решетчатой диаграмме *ошибка* соответствует выжившему пути, который сначала *расходится*, а затем снова *сливается* с правильным путем. Почему пути должны повторно сливаться (см. раздел 7.4.1)?

ГЛАВА 8

Канальное кодирование: часть 3



8.1. Коды Рида-Соломона

Коды Рида-Соломона (Reed-Solomon code, R-S code) — это *недвоичные циклические* коды, символы которых представляют собой m-битовые последовательности, где m — положительное целое число, большее 2. Код (n, k) определен на m-битовых символах при всех n и k, для которых

$$0 < k < n < 2^m + 2, \tag{8.1}$$

где k — число информационных битов, подлежащих кодированию, а n — число кодовых символов в кодируемом блоке. Для большинства сверточных кодов Рида-Соломона (n, k)

$$(n, k) = (2^m - 1, 2^m - 1 - 2t), \tag{8.2}$$

где t — количество ошибочных битов в символе, которые может исправить код, а $n - k = 2t$ — число контрольных символов. Расширенный код Рида-Соломона можно получить при $n = 2^m$ или $n = 2^m + 1$, но не более того.

Код Рида-Соломона обладает *наибольшим* минимальным расстоянием, возможным для линейного кода с одинаковой длиной входных и выходных блоков кодера. Для недвоичных кодов расстояние между двумя кодовыми словами определяется (по аналогии с расстоянием Хэмминга) как число символов, которыми отличаются последовательности. Для кодов Рида-Соломона минимальное расстояние определяется следующим образом [1].

$$d_{\min} = n - k + 1 \tag{8.3}$$

Код, который исправляет все искаженные символы, содержащие ошибку в t или меньшем числе бит, где t приведено в уравнении (6.44), можно выразить следующим образом.

$$t = \left\lfloor \frac{d_{\min} - 1}{2} \right\rfloor = \left\lfloor \frac{n-k}{2} \right\rfloor \tag{8.4}$$

Здесь $\lfloor x \rfloor$ означает наибольшее целое, не превышающее x. Из уравнения (8.4) видно, что коды Рида-Соломона, исправляющие t символьных ошибок, требуют не более $2t$ контрольных символов. Из уравнения (8.4) следует, что декодер имеет $n - k$ "используемых" избыточных символов, количество которых вдвое превышает количество исправляемых ошибок. Для каждой ошибки один избыточный символ используется для обнаружения ошибки и один — для определения правильного значения.

Способность кода к коррекции стираний выражается следующим образом.

$$\rho = d_{\min} - 1 = n - k \tag{8.5}$$

Возможность одновременной коррекции ошибок и стираний можно выразить как требование.

$$2\alpha + \gamma < d_{\min} < n - k \tag{8.6}$$

Здесь α — число символьных ошибочных комбинаций, которые можно исправить, а γ — количество комбинаций символьных стираний, которые могут быть исправлены. Преимущества недвоичных кодов, подобных кодам Рида-Соломона, можно увидеть в следующем сравнении. Рассмотрим двоичный код $(n, k) = (7, 3)$. Полное пространство

n-кортежей содержит $2^n = 2^7 = 128$ n-кортежей, из которых $2^k = 2^3 = 8$ (или 1/16 часть всех n-кортежей) являются кодовыми словами. Затем рассмотрим недвоичный код $(n, k) = (7, 3)$, где каждый символ состоит из $m = 3$ бит. Пространство n-кортежей содержит $2^{nm} = 2^{21} = 2\,097\,152$ n-кортежа, из которых $2^{km} = 2^9 = 512$ (или 1/4096 часть всех n-кортежей) являются кодовыми словами. Если операции производятся над недвоичными символами, каждый из которых образован m битами, то только незначительная часть (т.е. 2^{km} из большого числа 2^{nm}) возможных n-кортежей является кодовыми словами. Эта часть уменьшается с ростом m. Здесь важным является то, что если в качестве кодовых слов используется незначительная часть пространства n-кортежей, то можно достичь большего $d_{\min}$.

Любой линейный код дает возможность исправить $n - k$ комбинаций символьных стираний, если все $n - k$ стертых символов приходятся на контрольные символы. Однако коды Рида-Соломона имеют замечательное свойство, выражающееся в том, что они могут исправить *любой* набор $n - k$ символов стираний в блоке. Можно сконструировать коды с любой избыточностью. Впрочем, с увеличением избыточности растет сложность ее высокоскоростной реализации. Поэтому наиболее привлекательные коды Рида-Соломона обладают высокой степенью кодирования (низкой избыточностью).

8.1.1. Вероятность появления ошибок для кодов Рида-Соломона

Коды Рида-Соломона чрезвычайно эффективны для *исправления пакетов ошибок*, т.е. они оказываются эффективными в каналах с памятью. Также они хорошо зарекомендовали себя в каналах с большим набором входных символов. Особенностью кода Рида-Соломона является то, что к коду длины n можно добавить два информационных символа, не уменьшая при этом минимального расстояния. Такой расширенный код имеет длину $n + 2$ и то же количество символов контроля четности, что и исходный код. Из уравнения (6.46) вероятность появления ошибки в декодированном символе, P_E, можно записать через вероятность появления ошибки в канальном символе, p [2].

$$P_E \approx \frac{1}{2^m - 1} \sum_{j=t+1}^{2^m-1} j \binom{2^m - 1}{j} p^j (1-p)^{2^m - 1 - j} \tag{8.7}$$

Здесь t — количество ошибочных битов в символе, которые может исправить код, а символы содержат m битов каждый.

Для некоторых типов модуляции вероятность битовой ошибки можно ограничить сверху вероятностью символьной ошибки. Для модуляции MFSK с $M = 2^m$ связь P_B и P_E выражается формулой (4.112).

$$\frac{P_B}{P_E} = \frac{2^{m-1}}{2^m - 1} \tag{8.8}$$

На рис. 8.1 показана зависимость P_B от вероятности появления ошибки в канальном символе p, полученная из уравнений (8.7) и (8.8) для различных ортогональных 32-ричных кодов Рида-Соломона с возможностью коррекции t ошибочных бит в символе и $n = 31$ (тридцать один 5-битовый символ в кодовом блоке). На рис. 8.2 показана зависимость P_B от E_b/N_0 для таких систем кодирования при ис-

пользовании модуляции MFSK и некогерентной демодуляции в канале AWGN [2]. Для кодов Рида-Соломона вероятность появления ошибок является убывающей степенной функцией длины блока, n, а сложность декодирования пропорциональна небольшой степени длины блока [1]. Иногда коды Рида-Соломона применяются в каскадных схемах. В таких системах внутренний сверточный декодер сначала осуществляет некоторую защиту от ошибок за счет мягкой схемы решений на выходе демодулятора; затем сверточный декодер передает данные, оформленные согласно жесткой схеме, на внешний декодер Рида-Соломона, что снижает вероятность появления ошибок. В разделах 8.2.3 и 8.3 мы рассмотрим каскадное декодирование и декодирование Рида-Соломона на примере системы цифровой записи данных на аудиокомпакт-дисках (compact disc — CD).

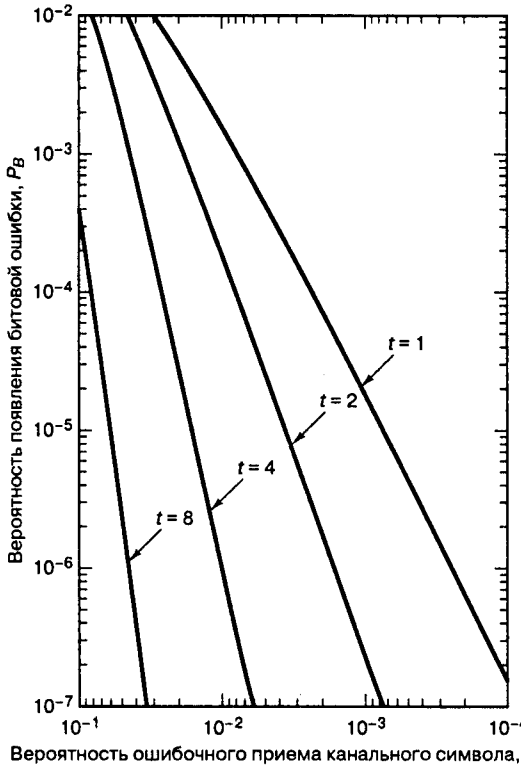


Рис. 8.1. Зависимость P_B от p для различных ортогональных 32-ричных кодов Рида-Соломона с возможностью коррекции t бит в символе и $n = 31$. (Перепечатано с разрешения автора из Data Communications, Network, and Systems, *ed. Thomas C. Bartee, Howard W. Sams Company, Indianapolis, Ind., 1985, p. 311. Ранее публиковалось в J. P. Odenwalder,* Error Control Coding Handbook, *M/A-COM LINKABIT, Inc., San Diego, Calif., July, 15, 1976, p. 91.)*

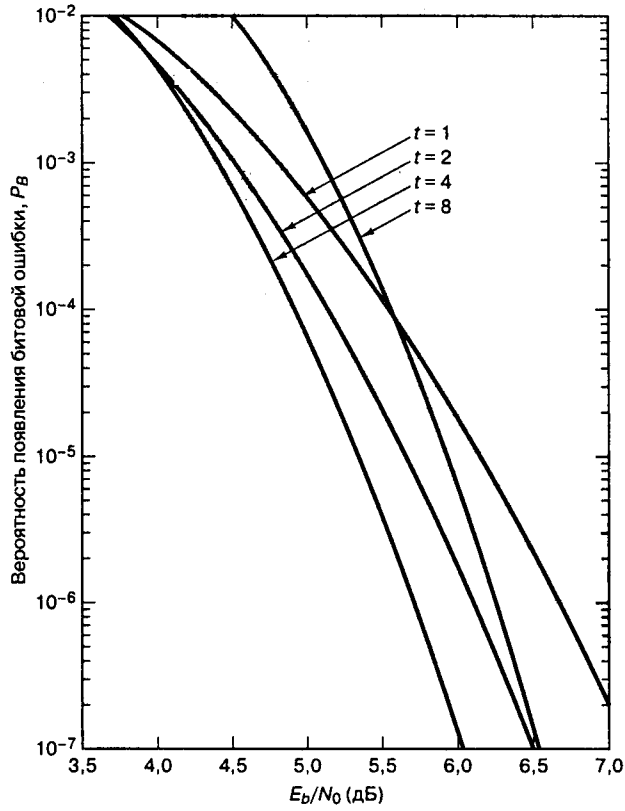


Рис. 8.2. Зависимость P_B от E_b/N_0 для различных ортогональных кодов Рида-Соломона с возможностью коррекции t бит в символе и n = 31, при 32-ричной модуляции MFSK в канале AWGN. (Перепечатано с разрешения автора из Data Communications, Network, and Systems, *ed. Thomas C. Bartee, Howard W. Sams Company, Indianapolis, Ind., 1985, p. 312. Ранее публиковалось в J. P. Odenwalder,* Error Control Coding Handbook, *M/A-COM LINKABIT, Inc., San Diego, Calif., July, 15, 1976, p. 92.)*

8.1.2. Почему коды Рида-Соломона эффективны при борьбе с импульсными помехами

Давайте рассмотрим код $(n, k) = (255, 247)$, в котором каждый символ состоит из $m = 8$ бит (такие символы принято называть *байтами*). Поскольку $n - k = 8$, из уравнения (8.4) можно видеть, что этот код может исправлять любые 4-символьные ошибки в блоке длиной до 255. Пусть блок длительностью 25 бит в ходе передачи поражается помехами, как показано на рис. 8.3. В этом примере пакет шума, который попадает на 25 последовательных битов, исказит точно 4 символа. Декодер для кода (255, 247) исправит *любые* 4-символьные ошибки без учета характера повреждений, причиненных символу. Другими словами, если декодер исправляет байт (заменяет неправильный правильным), то ошибка может быть вызвана искажением одного или всех восьми битов. Поэтому, если символ неправильный, он может быть искажен на всех двоичных позициях. Это дает коду Рида-Соломона огромное

преимущество при наличии импульсных помех по сравнению с двоичными кодами (даже при использовании в двоичном коде чередования). В этом примере, если наблюдается 25-битовая случайная помеха, ясно, что искаженными могут оказаться более чем 4 символа (искаженными могут оказаться до 25 символов). Конечно, исправление такого числа ошибок окажется вне возможностей кода (255, 247).

8.1.3. Рабочие характеристики кода Рида-Соломона как функция размера, избыточности и степени кодирования

Для того чтобы код успешно противостоял шумовым эффектам, длительность помех должна составлять относительно небольшой процент от количества кодовых слов. Чтобы быть уверенным, что так будет большую часть времени, принятый шум необходимо усреднить за большой промежуток времени, что снизит эффект от неожиданной или необычной полосы плохого приема. Следовательно, можно предвидеть, что код с коррекцией ошибок будет более эффективен (повысится надежность передачи) при увеличении размера передаваемого блока, что делает код Рида-Соломона более привлекательным, если желательна большая длина блока [3]. Это можно оценить по семейству кривых, показанному на рис. 8.4, где степень кодирования взята равной 7/8, при этом длина блока возрастает с $n = 32$ символов (при $m = 5$ бит на символ) до $n = 256$ символов (при $m = 8$ бит на символ). Таким образом, размер блока возрастает с 160 бит до 2048 бит.

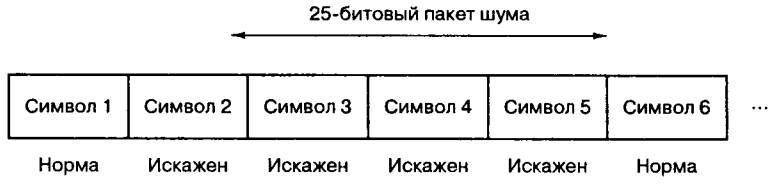


Рис. 8.3. Блок данных, искаженный 25-битовым пакетом ошибок

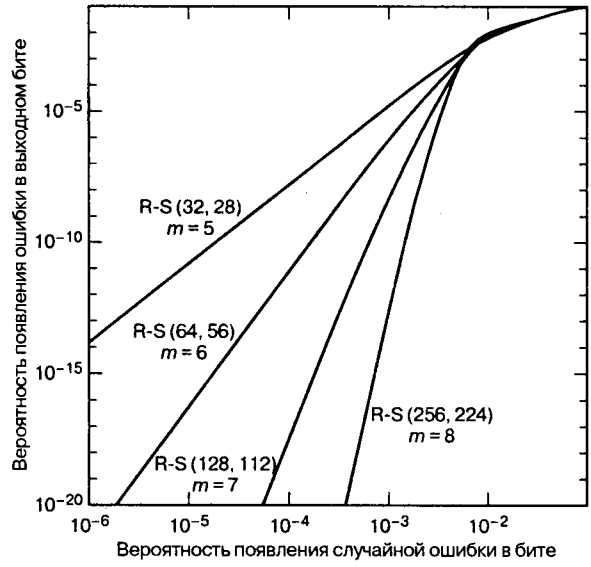


Рис. 8.4. Характеристики декодера Рида-Соломона как функция размера символов (степень кодирования = 7/8)

По мере увеличения избыточности кода (и снижения его степени кодирования), сложность реализации этого кода повышается (особенно для высокоскоростных устройств). При этом для систем связи реального времени должна увеличиться ширина полосы пропускания. Увеличение избыточности, например увеличение размера символа, приводит к уменьшению вероятности появления битовых ошибок, как можно видеть на рис. 8.5, где кодовая длина n равна постоянному значению 64 при снижении числа символов данных с $k = 60$ до $k = 4$ (избыточность возрастает с 4 до 60 символов).

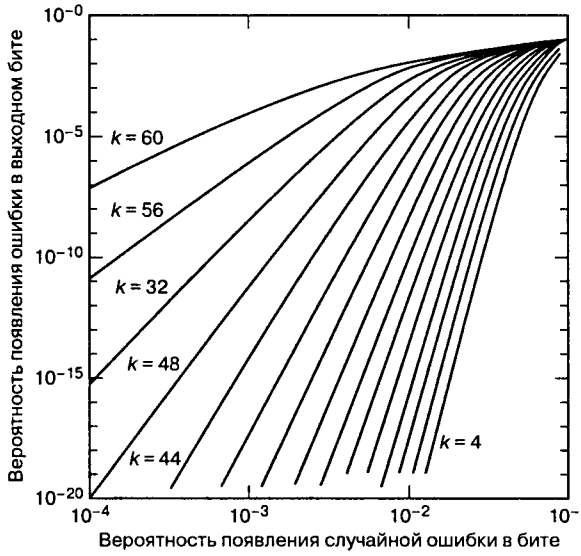


Рис. 8.5. Характеристики декодера Рида-Соломона (64, k) как функция избыточности

На рис. 8.5 показана передаточная функция (выходная вероятность появления битовой ошибки, зависящая от входной вероятности появления символьной ошибки) гипотетических декодеров. Поскольку здесь не имеется в виду определенная система или канал (лишь вход-выход декодера), можно заключить, что надежность передачи является монотонной функцией избыточности и будет неуклонно возрастать с приближением степени кодирования к нулю. Однако это не так для кодов, используемых в системах связи реального времени. По мере изменения степени кодирования кода от максимального значения до минимального (от 0 до 1), интересно было бы понаблюдать за эффектами, показанными на рис. 8.6. Здесь кривые рабочих характеристик показаны при модуляции BPSK и кодах (31, k) для разных типов каналов. На рис. 8.6 показаны системы связи реального времени, в которых за кодирование с коррекцией ошибок приходится платить расширением полосы пропускания, пропорциональным величине, равной обратной степени кодирования. Приведенные кривые показывают четкий оптимум степени кодирования, минимизирующий требуемое значение E_b/N_0 [4]. Для гауссова канала оптимальное значение степени кодирования находится где-то между 0,6 и 0,7, для канала с райсовским замиранием — около 0,5 (с отношением мощности прямого сигнала к мощности отраженного $K = 7$ дБ) и 0,3 — для канала с релеевским замиранием. (Каналы с замиранием будут рассматриваться в главе 15.) Почему здесь как при очень высоких степенях кодирования (малой избыточности), так и при очень низких (значительной из-

быточности) наблюдается ухудшение E_b/N_0? Для высоких степеней кодирования это легко объяснить, сравнивая высокие степени кодирования с оптимальной степенью кодирования. Любой код в целом обеспечивает все преимущества кодирования; следовательно, как только степень кодирования приближается к единице (нет кодирования), система проигрывает в надежности передачи. Ухудшение характеристик при низких степенях кодирования является более тонким вопросом, поскольку в системах связи реального времени используется и модуляция, и кодирование, т.е. работает два механизма. Один механизм направлен на снижение вероятности появления ошибок, другой повышает ее. Механизм, снижающий вероятность появления ошибки, — это кодирование; чем больше избыточность, тем больше возможности кода в коррекции ошибок. Механизм, повышающий эту вероятность, — это снижение энергии, приходящейся на канальный символ (по сравнению с информационным символом), что следует из увеличения избыточности (и более быстрой передачи сигналов в системах связи реального времени). Уменьшенная энергия символа вынуждает демодулятор совершать больше ошибок. В конечном счете второй механизм подавляет первый, поэтому очень низкие степени кодирования вызывают ухудшение характеристик кода.

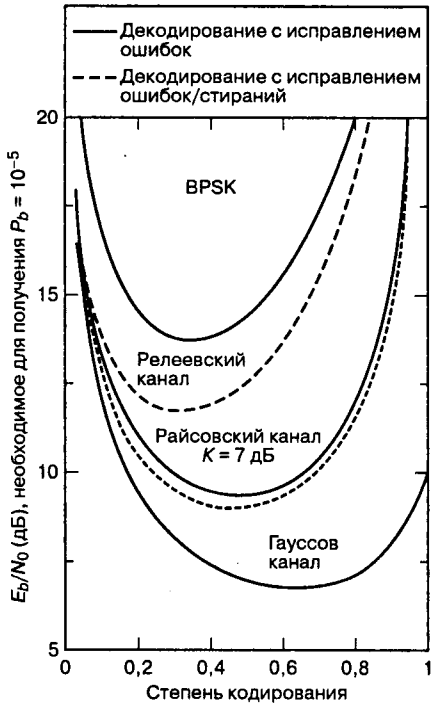


Рис. 8.6. Характеристики декодера Рида-Соломона (31, *k*) *как функция степени кодирования (модуляция BPSK)*

Давайте попробуем подтвердить зависимость вероятности появления ошибок от степени кодирования, показанную на рис. 8.6, с помощью кривых, изображенных на рис. 8.2. Непосредственно сравнить рисунки не удастся, поскольку на рис. 8.6 применяется модуляция BPSK, а на рис. 8.2 — 32-ричная модуляция MFSK. Однако, пожалуй, нам удастся показать, что зависимость характеристик кода Рида-Соломона от его степени кодирования

выглядит одинаково как при BPSK, так и при MFSK. На рис. 8.2 вероятность появления ошибки в канале AWGN снижается при увеличении способности кода t к коррекции символьных ошибок с $t=1$ до $t=4$; случаи $t=1$ и $t=4$ относятся к кодам (31, 29) и (31, 23) со степенями кодирования 0,94 и 0,74. Хотя при $t=8$, что отвечает коду (31, 15) со степенью кодирования 0,48, достоверность передачи $P_B=10^{-5}$ достигается при примерно на 0,5 дБ большем отношении E_b/N_0, по сравнению со случаем $t=4$. Из рис. 8.2 можно сделать вывод, что если нарисовать график зависимости достоверности передачи от степени кодирования кода, то кривая будет иметь вид, подобный приведенному на рис. 8.6. Заметим, что это утверждение нельзя получить из рис. 8.1, поскольку там представлена передаточная функция декодера, которая несет в себе сведения о канале и демодуляции. Поэтому из двух механизмов, работающих в канале, передаточная функция (рис. 8.1) представляет только выгоды, которые проявляются на входе/выходе декодера, и ничего не говорит о потерях энергии как функции низкой степени кодирования. В разделе 9.7.7 будет более подробно рассказано о выборе кода в соответствии с типом модуляции.

8.1.4. Конечные поля

Для понимания принципов кодирования и декодирования недвоичных кодов, таких как коды Рида-Соломона, нужно сделать экскурс в понятие конечных полей, известных как *поля Галуа* (Galois fields — GF). Для любого простого числа p существует конечное поле, которое обозначается GF(p) и содержит p элементов. Понятие GF(p) можно обобщить на поле из p^m элементов, именуемое *полем расширения* GF(p); это поле обозначается GF(p^m), где m — положительное целое число. Заметим, что GF(p^m) содержит в качестве подмножества все элементы GF(p). Символы из поля расширения GF(2^m) используются при построении кодов Рида-Соломона.

Двоичное поле GF(2) является подполем поля расширения GF(2^m), точно так же как поле вещественных чисел является подполем поля комплексных чисел. Кроме чисел 0 и 1, в поле расширения существуют дополнительные однозначные элементы, которые будут представлены новым символом α. Каждый ненулевой элемент в GF(2^m) можно представить как степень α. Бесконечное множество элементов, F, образуется из стартового множества $\{0, 1, \alpha\}$ и генерируется дополнительными элементами путем последовательного умножения последней записи на α.

$$F=\{0, 1, \alpha, \alpha^2, \ldots, \alpha^j, \ldots\}=\{0, \alpha^0, \alpha^1, \alpha^2, \ldots, \alpha^j, \ldots\} \tag{8.9}$$

Для вычисления из F *конечного* множества элементов GF(2^m) на F нужно наложить условия: оно может содержать только 2^m элемента и быть замкнутым относительно операции умножения. Условие замыкания множества элементов поля по отношению к операции умножения имеет вид неприводимого полинома.

$$\alpha^{(2^m-1)}+1=0$$

или, что то же самое,

$$\alpha^{(2^m-1)}=1=\alpha^0 \tag{8.10}$$

С помощью полиномиального ограничения любой элемент со степенью, большей или равной 2^m-1, можно следующим образом понизить до элемента со степенью, меньшей 2^m-1.

$$\alpha^{(2^m+n)} = \alpha^{(2^m-1)}\alpha^{n+1} = \alpha^{n+1} \tag{8.11}$$

Таким образом, как показано ниже, уравнение (8.10) можно использовать для формирования конечной последовательности F^* из бесконечной последовательности F.

$$\begin{aligned} F^* &= \{0, 1, \alpha, \alpha^2, \ldots, \alpha^{2^m-2}, \alpha^{2^m-1}, \alpha^{2^m}, \ldots\} = \\ &= \{0, \alpha^0, \alpha^1, \alpha^2, \ldots, \alpha^{2^m-2}, \alpha^0, \alpha^1, \alpha^2, \ldots\} \end{aligned} \tag{8.12}$$

Следовательно, из уравнения (8.12) можно видеть, что элементы конечного поля GF(2^m) даются следующим выражением.

$$\mathrm{GF}(2^m) = \{0, \alpha^0, \alpha^1, \alpha^2, \ldots, \alpha^{2^m-2}\} \tag{8.13}$$

8.1.4.1. Операция сложения в поле расширения GF(2^m)

Каждый из 2^m элементов конечного поля GF(2^m) можно представить как отдельный полином *степени* $m-1$ или меньше. Степенью полинома называется степень члена максимального порядка. Обозначим каждый ненулевой элемент GF(2^m) полиномом $a_i(X)$, в котором последние m коэффициентов $a_i(X)$ нулевые. Для $i = 0, 1, 2, \ldots, 2^m - 2$,

$$\alpha_i = a_i(X) = a_{i,0} + a_{i,1}X + a_{i,2}X^2 + \ldots + a_{i,m-1}X^{m-1}. \tag{8.14}$$

Рассмотрим случай $m = 3$, в котором конечное поле обозначается GF(2^3). На рис. 8.7 показано отображение семи элементов $\{\alpha_i\}$ и нулевого элемента в слагаемые базисных элементов $\{X^0, X^1, X^2\}$, описываемых уравнением (8.14). Поскольку из уравнения (8.10) $\alpha^0 = \alpha^7$, в этом поле имеется семь ненулевых элементов или всего восемь элементов. Каждая строка на рис. 8.7 содержит последовательность двоичных величин, представляющих коэффициенты $\alpha_{i,0}$, $\alpha_{i,1}$ и $\alpha_{i,2}$ из уравнения (8.14). Одним из преимуществ использования элементов $\{\alpha^i\}$ поля расширения, вместо двоичных элементов, является компактность записи, что оказывается удобным при математическом описании процессов недвоичного кодирования и декодирования. Сложение двух элементов конечного поля, следовательно, определяется как суммирование по модулю 2 всех коэффициентов при элементах одинаковых степеней.

$$\alpha_i + \alpha_j = (a_{i,0} + a_{j,0}) + (a_{i,1} + a_{j,1})X + \ldots + (a_{i,m-1} + a_{j,m-1})X^{m-1} \tag{8.15}$$

8.1.4.2. Описание конечного поля с помощью примитивного полинома

Класс полиномов, называемых *примитивными полиномами*, интересует нас, поскольку такие объекты определяют конечные поля GF(2^m), которые, в свою очередь, нужны для описания кодов Рида-Соломона. Следующее утверждение является необходимым и достаточным условием примитивности полинома. Неприводимый полином $f(X)$ порядка m будет примитивным, если наименьшим положительным целым числом n, для которого $X^n + 1$ делится на $f(X)$, будет $n = 2^m - 1$. Заметим, что неприводимый полином — это такой полином, который нельзя представить в виде произведения полиномов меньшего порядка; делимость A на B означает, что A делится на B с нулевым остатком и ненулевым частным. Обычно полином записывают в порядке

возрастания степеней. Иногда более удобным является обратный формат записи (например, при выполнении полиномиального деления).

Элементы поля	Образующие элементы X^0	X^1	X^2
	0	0	0
α^0	1	0	0
α^1	0	1	0
α^2	0	0	1
α^3	1	1	0
α^4	0	1	1
α^5	1	1	1
α^6	1	0	1
α^7	1	0	0

Рис. 8.7. Отображение элементов поля в базисные элементы GF(8) *с помощью* $f(X) = 1 + X + X^3$

Пример 8.1. Проверка полинома на примитивность

Основываясь на предыдущем определении примитивного полинома, укажите, какие из следующих неприводимых полиномов будут примитивными.

а) $1 + X + X^4$

б) $1 + X + X^2 + X^3 + X^4$

Решение

а) Мы можем проверить этот полином порядка $m = 4$, определив, будет ли он делителем $X^n + 1 = X^{(2^m - 1)} + 1 = X^{15} + 1$ для значений n из диапазона $1 \le n \le 15$. Нетрудно убедиться, что $X^{15} + 1$ делится на $1 + X + X^4$ (см. раздел 6.8.1), и после повторения вычислений можно проверить, что при любых значениях n из диапазона $1 \le n < 15$ полином $X^n + 1$ не делится на $1 + X + X^4$. Следовательно, $1 + X + X^4$ является примитивным полиномом.

б) Легко проверить, что полином является делителем $X^{15} + 1$. Проверив, делится ли $X^n + 1$ на $1 + X + X^2 + X^3 + X^4$, для значений n, меньших 15, можно также видеть, что указанный полином является делителем $X^5 + 1$. Следовательно, несмотря на то что полином $1 + X + X^2 + X^3 + X^4$ является неприводимым, он не будет примитивным.

8.1.4.3. Поле расширения GF(2^3)

Рассмотрим пример, в котором будут задействованы примитивный полином и конечное поле, которое он определяет. В табл. 8.1 содержатся примеры некоторых примитивных полиномов. Мы выберем первый из указанных там полиномов, $f(X) = 1 + X + X^3$, который определяет конечное поле GF(2^m), где степень полинома $m = 3$. Таким образом, в поле, определяемом полиномом $f(X)$, имеется $2^m = 2^3 = 8$ элементов. Поиск корней полинома $f(X)$ — это поиск таких значений X, при которых $f(X) = 0$. Привычные нам двоичные элементы 0 и 1 не подходят полиному $f(X) = 1 + X + X^3$ (они не являются корнями),

поскольку $f(1) = 1$ и $f(0) = 1$ (в рамках операций по модулю 2). Кроме того, основная теорема алгебры утверждает, что полином порядка m должен иметь в точности m корней. Следовательно, в этом примере выражение $f(X) = 0$ должно иметь 3 корня. Возникает определенная проблема, поскольку 3 корня не лежат в том же конечном поле, что и коэффициенты $f(X)$. А если они находятся где-то еще, то, наверняка, в поле расширения $GF(2^3)$. Пусть α, элемент поля расширения, определяется как корень полинома $f(X)$. Следовательно, можно записать следующее.

$$\begin{aligned} f(\alpha) &= 0 \\ 1 + \alpha + \alpha^3 &= 0 \\ \alpha^3 &= -1 - \alpha \end{aligned} \tag{8.16}$$

Поскольку при операциях над двоичным полем $+1 = -1$, то α^3 можно представить следующим образом.

$$\alpha^3 = 1 + \alpha \tag{8.17}$$

Таблица 8.1. Некоторые примитивные полиномы

m		m	
3	$1 + X + X^3$	14	$1 + X + X^6 + X^{10} + X^{14}$
4	$1 + X + X^4$	15	$1 + X + X^{15}$
5	$1 + X^2 + X^5$	16	$1 + X + X^3 + X^{12} + X^{16}$
6	$1 + X + X^6$	17	$1 + X^3 + X^{17}$
7	$1 + X^3 + X^7$	18	$1 + X^7 + X^{18}$
8	$1 + X^2 + X^3 + X^4 + X^8$	19	$1 + X + X^2 + X^5 + X^{19}$
9	$1 + X^4 + X^9$	20	$1 + X^3 + X^{20}$
10	$1 + X^3 + X^{10}$	21	$1 + X^2 + X^{21}$
11	$1 + X^2 + X^{11}$	22	$1 + X + X^{22}$
12	$1 + X + X^4 + X^6 + X^{12}$	23	$1 + X^5 + X^{23}$
13	$1 + X + X^3 + X^4 + X^{13}$	24	$1 + X + X^2 + X^7 + X^{24}$

Таким образом, α^3 представляется в виде взвешенной суммы всех α-членов более низкого порядка. Фактически так можно представить все степени α. Например, рассмотрим следующее.

$$\alpha^4 = \alpha \cdot \alpha^3 = \alpha \cdot (1 + \alpha) = \alpha + \alpha^2 \tag{8.18,а}$$

А теперь взглянем на следующий случай.

$$\alpha^5 = \alpha \cdot \alpha^4 = \alpha \cdot (\alpha + \alpha^2) = \alpha^2 + \alpha^3 \tag{8.18,б}$$

Из уравнений (8.17) и (8.18) получаем следующее.

$$\alpha^5 = 1 + \alpha + \alpha^2 \tag{8.18,в}$$

Используя уравнение (8.18,в), получаем следующее.

$$\alpha^6 = \alpha \cdot \alpha^5 = \alpha \cdot (1 + \alpha + \alpha^2) = \alpha + \alpha^2 + \alpha^3 = 1 + \alpha^2 \tag{8.18,г}$$

А теперь из уравнения (8.18,г) вычисляем

$$\alpha^7 = \alpha \cdot \alpha^6 = \alpha \cdot (1 + \alpha^2) = \alpha + \alpha^3 = 1 = \alpha^0. \qquad (8.18,д)$$

Заметьте, что $\alpha^7 = \alpha^0$ и, следовательно, восемью элементами конечного поля GF(2^3) будут

$$\{0, \alpha^0, \alpha^1, \alpha^2, \alpha^3, \alpha^4, \alpha^5, \alpha^6\}. \qquad (8.19)$$

Отображение элементов поля в базисные элементы, которое описывается уравнением (8.14), можно проиллюстрировать с помощью схемы линейного регистра сдвига с обратной связью (linear feedback shift register — LFSR) (рис. 8.8). Схема генерирует (при $m = 3$) $2^m - 1$ ненулевых элементов поля и, таким образом, обобщает процедуры, описанные в уравнениях (8.17)–(8.19). Следует отметить, что показанная на рис. 8.8 обратная связь соответствует коэффициентам полинома $f(X) = 1 + X + X^3$, как и в случае двоичных циклических кодов (см. раздел 6.7.5). Пусть вначале схема находится в некотором состоянии, например 1 0 0; при выполнении правого сдвига на один такт можно убедиться, что каждый из элементов поля (за исключением нулевого), показанных на рис. 8.7, циклически будет появляться в разрядах регистра сдвига. На данном конечном поле GF(2^3) можно определить две арифметические операции — сложение и умножение. В табл. 8.2 показана операция сложения, а в табл. 8.3 — операция умножения, но только для ненулевых элементов. Правила суммирования следуют из уравнений (8.17) и (8.18,д); и их можно доказать, обратившись к рис. 8.7, поскольку сумму двух элементов поля можно рассчитать путем сложения (по модулю 2) соответствующих коэффициентов их базисных элементов. Правила умножения, указанные в табл. 8.3, следуют из обычной процедуры, в которой произведение элементов поля вычисляется путем сложения по модулю $(2^m - 1)$ их показателей степеней или, для данного случая, по модулю 7.

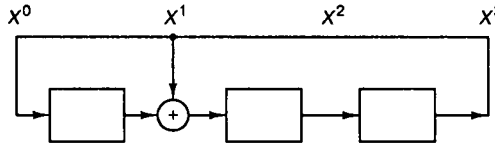


Рис. 8.8. Отображение элементов поля в базисные элементы можно представить с помощью схемы линейного регистра сдвига с обратной связью (linear feedback shift register — LFSR), построенного на примитивном полиноме

Таблица 8.2. Таблица сложения для GF(8) при $f(X) = 1 + X + X^3$

	α^0	α^1	α^2	α^3	α^4	α^5	α^6
α^0	0	α^3	α^6	α^1	α^5	α^4	α^2
α^1	α^3	0	α^4	α^0	α^2	α^6	α^5
α^2	α^6	α^4	0	α^5	α^1	α^3	α^0
α^3	α^1	α^0	α^5	0	α^6	α^2	α^4
α^4	α^5	α^2	α^1	α^6	0	α^0	α^3
α^5	α^4	α^6	α^3	α^2	α^0	0	α^1
α^6	α^2	α^5	α^0	α^4	α^3	α^1	0

Таблица 8.3. Таблица умножения для GF(8) при $f(X) = 1 + X + X^3$

	α^0	α^1	α^2	α^3	α^4	α^5	α^6
α^0	α^0	α^1	α^2	α^3	α^4	α^5	α^6
α^1	α^1	α^2	α^3	α^4	α^5	α^6	α^0
α^2	α^2	α^3	α^4	α^5	α^6	α^0	α^1
α^3	α^3	α^4	α^5	α^6	α^0	α^1	α^2
α^4	α^4	α^5	α^6	α^0	α^1	α^2	α^3
α^5	α^5	α^6	α^0	α^1	α^2	α^3	α^4
α^6	α^6	α^0	α^1	α^2	α^3	α^4	α^5

8.1.4.4. Простой тест для проверки полинома на примитивность

Существует еще один, чрезвычайно простой способ проверки, является ли полином примитивным. У неприводимого полинома, который является примитивным, по крайней мере, хотя бы один из корней должен быть примитивным элементом. *Примитивным элементом* называется такой элемент поля, который, будучи возведенным в более высокие степени, даст все ненулевые элементы поля. Поскольку данное поле является конечным, количество таких элементов также конечно.

Пример 8.2. Примитивный полином должен иметь, по крайней мере, хотя бы один примитивный элемент

Найдите $m = 3$ корня полинома $f(X) = 1 + X + X^3$ и определите, примитивен ли полином. Для этого проверьте, имеется ли среди корней полинома хотя бы один примитивный элемент. Каковы корни полинома? Какие из них примитивны?

Решение

Корни будут найдены прямым перебором. Итак, $\alpha^0 = 1$ не будет корнем, поскольку $f(\alpha^0) = 1$. Теперь, чтобы проверить, является ли корнем α^1, воспользуемся табл. 8.2. Поскольку $f(\alpha) = 1 + \alpha + \alpha^3 = 1 + \alpha^0 = 0$, значит, α будет корнем полинома. Далее поверим, будет ли корнем α^2. $f(\alpha^2) = 1 + \alpha^2 + \alpha^6 = 1 + \alpha^0 = 0$. Значит, и α^2 также будет корнем полинома. Теперь проверим α^3. $f(\alpha^3) = 1 + \alpha^3 + \alpha^2 = 1 + \alpha^5 = \alpha^4 \neq 0$. Следовательно, α^3 корнем полинома не является. Будет ли корнем α^4? $f(\alpha^4) = \alpha^{12} + \alpha^4 + 1 = 1 + \alpha^0 = 0$. Да, α^4 будет корнем полинома. Значит, корнями полинома $f(X) = 1 + X + X^3$ будут α, α^2 и α^4. Нетрудно убедиться, что последовательно возводя в степень любой из этих корней, можно получить все 7 ненулевых элементов поля. Таким образом, все корни будут примитивными элементами. Поскольку в определении требуется, чтобы по крайней мере один из корней был примитивным, полином является примитивным.

В этом примере описан относительно простой метод проверки полинома на примитивность. Для проверяемого полинома нужно составить регистр LFSR с контуром обратной связи, соответствующим коэффициентам полинома, как показано на рис. 8.8. Затем в схему регистра следует загрузить любое ненулевое состояние и выполнять за каждый такт правый сдвиг. Если за один период схема сгенерирует все ненулевые элементы поля, то данный полином с полем GF(2^m) будет примитивным.

8.1.5. Кодирование Рида-Соломона

В уравнении (8.2) представлена наиболее распространенная форма кодов Рида-Соломона через параметры n, k, t и некоторое положительное число $m > 2$. Приведем это уравнение повторно.

$$(n, k) = (2^m - 1, 2^m - 1 - 2t) \tag{8.20}$$

Здесь $n - k = 2t$ — число контрольных символов, а t — количество ошибочных битов в символе, которые может исправить код. Генерирующий полином для кода Рида-Соломона имеет следующий вид.

$$\mathbf{g}(X) = g_0 + g_1X + g_2X^2 + \ldots + g_{2t-1}X^{2t-1} + X^{2t} \tag{8.21}$$

Степень полиномиального генератора равна числу контрольных символов. Коды Рида-Соломона являются подмножеством кодов БХЧ, которые обсуждались в разделе 6.8.3 и показаны в табл. 6.4. Поэтому связь между степенью полиномиального генератора и числом контрольных символов, как и в кодах БХЧ, не должна оказаться неожиданностью. В этом можно убедиться, подвергнув проверке любой генератор из табл. 6.4. Поскольку полиномиальный генератор имеет порядок $2t$, мы должны иметь в точности $2t$ последовательные степени α, которые являются корнями полинома. Обозначим корни $\mathbf{g}(X)$ как: α, α^2, ..., α^{2t}. Нет необходимости начинать именно с корня α, это можно сделать с помощью любой степени α. Возьмем к примеру код (7, 3) с возможностью коррекции двухсимвольных ошибок. Мы выразим полиномиальный генератор через $2t = n - k = 4$ корня следующим образом.

$$\begin{aligned}\mathbf{g}(X) &= (X - \alpha)(X - \alpha^2)(X - \alpha^3)(X - \alpha^4) = \\ &= (X^2 - (\alpha + \alpha^2)X + \alpha^3)(X^2 - (\alpha^3 + \alpha^4)X + \alpha^7) = \\ &= (X^2 - \alpha^4X + \alpha^3)(X^2 - \alpha^6X + \alpha^0) = \\ &= X^4 - (\alpha^4 + \alpha^6)X^3 + (\alpha^3 + \alpha^{10} + \alpha^0)X^2 - (\alpha^4 + \alpha^9)X + \alpha^3 = \\ &= X^4 - \alpha^3X^3 + \alpha^0X^2 - \alpha^1X + \alpha^3\end{aligned}$$

Поменяв порядок расположения членов полинома на обратный и заменив знаки "минус" на "плюс", так как над двоичным полем $+1 = -1$, генератор $\mathbf{g}(X)$ можно будет представить следующим образом.

$$\mathbf{g}(X) = \alpha^3 + \alpha^1X + \alpha^0X^2 + \alpha^3X^3 + X^4 \tag{8.22}$$

8.1.5.1. Кодирование в систематической форме

Так как код Рида-Соломона является циклическим, кодирование в систематической форме аналогично процедуре двоичного кодирования, разработанной в разделе 6.7.3. Мы можем осуществить сдвиг полинома сообщения $\mathbf{m}(X)$ в крайние правые k разряды регистра кодового слова и произвести последующее прибавление полинома четности $\mathbf{p}(X)$ в крайние левые $n - k$ разряды. Поэтому мы умножаем $\mathbf{m}(X)$ на X^{n-k}, проделав алгебраическую операцию таким образом, что $\mathbf{m}(X)$ оказывается сдвинутым вправо на $n - k$ позиций. В главе 6 это показано в уравнении (6.61) на примере двоичного кодирования. Далее мы делим $X^{n-k}\mathbf{m}(X)$ на полиномиальный генератор $\mathbf{g}(X)$, что можно записать следующим образом.

$$X^{n-k}\mathbf{m}(X) = \mathbf{q}(X)\mathbf{g}(X) + \mathbf{p}(X) \tag{8.23}$$

Здесь $\mathbf{q}(X)$ и $\mathbf{p}(X)$ — это частное и остаток от полиномиального деления. Как и в случае двоичного кодирования, остаток будет четным. Уравнение (8.23) можно переписать следующим образом.

$$\mathbf{p}(X) = X^{n-k}\mathbf{m}(X) \text{ по модулю } \mathbf{g}(X) \tag{8.24}$$

Результирующий полином кодового слова $\mathbf{U}(X)$, показанный в уравнении (6.64), можно переписать следующим образом.

$$\mathbf{U}(X) = \mathbf{p}(X) + X^{n-k}\mathbf{m}(X) \tag{8.25}$$

Продемонстрируем шаги, подразумеваемые уравнениями (8.24) и (8.25), закодировав сообщение из трех символов

$$\underbrace{010}_{\alpha^1}\ \underbrace{110}_{\alpha^3}\ \underbrace{111}_{\alpha^5}$$

с помощью кода (7, 3), генератор которого определяется уравнением (8.22). Сначала мы умножаем (сдвиг вверх) полином сообщения $\alpha^1 + \alpha^3 X + \alpha^5 X^2$ на $X^{n-k} = X^4$, что дает $\alpha^1 X + \alpha^3 X^5 + \alpha^5 X^6$. Далее мы делим такой сдвинутый вверх полином сообщения на полиномиальный генератор из уравнения (8.22), $\alpha^3 + \alpha^1 X + \alpha^0 X^2 + \alpha^3 X^3 + X^4$. Полиномиальное деление недвоичных коэффициентов — это еще более утомительная процедура, чем ее двоичный аналог (см. пример 6.9), поскольку операции сложения (вычитания) и умножения (деления) выполняются согласно табл. 8.2 и 8.3. Мы оставим читателю в качестве самостоятельного упражнения проверку того, что полиномиальное деление даст в результате следующий полиномиальный остаток (полином четности).

$$\mathbf{p}(X) = \alpha^0 + \alpha^2 X + \alpha^4 X^2 + \alpha^6 X^3$$

Затем, из уравнения (8.25), полином кодового слова можно записать следующим образом.

$$\mathbf{U}(X) = \alpha^0 + \alpha^2 X + \alpha^4 X^2 + \alpha^6 X^3 + \alpha^1 X^4 + \alpha^3 X^5 + \alpha^5 X^6$$

8.1.5.2. Систематическое кодирование с помощью $(n-k)$-разрядного регистра сдвига

Как показано на рис. 8.9, кодирование последовательности из 3 символов в систематической форме на основе кода (7, 3), определяемого генератором $\mathbf{g}(X)$ из уравнения (8.22), требует реализации регистра LFSR. Нетрудно убедиться, что элементы умножителя на рис. 8.9, взятые справа налево, соответствуют коэффициентам полинома в уравнении (8.22). Этот процесс кодирования является недвоичным аналогом циклического кодирования, которое описывалось в разделе 6.7.5. Здесь, в соответствии с уравнением (8.20), ненулевые кодовые слова образованы $2^m - 1 = 7$ символами, и каждый символ состоит из $m = 3$ бит.

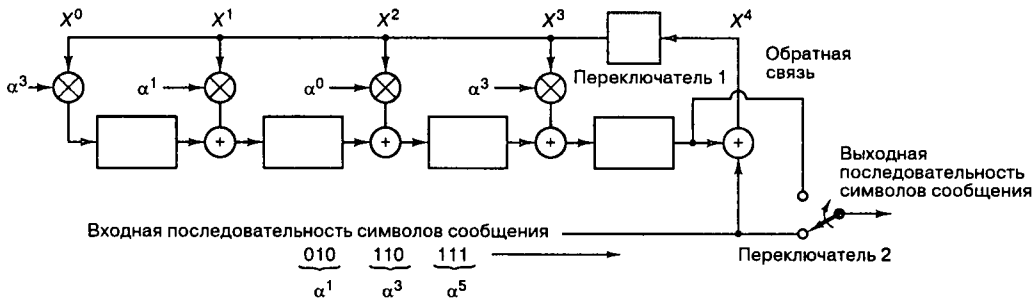


Рис. 8.9. Кодер LFSR для кода (7, 3)

Следует отметить сходство между рис. 8.9, 6.18 и 6.19. Во всех трех случаях количество разрядов в регистре равно $n - k$. Рисунки в главе 6 отображают пример двоичного кодирования, где каждый разряд содержит 1 бит. В данной главе приведен пример недвоичного кодирования, так что каждый разряд регистра сдвига, изображенного на рис. 8.9, содержит 3-битовый символ. На рис. 6.18 коэффициенты, обозначенные g_1, g_2, ..., являются двоичными. Поэтому они принимают одно из значений 0 или 1, просто указывая на наличие или отсутствие связи в LFSR. На рис. 8.9 каждый коэффициент является 3-битовым, так что они могут принимать одно из 8 значений.

Недвоичные операции, осуществляемые кодером, показанным на рис. 8.9, создают кодовые слова в систематической форме, так же как и в двоичном случае. Эти операции определяются следующими шагами.

1. Переключатель 1 в течение первых k тактовых импульсов закрыт, для того чтобы подавать символы сообщения в $(n - k)$-разрядный регистр сдвига.
2. В течение первых k тактовых импульсов переключатель 2 находится в нижнем положении, что обеспечивает одновременную передачу всех символов сообщения непосредственно на регистр выхода (на рис. 8.9 не показан).
3. После передачи k-го символа на регистр выхода, переключатель 1 открывается, а переключатель 2 переходит в верхнее положение.
4. Остальные $(n - k)$ тактовых импульсов очищают контрольные символы, содержащиеся в регистре, подавая их на регистр выхода.
5. Общее число тактовых импульсов равно n, и содержимое регистра выхода является полиномом кодового слова $\mathbf{p}(X) + X^{n-k}\mathbf{m}(X)$, где $\mathbf{p}(X)$ представляет собой кодовые символы, а $\mathbf{m}(X)$ — символы сообщения в полиномиальной форме.

Для проверки возьмем ту же последовательность символов, что и в разделе 8.1.5.1.

$$\underbrace{010}_{\alpha^1}\ \underbrace{110}_{\alpha^3}\ \underbrace{111}_{\alpha^5}$$

Здесь крайний правый символ является самым первым и крайний правый бит также является самым первым. Последовательность действий в течение первых $k = 3$ сдвигов в цепи кодирования на рис. 8.9 будет иметь следующий вид.

ОЧЕРЕДЬ ВВОДА			ТАКТ	СОДЕРЖИМОЕ РЕГИСТРА				ОБРАТНАЯ СВЯЗЬ
α^1	α^3	α^5	0	0	0	0	0	α^5
	α^1	α^3	1	α^1	α^6	α^5	α^1	α^0
		α^1	2	α^3	0	α^2	α^2	α^4
		—	3	α^0	α^2	α^4	α^6	—

Как можно видеть, после третьего такта регистр содержит 4 контрольных символа, α^0, α^2, α^4 и α^6. Затем переключатель 1 переходит в верхнее положение, и контрольные символы, содержащиеся в регистре, подаются на выход. Поэтому выходное кодовое слово, записанное в полиномиальной форме, можно представить в следующем виде.

$$\mathbf{U}(X) = \sum_{n=0}^{6} u_n X^n$$

$$\mathbf{U}(X) = \alpha^0 + \alpha^2 X + \alpha^4 X^2 + \alpha^6 X^3 + \alpha^1 X^4 + \alpha^3 X^5 + \alpha^5 X^6 = \tag{8.26}$$

$$= (100) + (001)X + (011)X^2 + (101)X^3 + (010)X^4 + (110)X^5 + (111)X^6$$

Процесс проверки содержимого регистра во время разных тактов несколько сложнее, чем в случае бинарного кодирования. Здесь сложение и умножение элементов поля должны выполняться согласно табл. 8.2 и 8.3.

Корни полиномиального генератора $\mathbf{g}(X)$ должны быть и корнями кодового слова, генерируемого $\mathbf{g}(X)$, поскольку правильное кодовое слово имеет следующий вид.

$$\mathbf{U}(X) = \mathbf{m}(X)\mathbf{g}(X) \tag{8.27}$$

Следовательно, произвольное кодовое слово, выражаемое через корень генератора $\mathbf{g}(X)$, должно давать нуль. Представляется интересным, действительно ли полином кодового слова в уравнении (8.26) дает нуль, когда он выражается через какой-либо из четырех корней $\mathbf{g}(X)$. Иными словами, это означает проверку следующего.

$$\mathbf{U}(\alpha) = \mathbf{U}(\alpha^2) = \mathbf{U}(\alpha^3) = \mathbf{U}(\alpha^4) = \mathbf{0}$$

Независимо выполнив вычисления для разных корней, получим следующее.

$$\begin{aligned}\mathbf{U}(\alpha) &= \alpha^0 + \alpha^3 + \alpha^6 + \alpha^9 + \alpha^5 + \alpha^8 + \alpha^{11} = \\ &= \alpha^0 + \alpha^3 + \alpha^6 + \alpha^2 + \alpha^5 + \alpha^1 + \alpha^4 = \\ &= \alpha^1 + \alpha^0 + \alpha^6 + \alpha^4 = \\ &= \alpha^3 + \alpha^3 = \mathbf{0}\end{aligned}$$

$$\begin{aligned}\mathbf{U}(\alpha^2) &= \alpha^0 + \alpha^4 + \alpha^8 + \alpha^{12} + \alpha^9 + \alpha^{13} + \alpha^{17} = \\ &= \alpha^0 + \alpha^4 + \alpha^1 + \alpha^5 + \alpha^2 + \alpha^6 + \alpha^3 = \\ &= \alpha^5 + \alpha^6 + \alpha^0 + \alpha^3 = \\ &= \alpha^1 + \alpha^1 = \mathbf{0}\end{aligned}$$

$$\begin{aligned}\mathbf{U}(\alpha^3) &= \alpha^0 + \alpha^5 + \alpha^{10} + \alpha^{15} + \alpha^{13} + \alpha^{18} + \alpha^{23} = \\ &= \alpha^0 + \alpha^5 + \alpha^3 + \alpha^1 + \alpha^6 + \alpha^4 + \alpha^2 = \\ &= \alpha^4 + \alpha^0 + \alpha^3 + \alpha^2 = \\ &= \alpha^5 + \alpha^5 = \mathbf{0}\end{aligned}$$

$$\begin{aligned}\mathbf{U}(\alpha^4) &= \alpha^0 + \alpha^6 + \alpha^{12} + \alpha^{18} + \alpha^{17} + \alpha^{23} + \alpha^{29} = \\ &= \alpha^0 + \alpha^6 + \alpha^5 + \alpha^4 + \alpha^3 + \alpha^2 + \alpha^1 = \\ &= \alpha^2 + \alpha^0 + \alpha^5 + \alpha^1 = \\ &= \alpha^6 + \alpha^6 = \mathbf{0}\end{aligned}$$

Эти вычисления показывают, что, как и ожидалось, кодовое слово, выражаемое через любой корень генератора $\mathbf{g}(X)$, должно давать нуль.

8.1.6. Декодирование Рида-Соломона

В разделе 8.1.5 тестовое сообщение кодируется в систематической форме с помощью кода (7, 3), что дает в результате полином кодового слова, описываемый уравнением (8.26). Допустим, что в ходе передачи это кодовое слово подверглось искажению: 2 символа были приняты с ошибкой. (Такое количество ошибок соответствует макси-

мальной способности кода к коррекции ошибок.) При использовании 7-символьного кодового слова ошибочную комбинацию можно представить в полиномиальной форме следующим образом.

$$\mathbf{e}(X) = \sum_{n=0}^{6} e_n X^n \tag{8.28}$$

Пусть двухсимвольная ошибка будет такой, что

$$\begin{aligned}\mathbf{e}(X) &= 0 + 0X + 0X^2 + \alpha^2X^3 + \alpha^5X^4 + 0X^5 + 0X^6 = \\ &= (000) + (000)X + (000)X^2 + (001)X^3 + (111)X^4 + (000)X^5 + (000)X^6.\end{aligned} \tag{8.29}$$

Другими словами, контрольный символ искажен 1-битовой ошибкой (представленной как α^2), а символ сообщения — 3-битовой ошибкой (представленной как α^5). В данном случае принятый полином поврежденного кодового слова $\mathbf{r}(X)$ представляется в виде суммы полинома переданного кодового слова и полинома ошибочной комбинации, как показано ниже.

$$\mathbf{r}(X) = \mathbf{U}(X) + \mathbf{e}(X) \tag{8.30}$$

Следуя уравнению (8.30), мы суммируем $\mathbf{U}(X)$ из уравнения (8.26) и $\mathbf{e}(X)$ из уравнения (8.29) и имеем следующее.

$$\begin{aligned}\mathbf{r}(X) &= (100) + (001)X + (011)X^2 + (100)X^3 + (101)X^4 + (110)X^5 + (111)X^6 = \\ &= \alpha^0 + \alpha^2X + \alpha^4X^2 + \alpha^0X^3 + \alpha^6X^4 + \alpha^3X^5 + \alpha^5X^6\end{aligned} \tag{8.31}$$

В данном примере исправления 2-символьной ошибки имеется четыре неизвестных — два относятся к расположению ошибки, а два касаются ошибочных значений. Отметим важное различие между недвоичным декодированием $\mathbf{r}(X)$, которое мы показали в уравнении (8.31), и двоичным, которое описывалось в главе 6. При двоичном декодировании декодеру нужно знать лишь расположение ошибки. Если известно, где находится ошибка, бит нужно поменять с 1 на 0 или наоборот. Но здесь недвоичные символы требуют, чтобы мы не только узнали расположение ошибки, но и определили правильное значение символа, расположенного на этой позиции. Поскольку в данном примере у нас имеется четыре неизвестных, нам нужно четыре уравнения, чтобы найти их.

8.1.6.1. Вычисление синдрома

Вернемся к разделу 6.4.7 и напомним, что *синдром* — это результат проверки четности, выполняемой над **r**, чтобы определить, принадлежит ли **r** набору кодовых слов. Если **r** является членом набора, то синдром **S** имеет значение, равное **0**. Любое ненулевое значение **S** означает наличие ошибок. Точно так же, как и в двоичном случае, синдром **S** состоит из $n - k$ символов, $\{S_i\}$ ($i = 1, \ldots, n - k$). Таким образом, для нашего кода (7, 3) имеется по четыре символа в каждом векторе синдрома; их значения можно рассчитать из принятого полинома $\mathbf{r}(X)$. Заметим, как облегчаются вычисления благодаря самой структуре кода, определяемой уравнением (8.27).

$$\mathbf{U}(X) = \mathbf{m}(X)\mathbf{g}(X)$$

Из этой структуры можно видеть, что каждый правильный полином кодового слова $\mathbf{U}(X)$ является кратным полиномиальному генератору $\mathbf{g}(X)$. Следовательно, корни $\mathbf{g}(X)$ также должны быть корнями $\mathbf{U}(X)$. Поскольку $\mathbf{r}(X) = \mathbf{U}(X) + \mathbf{e}(X)$, то $\mathbf{r}(X)$, вычисляемый с каждым корнем $\mathbf{g}(X)$, должен давать нуль, только если $\mathbf{r}(X)$ будет правильным кодовым словом. Любые ошибки приведут в итоге к ненулевому результату в одном (или более) случае. Вычисления символов синдрома можно записать следующим образом.

$$S_i = \mathbf{r}(X)\Big|_{X=\alpha^i} = \mathbf{r}(\alpha^i) \quad i = 1, \ldots, n-k \tag{8.32}$$

Здесь, как было показано в уравнении (8.29), $\mathbf{r}(X)$ содержит 2-символьные ошибки. Если $\mathbf{r}(X)$ окажется правильным кодовым словом, то это приведет к тому, что все символы синдрома S_i будут равны нулю. В данном примере четыре символа синдрома находятся следующим образом.

$$\begin{aligned} S_1 = \mathbf{r}(\alpha) &= \alpha^0 + \alpha^3 + \alpha^6 + \alpha^3 + \alpha^{10} + \alpha^8 + \alpha^{11} = \\ &= \alpha^0 + \alpha^3 + \alpha^6 + \alpha^3 + \alpha^2 + \alpha^1 + \alpha^4 = \\ &= \alpha^3 \end{aligned} \tag{8.33}$$

$$\begin{aligned} S_2 = \mathbf{r}(\alpha^2) &= \alpha^0 + \alpha^4 + \alpha^8 + \alpha^6 + \alpha^{14} + \alpha^{13} + \alpha^{17} = \\ &= \alpha^0 + \alpha^4 + \alpha^1 + \alpha^6 + \alpha^0 + \alpha^6 + \alpha^3 = \\ &= \alpha^5 \end{aligned} \tag{8.34}$$

$$\begin{aligned} S_3 = \mathbf{r}(\alpha^3) &= \alpha^0 + \alpha^5 + \alpha^{10} + \alpha^9 + \alpha^{18} + \alpha^{18} + \alpha^{23} = \\ &= \alpha^0 + \alpha^5 + \alpha^3 + \alpha^2 + \alpha^4 + \alpha^4 + \alpha^2 = \\ &= \alpha^6 \end{aligned} \tag{8.35}$$

$$\begin{aligned} S_4 = \mathbf{r}(\alpha^4) &= \alpha^0 + \alpha^6 + \alpha^{12} + \alpha^{12} + \alpha^{22} + \alpha^{23} + \alpha^{29} = \\ &= \alpha^0 + \alpha^6 + \alpha^5 + \alpha^5 + \alpha^1 + \alpha^2 + \alpha^1 = \\ &= 0 \end{aligned} \tag{8.36}$$

Результат подтверждает, что принятое кодовое слово содержит ошибку (введенную нами), поскольку $\mathbf{S} \neq \mathbf{0}$.

Пример 8.3. Повторная проверка значений синдрома

Для рассматриваемого кода (7, 3) ошибочная комбинация известна, поскольку мы выбрали ее заранее. Вспомним свойство кодов, обсуждаемое в разделе 6.4.8.1, когда была введена нормальная матрица. Все элементы класса смежности (строка) нормальной матрицы имеют один и тот же синдром. Нужно показать, что это свойство справедливо и для кода Рида-Соломона, путем вычисления полинома ошибок $\mathbf{e}(X)$ со значениями корней $\mathbf{g}(X)$. Это должно дать те же значения синдрома, что и вычисление $\mathbf{r}(X)$ со значениями корней $\mathbf{g}(X)$. Другими словами, это должно дать те же значения, которые были получены в уравнениях (8.33)–(8.36).

Решение

$$S_i = \mathbf{r}(X)\Big|_{X=\alpha^i} = \mathbf{r}(\alpha^i) \quad i = 1, 2, \ldots, n-k$$

$$S_i = \left[\mathbf{U}(X) + \mathbf{e}(X)\right]\Big|_{X=\alpha^i} = \mathbf{U}(\alpha^i) + \mathbf{e}(\alpha^i)$$

$$S_i = \mathbf{r}(\alpha^i) = \mathbf{U}(\alpha^i) + \mathbf{e}(\alpha^i) = 0 + \mathbf{e}(\alpha^i)$$

Из уравнения (8.29) следует, что $\mathbf{e}(X) = \alpha^2 X^3 + \alpha^5 X^4$, поэтому

$$S_1 = \mathbf{e}(\alpha^1) = \alpha^5 + \alpha^9 =$$
$$= \alpha^5 + \alpha^2 =$$
$$= \alpha^3$$

$$S_2 = \mathbf{e}(\alpha^2) = \alpha^8 + \alpha^{13} =$$
$$= \alpha^1 + \alpha^6 =$$
$$= \alpha^5$$

$$S_3 = \mathbf{e}(\alpha^3) = \alpha^{11} + \alpha^{17} =$$
$$= \alpha^4 + \alpha^3 =$$
$$= \alpha^6$$

$$S_4 = \mathbf{e}(\alpha^4) = \alpha^{14} + \alpha^{21} =$$
$$= \alpha^0 + \alpha^0 =$$
$$= 0$$

Из этих результатов можно заключить, что значения синдрома одинаковы — как полученные путем вычисления $\mathbf{e}(X)$ со значениями корней $\mathbf{g}(X)$, так и полученные путем вычисления $\mathbf{r}(X)$ с теми же значениями корней $\mathbf{g}(X)$.

8.1.6.2. Локализация ошибки

Допустим, в кодовом слове имеется v ошибок, расположенных на позициях $X^{j_1}, X^{j_2}, \ldots, X^{j_v}$. Тогда полином ошибок, определяемый уравнениями (8.28) и (8.29), можно записать следующим образом.

$$\mathbf{e}(X) = e_{j_1} X^{j_1} + e_{j_2} X^{j_2} + \ldots + e_{j_v} X^{j_v} \tag{8.37}$$

Индексы 1, 2, ..., v обозначают 1-ю, 2-ю, ..., v-ю ошибки, а индекс j — расположение ошибки. Для коррекции искаженного кодового слова нужно определить каждое значение ошибки e_{j_l} и ее расположение X^{j_l}, где $l = 1, 2, \ldots, v$. Обозначим номер ло-

катора ошибки как $\beta_l = \alpha^{j_l}$. Далее вычисляем $n-k=2t$ символа синдрома, подставляя α_i в принятый полином при $i=1, 2, \ldots, 2t$.

$$\begin{aligned} S_1 &= \mathbf{r}(\alpha) = e_{j_1}\beta_1 + e_{j_2}\beta_2 + \cdots + e_{j_v}\beta_v \\ S_2 &= \mathbf{r}(\alpha^2) = e_{j_1}\beta_1^2 + e_{j_2}\beta_2^2 + \cdots + e_{j_v}\beta_v^2 \\ &\vdots \\ S_{2t} &= \mathbf{r}(\alpha^{2t}) = e_{j_1}\beta_1^{2t} + e_{j_2}\beta_2^{2t} + \cdots + e_{j_v}\beta_v^{2t} \end{aligned} \tag{8.38}$$

У нас имеется $2t$ неизвестных (t значений ошибок и t расположений) и система $2t$ уравнений. Впрочем, эту систему $2t$ уравнений нельзя решить обычным путем, поскольку уравнения в ней нелинейны (некоторые неизвестные входят в уравнение в степени). Методика, позволяющая решить эту систему уравнений, называется алгоритмом декодирования Рида-Соломона.

Если вычислен ненулевой вектор синдрома (один или более его символов не равны нулю), это означает, что была принята ошибка. Далее нужно узнать расположение ошибки (или ошибок). Полином локатора ошибок можно определить следующим образом.

$$\begin{aligned} \sigma(X) &= (1+\beta_1 X)(1+\beta_2 X)\ldots(1+\beta_v X) = \\ &= 1+\sigma_1 X+\sigma_2 X^2+\ldots+\sigma_v X^v \end{aligned} \tag{8.39}$$

Корнями $\sigma(X)$ будут $1/\beta_1$, $1/\beta_2$, ..., $1/\beta_v$. Величины, обратные корням $\sigma(X)$, будут представлять номера расположений ошибочной комбинации $\mathbf{e}(X)$. Тогда, воспользовавшись авторегрессионной техникой моделирования [5], мы составим из синдромов матрицу, в которой первые t синдромов будут использоваться для предсказания следующего синдрома.

$$\begin{bmatrix} S_1 & S_2 & S_3 & \cdots & S_{t-1} & S_t \\ S_2 & S_3 & S_4 & \cdots & S_t & S_{t+1} \\ & & \vdots & & & \\ S_{t-1} & S_t & S_{t+1} & \cdots & S_{2t-3} & S_{2t-2} \\ S_t & S_{t+1} & S_{t+2} & \cdots & S_{2t-2} & S_{2t-1} \end{bmatrix} \begin{bmatrix} \sigma_t \\ \sigma_{t-1} \\ \vdots \\ \sigma_2 \\ \sigma_1 \end{bmatrix} = \begin{bmatrix} -S_{t+1} \\ -S_{t+2} \\ \vdots \\ -S_{2t-1} \\ -S_{2t} \end{bmatrix} \tag{8.40}$$

Мы воспользовались авторегрессионной моделью уравнения (8.40), взяв матрицу наибольшей размерности с ненулевым определителем. Для кода (7, 3) с коррекцией двухсимвольных ошибок матрица будет иметь размерность 2×2, и модель запишется следующим образом.

$$\begin{bmatrix} S_1 & S_2 \\ S_2 & S_3 \end{bmatrix} \begin{bmatrix} \sigma_2 \\ \sigma_1 \end{bmatrix} = \begin{bmatrix} S_3 \\ S_4 \end{bmatrix} \tag{8.41}$$

$$\begin{bmatrix} \alpha^3 & \alpha^5 \\ \alpha^5 & \alpha^6 \end{bmatrix} \begin{bmatrix} \sigma_2 \\ \sigma_1 \end{bmatrix} = \begin{bmatrix} \alpha^6 \\ 0 \end{bmatrix} \tag{8.42}$$

Чтобы найти коэффициенты σ_1 и σ_2 полинома локатора ошибок $\sigma(X)$, сначала необходимо вычислить обратную матрицу для уравнения (8.42). Обратная матрица для матрицы $[A]$ определяется следующим образом.

$$\text{Inv}\,[A] = \frac{\text{cofactor}\,[A]}{\det\,[A]}$$

Следовательно,

$$\det\begin{bmatrix}\alpha^3 & \alpha^5\\ \alpha^5 & \alpha^6\end{bmatrix} = \alpha^3\alpha^6 - \alpha^5\alpha^5 = \alpha^9 + \alpha^{10} = \\ = \alpha^2 + \alpha^3 = \alpha^5 \tag{8.43}$$

$$\text{cofactor}\begin{bmatrix}\alpha^3 & \alpha^5\\ \alpha^5 & \alpha^6\end{bmatrix} = \begin{bmatrix}\alpha^6 & \alpha^5\\ \alpha^5 & \alpha^3\end{bmatrix} \tag{8.44}$$

и

$$\text{Inv}\begin{bmatrix}\alpha^3 & \alpha^5\\ \alpha^5 & \alpha^6\end{bmatrix} = \frac{\begin{bmatrix}\alpha^6 & \alpha^5\\ \alpha^5 & \alpha^3\end{bmatrix}}{\alpha^5} = \alpha^{-5}\begin{bmatrix}\alpha^6 & \alpha^5\\ \alpha^5 & \alpha^3\end{bmatrix} = \\ = \alpha^2\begin{bmatrix}\alpha^6 & \alpha^5\\ \alpha^5 & \alpha^3\end{bmatrix} = \begin{bmatrix}\alpha^8 & \alpha^7\\ \alpha^7 & \alpha^5\end{bmatrix} = \begin{bmatrix}\alpha^1 & \alpha^0\\ \alpha^0 & \alpha^5\end{bmatrix} \tag{8.45}$$

Проверка надежности

Если обратная матрица вычислена правильно, то произведение исходной и обратной матрицы должно дать единичную матрицу.

$$\begin{bmatrix}\alpha^3 & \alpha^5\\ \alpha^5 & \alpha^6\end{bmatrix}\begin{bmatrix}\alpha^1 & \alpha^0\\ \alpha^0 & \alpha^5\end{bmatrix} = \begin{bmatrix}\alpha^4 + \alpha^5 & \alpha^3 + \alpha^{10}\\ \alpha^6 + \alpha^6 & \alpha^5 + \alpha^{11}\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \tag{8.46}$$

С помощью уравнения (8.42) начнем поиск положений ошибок с вычисления коэффициентов полинома локатора ошибок $\sigma(X)$, как показано далее.

$$\begin{bmatrix}\sigma_2\\ \sigma_1\end{bmatrix} = \begin{bmatrix}\alpha^1 & \alpha^0\\ \alpha^0 & \alpha^5\end{bmatrix}\begin{bmatrix}\alpha^6\\ 0\end{bmatrix} = \begin{bmatrix}\alpha^7\\ \alpha^6\end{bmatrix} = \begin{bmatrix}\alpha^0\\ \alpha^6\end{bmatrix} \tag{8.47}$$

Из уравнений (8.39) и (8.47)

$$\sigma(X) = \alpha^0 + \sigma_1 X + \sigma_2 X^2 = \\ = \alpha^0 + \alpha^6 X + \alpha^0 X^2 \tag{8.48}$$

Корни $\sigma(X)$ являются обратными числами к положениям ошибок. После того как эти корни найдены, мы знаем расположение ошибок. Вообще, корни $\sigma(X)$ могут быть одним или несколькими элементами поля. Определим эти корни путем полной проверки полинома $\sigma(X)$ со всеми элементами поля, как будет показано ниже. Любой элемент X, который дает $\sigma(X) = \mathbf{0}$, является корнем, что позволяет нам определить расположение ошибки.

$$\sigma(\alpha^0) = \alpha^0 + \alpha^6 + \alpha^0 = \alpha^6 \neq \mathbf{0}$$
$$\sigma(\alpha^1) = \alpha^2 + \alpha^7 + \alpha^0 = \alpha^2 \neq \mathbf{0}$$
$$\sigma(\alpha^2) = \alpha^4 + \alpha^8 + \alpha^0 = \alpha^6 \neq \mathbf{0}$$
$$\sigma(\alpha^3) = \alpha^6 + \alpha^9 + \alpha^0 = \mathbf{0} \Rightarrow \text{ОШИБКА}$$

$$\sigma(\alpha^4) = \alpha^8 + \alpha^{10} + \alpha^0 = \mathbf{0} \Rightarrow \text{ОШИБКА}$$
$$\sigma(\alpha^5) = \alpha^{10} + \alpha^{11} + \alpha^0 = \alpha^2 \neq \mathbf{0}$$
$$\sigma(\alpha^6) = \alpha^{12} + \alpha^{12} + \alpha^0 = \alpha^0 \neq \mathbf{0}$$

Как видно из уравнения (8.39), расположение ошибок является обратной величиной к корням полинома. А значит, $\sigma(\alpha^3) = \mathbf{0}$ означает, что один корень получается при $1/\beta_l = \alpha^3$. Отсюда $\beta_l = 1/\alpha^3 = \alpha^4$. Аналогично $\sigma(\alpha^4) = \mathbf{0}$ означает, что другой корень появляется при $1/\beta_{l'} = 1/\alpha^4 = \alpha^3$, где (в данном примере) l и l' обозначают 1-ю и 2-ю ошибки. Поскольку мы имеем дело с 2-символьными ошибками, полином ошибок можно записать следующим образом.

$$\mathbf{e}(X) = e_{j_1} X^{j_1} + e_{j_2} X^{j_2} \tag{8.49}$$

Здесь были найдены две ошибки на позициях α^3 и α^4. Заметим, что индексация номеров расположения ошибок является сугубо произвольной. Итак, в этом примере мы обозначили величины $\beta_l = \alpha^{j_l}$ как $\beta_1 = \alpha^{j_1} = \alpha^3$ и $\beta_2 = \alpha^{j_2} = \alpha^4$.

8.1.6.3. Значения ошибок

Мы обозначили ошибки e_{j_l}, где индекс j обозначает расположение ошибки, а индекс l — l-ю ошибку. Поскольку каждое значение ошибки связано с конкретным месторасположением, систему обозначений можно упростить, обозначив e_{j_l} просто как e_l. Теперь, приготовившись к нахождению значений ошибок e_1 и e_2, связанных с позициями $\beta_1 = \alpha^3$ и $\beta_2 = \alpha^4$, можно использовать любое из четырех синдромных уравнений. Выразим из уравнения (8.38) S_1 и S_2.

$$\begin{aligned} S_1 &= \mathbf{r}(\alpha) = e_1\beta_1 + e_2\beta_2 \\ S_2 &= \mathbf{r}(\alpha^2) = e_1\beta_1^2 + e_2\beta_2^2 \end{aligned} \tag{8.50}$$

Эти уравнения можно переписать в матричной форме следующим образом.

$$\begin{bmatrix} \beta_1 & \beta_2 \\ \beta_1^2 & \beta_2^2 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \tag{8.51}$$

$$\begin{bmatrix} \alpha^3 & \alpha^4 \\ \alpha^6 & \alpha^8 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} \alpha^3 \\ \alpha^5 \end{bmatrix} \tag{8.52}$$

Чтобы найти значения ошибок e_1 и e_2, нужно, как обычно, выполнить поиск обратной матрицы для уравнения (8.52).

$$\begin{aligned} \text{Inv}\begin{bmatrix} \alpha^3 & \alpha^4 \\ \alpha^6 & \alpha^1 \end{bmatrix} &= \frac{\begin{bmatrix} \alpha^1 & \alpha^4 \\ \alpha^6 & \alpha^3 \end{bmatrix}}{\alpha^3\alpha^1 - \alpha^6\alpha^4} = \\ &= \frac{\begin{bmatrix} \alpha^1 & \alpha^4 \\ \alpha^6 & \alpha^3 \end{bmatrix}}{\alpha^4 + \alpha^3} = \alpha^{-6}\begin{bmatrix} \alpha^1 & \alpha^4 \\ \alpha^6 & \alpha^3 \end{bmatrix} = \alpha^1\begin{bmatrix} \alpha^1 & \alpha^4 \\ \alpha^6 & \alpha^3 \end{bmatrix} = \end{aligned} \tag{8.53}$$

$$= \begin{bmatrix} \alpha^2 & \alpha^5 \\ \alpha^7 & \alpha^4 \end{bmatrix} = \begin{bmatrix} \alpha^2 & \alpha^5 \\ \alpha^0 & \alpha^4 \end{bmatrix}$$

Теперь мы можем найти из уравнения (8.52) значения ошибок.

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} \alpha^2 & \alpha^5 \\ \alpha^0 & \alpha^4 \end{bmatrix} \begin{bmatrix} \alpha^3 \\ \alpha^5 \end{bmatrix} = \begin{bmatrix} \alpha^5 + \alpha^{10} \\ \alpha^3 + \alpha^9 \end{bmatrix} = \begin{bmatrix} \alpha^5 + \alpha^3 \\ \alpha^3 + \alpha^2 \end{bmatrix} = \begin{bmatrix} \alpha^2 \\ \alpha^5 \end{bmatrix} \tag{8.54}$$

8.1.6.4. Исправление принятого полинома с помощью найденного полинома ошибок

Из уравнений (8.49) и (8.54) мы находим полином ошибок.

$$\begin{aligned} \hat{\mathbf{e}}(X) &= e_1 X^{j_1} + e_2 X^{j_2} = \\ &= \alpha^2 X^3 + \alpha^5 X^4 \end{aligned} \tag{8.55}$$

Показанный алгоритм восстанавливает принятый полином, выдавая в итоге предполагаемое переданное кодовое слово и, в конечном счете, декодированное сообщение.

$$\hat{\mathbf{U}}(X) = \mathbf{r}(X) + \hat{\mathbf{e}}(X) = \mathbf{U}(X) + \mathbf{e}(X) + \hat{\mathbf{e}}(X) \tag{8.56}$$

$$\begin{aligned} \mathbf{r}(X) &= (100) + (001)X + (011)X^2 + (100)X^3 + (101)X^4 + (110)X^5 + (111)X^6 \\ \hat{\mathbf{e}}(X) &= (000) + (000)X + (000)X^2 + (001)X^3 + (111)X^4 + (000)X^5 + (000)X^6 \\ \hat{\mathbf{U}}(X) &= (100) + (001)X + (011)X^2 + (101)X^3 + (010)X^4 + (110)X^5 + (111)X^6 = \\ &= \alpha^0 + \alpha^2 X + \alpha^4 X^2 + \alpha^6 X^3 + \alpha^1 X^4 + \alpha^3 X^5 + \alpha^5 X^6 \end{aligned} \tag{8.57}$$

Поскольку символы сообщения содержатся в крайних правых $k = 3$ символах, декодированным будет следующее сообщение.

$$\underbrace{010}_{\alpha^1} \quad \underbrace{110}_{\alpha^3} \quad \underbrace{111}_{\alpha^5}$$

Это сообщение в точности соответствует тому, которое было выбрано для этого примера в разделе 8.1.5. (Для более детального знакомства с кодированием Рида-Соломона обратитесь к работе [6].)

8.2. Коды с чередованием и каскадные коды

В предыдущих главах мы подразумевали, что у канала *отсутствует память*, поскольку рассматривались коды, которые должны были противостоять случайным независимым ошибкам. Канал *с памятью* — это такой канал, в котором проявляется взаимная зависимость ухудшений передачи сигнала. Канал, в котором проявляется *замирание вследствие многолучевого распространения*, когда сигнал поступает на приемник по двум или более путям различной длины, есть примером канала с памятью. Следствием является различная фаза сигналов, и в итоге, суммарный сигнал оказывается искаженным. Таким эффектом обладают каналы мобильной беспроводной связи, так же как ионосферные и тропосферные каналы. (Более подробно о замирании см. главу 15.) В некоторых каналах также имеются коммутационные и другие импульсные помехи (например, телефонные каналы или каналы с создаваемыми импульсными помехами). Все эти ухудшения коррелируют во времени и, в результате, дают статистическую взаимную зависимость успешно переданных символов. Иными словами, искажения вызывают ошибки, имеющие вид пакетов, а не отдельных изолированных ошибок.

Если канал имеет память, то ошибки не являются независимыми, одиночными и случайно распределенными. Большинство блочных и сверточных кодов разрабатывается для борьбы с независимыми случайными ошибками. Влияние канала с памятью на кодированный таким образом сигнал приведет к ухудшению достоверности передачи. Существуют схемы кодирования для каналов с памятью, но наибольшую проблему в этом кодировании представляет расчет точных моделей сильно нестационарных статистик таких каналов. Один подход, при котором требуется знать только *объем памяти* канала, а не его точное статистическое описание, использует временное разнесение, или *чередование битов*.

Чередование битов кодированного сообщения перед передачей и обратная операция после приема приводят к рассеиванию пакета ошибок во времени: таким образом, они становятся для декодера случайно распределенными. Поскольку в реальной ситуации память канала уменьшается с временным разделением, идея, лежащая в основе метода чередования битов, заключается в разнесении символов кодовых слов во времени. Получаемые промежутки времени точно так же заполняются символами других кодовых слов. Разнесение символов во времени эффективно превращает канал с памятью в *канал без памяти* и, следовательно, позволяет использовать коды с коррекцией случайных ошибок в канале с импульсными помехами.

Устройство чередования смешивает кодовые символы в промежутке нескольких длин блоков (для блочных кодов) или нескольких длин кодового ограничения (для сверточных кодов). Требуемый промежуток определяется длительностью пакета. Подробности структуры битового перераспределения должны быть известны приемнику, чтобы иметь возможность выполнить восстановление порядка битов перед декодированием. На рис. 8.10 показан простой пример чередования. На рис. 8.10, *а* мы можем видеть кодовые слова, которые еще не подвергались описанной операции, от *A* до *G*. Каждое кодовое слово состоит из семи кодовых символов. Пусть наш код может исправлять однобитовые ошибки в любой 7-символьной последовательности. Если промежуток памяти канала равен длительности одного кодового слова, такой пакет, длительностью в семь символов, может уничтожить информацию в одном или двух кодовых словах. Тем не менее допустим, что после получения кодированных данных кодовые символы затем перемешиваются, как показано на рис. 8.10, *б*. Иными словами, каждый кодовый символ каждого кодового слова отделяется от своего соседа на расстояние из семи символьных периодов. Полученный поток затем преобразуется в модулированный сигнал и передается по каналу. Как можно видеть на рис. 8.10, *б*, последовательные канальные пакеты шума попадают на семь символьных промежутков, влияя на один кодовый символ каждого из семи исходных кодовых слов. Во время приема в потоке вначале восстанавливается исходный порядок битов, так что он становится похож на исходную кодированную последовательность, изображенную на рис. 8.10, *а*. Затем поток декодируется. Поскольку в каждом кодовом слове возможно исправление одиночной ошибки, импульсная помеха не оказывает никакого влияния на конечную последовательность.

Идея чередования битов используется во всех блочных и сверточных кодах, рассмотренных здесь и ранее в предыдущих главах. Обычно применяются два типа устройств чередования — блочные и сверточные (оба рассматриваются далее).

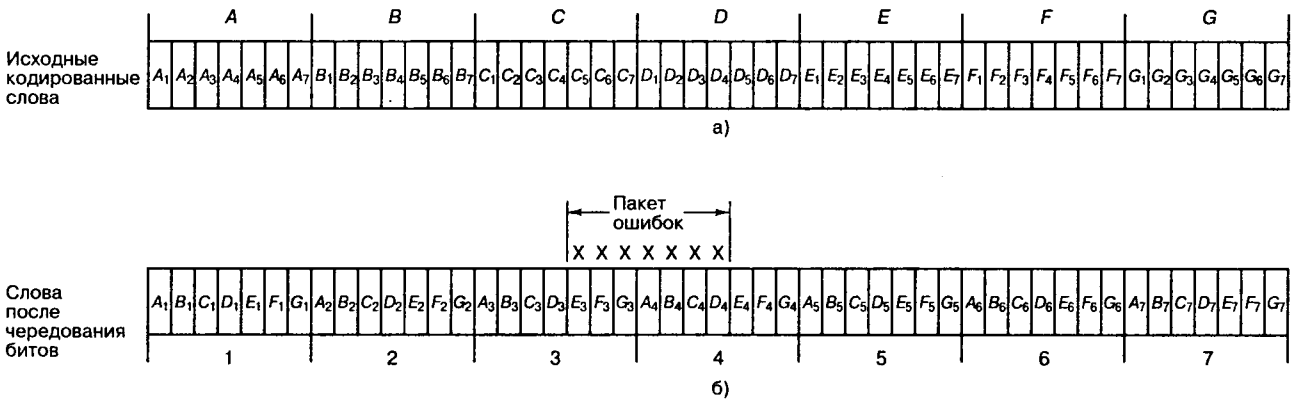


Рис. 8.10. Пример процедуры чередования битов: а) исходные кодовые слова, содержащие семь кодовых символов; б) полученные кодовые символы

8.2.1. Блочное чередование

Блочное устройство чередования принимает кодированные символы блоками от кодера, переставляет их, а затем передает измененные символы на модулятор. Как правило, перестановка блоков завершается заполнением столбцов матрицы M строками и N столбцами ($M \times N$) кодированной последовательности. После того как матрица полностью заполнена, символы подаются на модулятор (по одной строке за раз), а затем передаются по каналу. В приемнике устройство восстановления выполняет обратные операции; оно принимает символы из демодулятора, восстанавливает исходный порядок битов и передает их на декодер. Символы поступают в массив устройства восстановления по строкам и заменяются столбцами. На рис. 8.11, *а* приведен пример устройства чередования с $M = 4$ строками и $N = 6$ столбцами. Записи в массиве отображают порядок, в котором 24 кодовых символа попадают в устройство чередования. Выходная последовательность, предназначенная для передатчика, состоит из кодовых символов, которые построчно удалены из массива, как показано на рисунке. Ниже перечисляются наиболее важные характеристики такого блочного устройства.

1. Пакет, который содержит меньше N последовательных канальных символов, дает на выходе устройства восстановления исходного порядка символов ошибки, разнесенные между собой, по крайней мере, на M символов.
2. Пакет из bN ошибок, где $b > 1$, дает на выходе устройства восстановления пакет, который содержит не меньше $\lceil b \rceil$ символьных ошибок. Каждый из пакетов ошибок отделен от другого не меньше, чем на $M - \lfloor b \rfloor$ символов. Запись $\lceil x \rceil$ означает наименьшее целое число, не меньшее x, а запись $\lfloor x \rfloor$ — наибольшее целое число, не превышающее x.
3. Периодическая последовательность одиночных ошибок, разделенных N символами, дает на выходе устройства восстановления одиночные пакеты ошибок длиной M.
4. Прямая задержка между устройствами чередования и восстановления равна приблизительно длительности $2MN$ символов. Если быть точным, перед тем как начать передачу, нужно заполнить лишь $M(N - 1) + 1$ ячеек памяти (как только будет внесен первый символ последнего столбца массива $M \times N$). Соответствующее время нужно приемнику, чтобы начать декодирование. Значит, минимальная прямая задержка будет составлять длительность ($2MN - 2M + 2$) символов, не учитывая задержку на передачу по каналу.
5. Необходимая память составляет MN символов для каждого объекта (устройств чередования и восстановления исходного порядка). Однако массив $M \times N$ нужно заполнить (по большей части) до того, как он будет считан. Для каждого объекта нужно предусмотреть память для $2MN$ символов, чтобы опорожнить массив $M \times N$, пока другой будет наполняться, и наоборот.

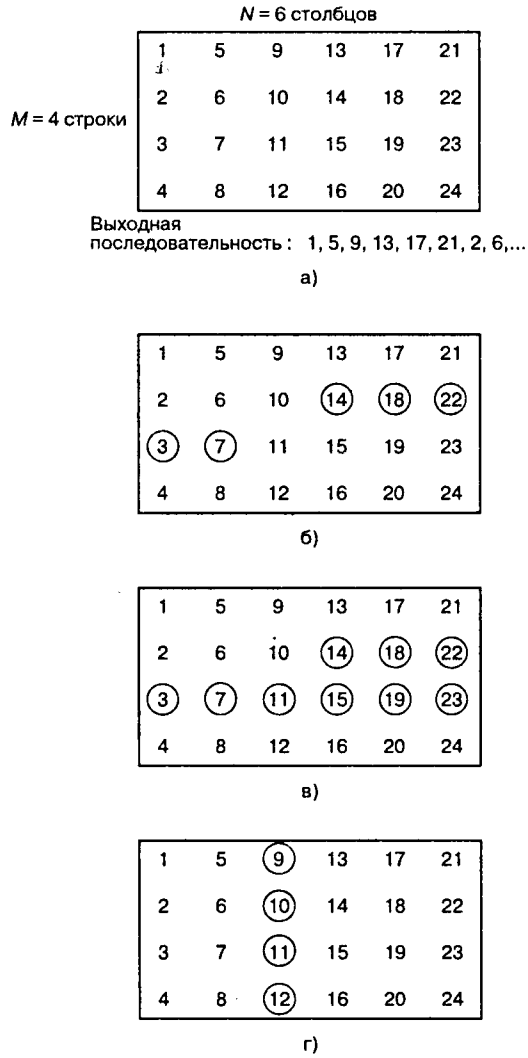


Рис. 8.11. Пример блочного чередования: а) блочное устройство чередования размером $M \times N$; б) пятисимвольный пакет ошибок; в) девятисимвольный пакет ошибок; г) периодическая последовательность одиночных ошибок, разнесенных на $N = 6$ символов

Пример 8.4. Характеристики устройства чередования

Используя структуру устройства чередования с $M = 4$, $N = 6$, изображенную на рис. 8.11, *а*, проверьте описанные выше характеристики.

Решение

1. Пусть имеется пакет шума длительностью в пять символьных интервалов; так что символы, выделенные на рис. 8.11, *б*, подвергнутся искажению во время передачи. После восстановления исходного порядка битов в приемнике, последовательность принимает следующий вид.

1 2 (3) 4 5 6 (7) 8 9 10 11 12

13 (14) 15 16 17 (18) 19 20 21 (22) 23 24

Здесь выделенные символы являются ошибочными. Можно видеть, что минимальное расстояние, разделяющее символы с ошибками, равно $M = 4$.

2. Пусть $b = 1{,}5$, так что $bN = 9$. Пример девятисимвольного пакета ошибок можно видеть на рис. 8.11, *в*. После того как в приемнике проведена процедура восстановления исходного порядка, последовательность примет следующий вид.

   1 2 (3) 4 5 6 (7) 8 9 10 (11) 12

   13 (14) (15) 16 17 (18) (19) 20 21 (22) (23) 24

   Снова выделенные символы являются ошибочными. Здесь можно видеть, что пакеты содержат не больше $\lceil 1{,}5 \rceil = 2$ символов подряд и разнесены, по крайней мере, на $M - \lfloor 1{,}5 \rfloor = 4 - 1 = 3$ символа.

3. На рис. 8.11, *г* показана последовательность одиночных ошибок, разделенных (каждый по отдельности) $N = 6$ символами. После восстановления исходного порядка в приемнике, последовательность принимает следующий вид.

   1 2 3 4 5 6 7 8 (9) (10) (11) (12)

   13 14 15 16 17 18 19 20 21 22 23 24

   Можно видеть, что после этого последовательность содержит пакет одиночных ошибок длиной $M = 4$ символа.

4. Прямая задержка: минимальная прямая задержка, вызванная обоими устройствами, составляет $(2MN - 2M + 2) = 42$ символьных периода.

5. Требуемый объем памяти: размерность массивов устройств чередования и восстановления составляет $M \times N$. Значит, требуется объем памяти для хранения $MN = 24$ символов на обоих концах канала. Как упоминалось ранее, в общем случае память реализуется для хранения $2MN = 48$ символов.

Как правило, параметры устройства чередования, используемого совместно с кодом с коррекцией одиночных ошибок, выбираются таким образом, чтобы число столбцов N превышало ожидаемую длину пакета. Выбираемое число строк зависит от того, какая схема кодирования будет использована. Для блочных кодов M должно быть больше длины кодового блока; для сверточных кодов M должно превышать длину кодового ограничения. Поэтому пакет длиной N может вызвать в блоке кода (самое большее) одиночную ошибку; аналогично в случае сверточных кодов в пределах одной длины кодового ограничения будет не более одной ошибки. Для кодов с коррекцией ошибок кратности t, выбираемое N должно лишь превышать ожидаемую длину пакета, деленную на t.

8.2.2. Сверточное чередование

Сверточные устройства чередования были предложены Рамси (Ramsey) [7] и Форни (Forney) [8]. Схема, предложенная Форни, показана на рис. 8.12. Кодовые символы последовательно подаются в блок из N регистров; каждый последующий регистр может хранить на J символов больше, чем предыдущий. Нулевой регистр не предназначен для хранения

(символ сразу же передается). С каждым новым кодовым символом коммутатор переключается на новый регистр, и кодовый символ подается на него до тех пор, пока наиболее старый кодовый символ в регистре не будет передан на модулятор/передатчик. После $(N-1)$-го регистра коммутатор возвращается к нулевому регистру и повторяет все снова. После приема операции повторяются в обратном порядке. И вход, и выход устройств чередования и восстановления должны быть синхронизированы.

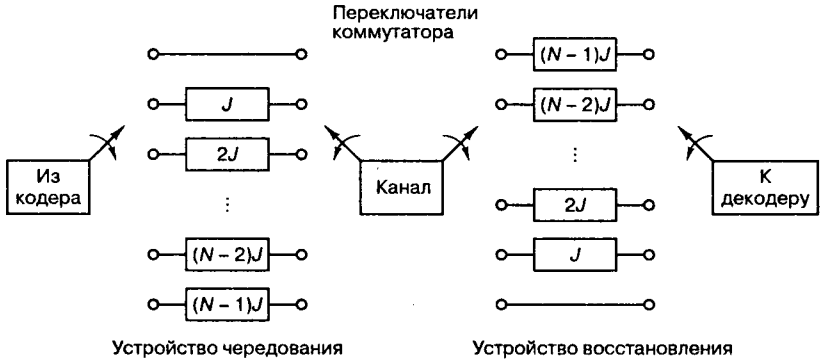


Рис. 8.12. Реализация регистра сдвига для сверточного устройства чередования/восстановления

На рис. 8.13 показан пример простого сверточного четырехрегистрового ($J = 1$) устройства чередования, загруженного последовательностью кодовых символов. Одновременно представлено синхронизированное устройство восстановления, которое передает обработанные символы на декодер. На рис. 8.13, *а* показана загрузка символов 1–4; знак "×" означает неизвестное состояние. На рис. 8.13, *б* представлены первые четыре символа, подаваемые в регистры, и показана передача символов 5–8 на выход устройства чередования. На рис. 8.13, *в* показаны поступающие в устройство символы 9–12. Теперь устройство восстановления заполнено символами сообщения, но еще не способно ничего передавать на декодер. И наконец, на рис. 8.13, *г* показаны символы 13–16, поступившие в устройство чередования, и символы 1–4, переданные на декодер. Процесс продолжается таким образом до тех пор, пока полная последовательность кодового слова не будет передана на декодер в своей исходной форме.

Рабочие характеристики сверточного устройства чередования сходны с параметрами блочного устройства. Важным преимуществом сверточного устройства перед блочным является то, что при сверточном чередовании прямая задержка составляет $M(N-1)$ символов при $M = NJ$, а требуемые объемы памяти — $M(N-1)/2$ на обоих концах канала. Очевидно, что требования к памяти и время задержки снижаются вдвое, по сравнению с блочным чередованием [9].

8.2.3. Каскадные коды

Каскадными называются коды, в которых кодирование осуществляется в два уровня; имеется внутренний и внешний коды, с помощью которых и достигается желаемая надежность передачи сообщений. На рис. 8.14 изображен порядок кодирования и декодирования. Внутренний код связан с модулятором (демодулятором) и каналом; он, как правило, настраивается для исправления большинства канальных ошибок. Внешний код, чаще всего высокоскоростной (с низкой избыточностью), снижает вероят-

ность появления ошибок до заданного значения. Основной причиной использования каскадного кода является низкая степень кодирования и общая сложность реализации, меньшая той, которая потребовалась бы для осуществления отдельной процедуры кодирования. На рис. 8.14 между двумя этапами кодирования располагается устройство чередования. Обычно это делается для того, чтобы разнести пакетные ошибки, которые могли бы появиться в результате внутреннего кодирования.

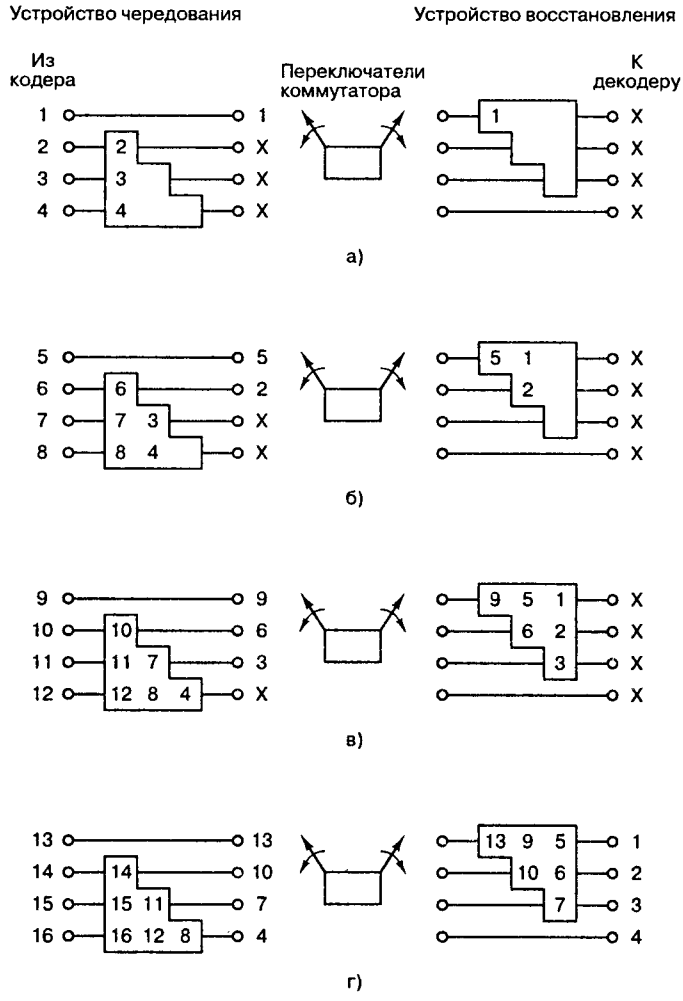


Рис. 8.13. Пример сверточного чередования/восстановления

В одной из наиболее популярных систем каскадного кодирования для внутреннего кода применяется сверточное кодирование по алгоритму Витерби, а для внешнего — код Рида-Соломона с чередованием между двумя этапами кодирования [2]. Функционирование таких систем при E_b/N_0, находящемся в пределах от 0,2 до 2,5 дБ, для достижения $P_B = 10^{-5}$ реально достижимо в прикладных задачах [9]. В этой системе демодулятор выдает мягко квантованные кодовые символы на внутренний сверточный декодер, который, в свою очередь, выдает жестко квантованные кодовые символы с пакетными ошибками на декодер Рида-Соломона.

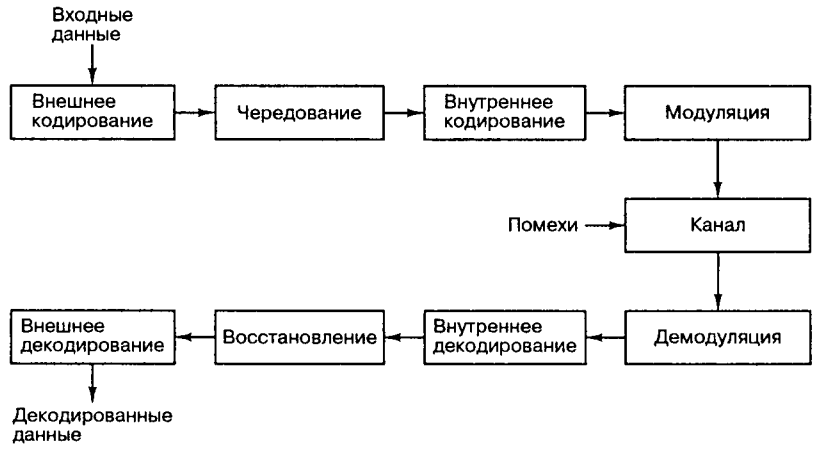


Рис. 8.14. Блочная диаграмма каскадной системы кодирования

(В системе декодирования по алгоритму Витерби выходные ошибки имеют тенденцию к появлению пакетами.) Внешний код Рида-Соломона образуется из m-битовых сегментов двоичного потока данных. Производительность такого (недвоичного) кода Рида-Соломона зависит только от числа *символьных ошибок* в блоке. Код не искажается пакетами ошибок внутри m-битового символа. Иными словами, для данной символьной ошибки производительность кода Рида-Соломона такова, как если бы символьная ошибка была вызвана одним битом или m бит. Тем не менее производительность каскадных систем несколько ухудшается за счет коррелирующих ошибок в последовательных символах. Поэтому чередование между кодированиями нужно выполнять на уровне символов (а не битов). Работа [10] представляет собой обзор каскадных кодов, которые были разработаны для дальней космической связи. В следующем разделе мы рассмотрим распространенную практическую реализацию символьного чередования в каскадных системах.

8.3. Кодирование и чередование в системах цифровой записи информации на компакт-дисках

В 1979 году компании Philips Corp. (Нидерланды) и Sony Corp. (Япония) запатентовали стандарт хранения и воспроизведения цифровой записи аудиосигналов, известный как *система цифровой записи на компакт-дисках* (compact disc digital audio — CD-DA). Эта система стала мировым стандартом, позволяющим достичь безукоризненной точности воспроизведения звука, и опередила другие методики. Для хранения оцифрованных аудиосигналов используется пластиковый диск диаметром 120 мм. Сигнал дискретизирован с частотой 44100 фрагментов/с для получения записи в полосе 20 кГц. Каждый аудиофрагмент однозначно квантован на один из 216 уровней (16 бит/фрагмент), что дает в результате динамический диапазон в 96 дБ и нелинейное искажение 0,005%. Отдельный диск (время звучания составляет порядка 70 минут) хранит порядка 10^{10} бит в виде коротких *впадин*, которые сканируются лазером.

В данном случае существует несколько источников канальных ошибок: 1) маленькие нежелательные частички или воздушные пузырьки в материале пластика или неточное расположение *впадин* при изготовлении диска; 2) отпечатки пальцев или цара-

пины, появившиеся при эксплуатации. Трудно предсказать, как в среднем можно повредить компакт-диск; но при наличии точной канальной модели можно со всей уверенностью сказать, что канал, в основном, склонен вносить *пакетоподобные* ошибки, поскольку царапины или пятна от пальцев будут вызывать ошибки в *нескольких* последовательных фрагментах данных. Важным элементом разработки системы получения высококачественных характеристик является каскадная схема защиты от ошибок, которая называется *кодом Рида-Соломона с перекрестным чередованием* (cross-interleave Reed-Solomon code — CIRC). Данные перемешиваются во времени так, что знаки, выходящие из последовательных фрагментов сигнала, оказываются *разнесенными во времени*. Таким образом, появление ошибок представляется в виде одиночных случайных ошибок (см. предыдущий раздел). Цифровая информация защищена посредством прибавления байтов четности, получаемых в двух кодерах Рида-Соломона. Защита от ошибок, осуществляемая на компакт-дисках, зависит обычно от кодирования Рида-Соломона и алгоритма чередования.

В прикладных задачах передачи цифровой аудиоинформации, невыявляемая ошибка декодирования очень значительна, поскольку является результатом щелчка при воспроизведении, в то время как *выявляемые* ошибки незначительны, так как их можно скрыть. Схема защиты от ошибок CIRC в системе CD-DA включает в себя и *исправление*, и *маскировку* ошибок. Технические характеристики схемы CIRC даются в табл. 8.4. Из данных таблицы должно быть ясно, что компакт-диск может выдержать сильные повреждения (например, 8-миллиметровые отверстия, пробитые в диске) без значительных потерь в качестве звучания.

Таблица 8.4. Спецификация кода Рида-Соломона с перекрестным чередованием, применяемого в аудиокомпакт-дисках

Максимальная длина исправимого пакета	≈ 4000 бит (2,5 мм длины дорожки на диске)
Максимальная длина пакета, который можно интерполировать	≈ 12000 бит (8 мм)
Скорость интерполяции фрагмента	1 фрагмент/10 часов при $P_B = 10^{-4}$; 1000 фрагментов/мин. при $P_B = 10^{-3}$
Необнаруженные фрагменты с ошибками (щелчки)	Менее чем 1 на 750 часов при $P_B = 10^{-3}$ Пренебрежимо малое количество при $P_B \leq 10^{-4}$
Качество нового диска	$P_B \approx 10^{-4}$

В системе CIRC защита от ошибок обеспечивается множеством способов.

1. Декодер обеспечивает нужный уровень коррекции ошибок.
2. Если исчерпывается способность к коррекции ошибок, то декодер переходит на уровень коррекции стираний (см. раздел 6.5.5).
3. Если исчерпывается и эта способность, декодер предпринимает попытки замаскировать ненадежные фрагменты данных путем *интерполяции* между ближайшими надежными фрагментами.
4. Если исчерпывается способность к интерполяции, декодер выключает или *подавляет* систему на период ненадежного фрагмента.

8.3.1. Кодирование по схеме CIRC

На рис. 8.15 показана основная блочная диаграмма кодера CIRC (с оборудованием для записи компакт-диска) и декодера (с оборудованием для воспроизведения компакт-диска). Процедура кодирования состоит из собственно кодирования и чередования, где введены следующие обозначения: Δ-чередование, C_2-кодирование, D^*-чередование, C_1-кодирование и D-чередование. Процедура декодирования состоит из этапов декодирования и восстановления исходного порядка битов, которые выполняются в *обратном* порядке; здесь идут D-восстановление, C_1-декодирование, D^*-восстановление, C_2-декодирование и Δ-восстановление.

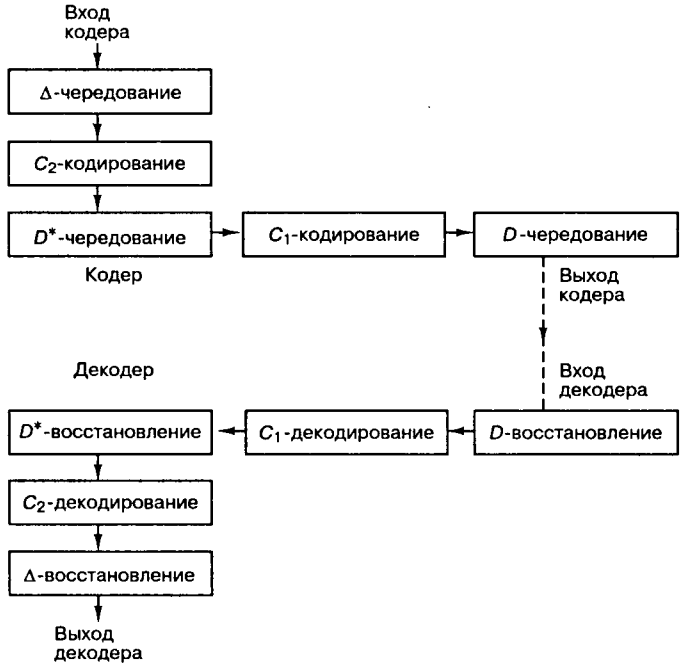


Рис. 8.15. Схема кодера и декодера CIRC

На рис. 8.16 показан элементарный период системного кадра и шесть периодов дискретизации, каждый из которых состоит из пары стереофрагментов (16-битовый левый фрагмент и 16-битовый правый фрагмент). Биты собраны в символы или байты размером 8 бит каждый. Следовательно, каждая пара фрагментов содержит 4 байт, а некодированный кадр — $k = 24$ байт. На рис. 8.16, *а—д* представлены *пять этапов кодирования*, которые характеризуют систему CIRC. Функции каждого этапа будут более понятны, если мы рассмотрим процедуру декодирования. Этапы выглядят следующим образом.

а) Δ-*чередование*. Четные фрагменты отделяются от нечетных двумя кадрами для перемешивания ошибок, которые определены, но нельзя исправить. Это облегчает процесс интерполяции.

б) C_2-*кодирование*. К Δ-чередованному 24-байтовому кадру прибавляется четыре байта четности Рида-Соломона, что дает в итоге $n = 28$ байт. Такой код (28, 24) называется *внешним*.

в) *D*-чередование.* Здесь каждый байт задерживается на разную длину; таким образом ошибки разбрасываются на несколько кодовых слов. C_2-кодирование совместно с *D**-чередованием нужно для исправления пакетных ошибок и ошибочных комбинаций, которые C_1-декодер не в состоянии исправить.

г) *C_1-кодирование.* К $k = 28$ байт *D**-чередованного кадра прибавляется четыре байта четности Рида-Соломона, что дает в итоге всего $n = 32$ байт. Такой код (32, 28) называется *внутренним.*

д) *D-чередование.* Осуществляется *перекрестное чередование четных байтов* кадра с *нечетными байтами* следующего кадра. После этой процедуры два последовательных байта на диске будут всегда расположены в двух разных кодовых словах. При декодировании это чередование даст возможность исправлять большинство случайных одиночных ошибок и обнаруживать более длинные пакеты ошибок.

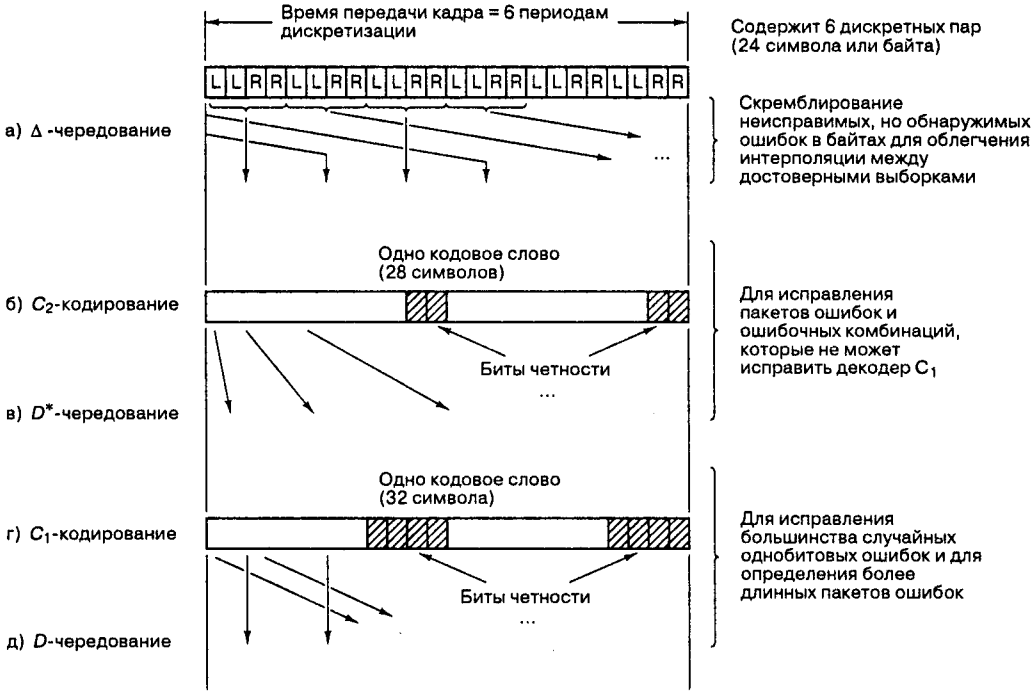


Рис. 8.16. Кодирование компакт-диска: а) Δ-чередование; б) C_2-кодирование; в)D-чередование; г) C_1-кодирование; д) D-чередование*

8.3.1.1. Укорачивание кода Рида-Соломона

В разделе 8.1 код (n, k) выражался через $n = 2^m - 1$ итоговых символов и $k = 2^m - 1 - 2t$ символов данных, где m представляет собой число битов в символе, а t — способность кода к коррекции ошибок в символах. Для системы CD-DA, где символ образован из 8 бит, код с коррекцией 2-битовых ошибок можно сконфигурировать как код (255, 251). Однако в системе CD-DA используется значительно меньшая длина блока. Любой блочный код (в систематической форме) можно укоротить без уменьшения числа ошибок, которые поддаются исправлению внутри блока. Представим себе, что в

терминах кода (255, 251), 227 из 251 информационного символа являются набором нулевых символов (которые в действительности не передавались и поэтому не содержат ошибок). Тогда код в действительности будет кодом (28, 24) с той же коррекцией 2-символьных ошибок. Это и делается в C_2-кодере системы CD-DA.

Мы можем представить 28 символов вне C_2-кодера как информационные символы в C_1-кодере. И снова можно сконфигурировать сокращенный код (255, 251) с коррекцией 2-символьных ошибок, отбросив 223 символа данных; результатом будет код (32, 28).

8.3.2. Декодирование по схеме CIRC

Внутренний и внешний коды Рида-Соломона с параметром (n, k), равным (32, 28) и (28, 24), используют четыре контрольных байта. Степень кодирования кода в схеме CIRC равна $(k_1/n_1)(k_1/n_1) = 24/32 = 3/4$. Из уравнения (8.3) следует, что минимальное расстояние C_1 и C_2 кодов Рида-Соломона будет $d_{\min} = n - k + 1 = 5$. Из уравнений (8.4) и (8.5) имеем следующее.

$$t \le \left\lfloor \frac{d_{\min} - 1}{2} \right\rfloor = \left\lfloor \frac{n-k}{2} \right\rfloor \tag{8.58}$$

$$\rho \le d_{\min} - 1 \tag{8.59}$$

Здесь t — способность к коррекции ошибок, а ρ — способность к коррекции стираний. Видно, что C_1- и C_2-декодеры могут исправить максимум 2 символьные ошибки или 4 символьных стирания на кодовое слово. Или, как определяется уравнением (8.6), имеется возможность исправлять α ошибок и γ стираний одновременно, если

$$2\alpha + \gamma < d_{\min} < n - k. \tag{8.60}$$

Существует компромисс между коррекцией ошибок и коррекцией стираний; чем больше возможностей задействовано в коррекции ошибок, тем меньше остается возможностей для коррекции стираний.

Преимущества схемы CIRC лучше видны на примере *декодера*. Рабочие этапы, изображенные на рис. 8.17, имеют обратный порядок по сравнению с кодером. Давайте рассмотрим этапы работы декодера.

1. *D-восстановление.* Этот этап нужен для чередования линий задержки, обозначенных символом D. 32 байт $(B_{i1}, \ldots, B_{i32})$ кодированного кадра выстраиваются для параллельной подачи на 32 входа D-восстановителя. Каждая задержка равна длительности 1 байт, так что происходит обращение перекрестного чередования информации *четных байтов* кадра с *нечетными байтами* следующего кадра.
2. *C_1-декодирование.* D-восстановитель и C_1-декодер разработаны для исправления однобайтовых ошибок в блоке из 32 байт и обнаружения больших пакетов ошибок. Если появляются многократные ошибки, то C_1-декодер пропускает их без изменений, приписывая ко всем 28 байт метку стирания и пересылая их по пунктирным линиям (четыре бита контроля четности используются в C_1-декодере и больше не сохраняются).

3. *D*-восстановление.* Из-за разности длины линий задержки $D^*(1, ..., 27)$ при восстановлении порядка битов, ошибки, возникающие в слове на выходе C_1-декодера, оказываются *разбросанными по большому количеству слов* на входе C_2-декодера, что позволяет C_2-декодеру заниматься исправлением этих ошибок.
4. *C_2-декодирование.* C_2-декодер применяется для исправления пакетов ошибок, которые не может исправить C_1-декодер. Если C_2-декодеру не удается исправить эти ошибки, то 24-байтовое кодовое слово пропускается без изменений на Δ-восстановитель и на соответствующие позиции ставится *метка стирания* по пунктирным линиям, B_{o1}, ..., B_{o24}.
5. *Δ-восстановление.* Это финальная операция, в ходе которой осуществляется обращение чередования неисправимых, но обнаружимых ошибок, в результате чего происходит интерполяция между соседними кадрами.

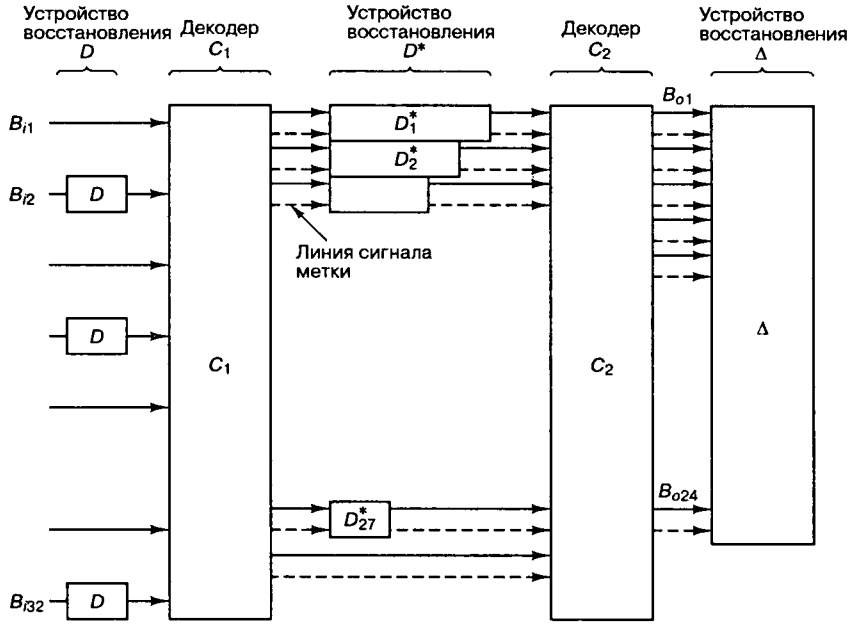


Рис. 8.17. Декодер системы воспроизведения компакт-дисков

На рис. 8.18 выделены 2-, 3- и 4-й этапы декодирования. На выходе C_1-декодера видна последовательность четырех 28-байтовых кодовых слов, которые превышают однобайтовую способность кода корректировать ошибки. Следовательно, каждый из символов в этих кодовых словах получает метку стирания (показана кружком). *D**-восстановитель выполняет разнесение линий задержки для каждого байта кодового слова так, что байты данного кодового слова попадают в разные кодовые слова на входе C_2-декодера. Если допустить, что коэффициент задержки *D**-восстановителя, изображенного на рис. 8.18, равен 1 байт, то можно исправить пакет ошибок четырех последовательных кодовых слов C_1 (поскольку C_2-декодер может исправить четыре стирания на кодовое слово). В прикладных системах CD-DA коэффициент задержки составляет 4 байт; поэтому максимальная способность кода к исправлению пакетных ошибок равняется 16 последовательным неисправленным C_1-словам.

8.3.3. Интерполяция и подавление

Фрагменты, которые нельзя исправить с помощью C_2-декодера, могут вызвать слышимые искажения. Роль процедуры *интерполяции* состоит в том, чтобы вставлять новые фрагменты, оцениваемые по ближайшим соседям, вместо ненадежных. Если полное слово признано C_2-ненадежным, то невозможно произвести интерполяцию без дополнительного чередования, поскольку и четные, и нечетные фрагменты одинаково ненадежны. Это может произойти, если C_1-декодер не обнаруживает ошибки, а C_2-декодер обнаруживает ее. Целью Δ-восстановления (в течение двух кадровых периодов) является вычисление структуры, в которой четные фрагменты можно интерполировать по нечетным или наоборот.

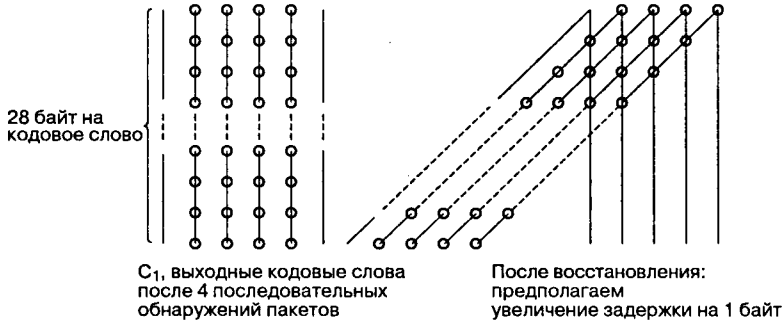


Рис. 8.18. Пример 4-байтовой возможности стираний (время показано справа налево)

На рис. 8.19 показаны два последовательных ненадежных слова, состоящих из 12 пар фрагментов. Пара фрагментов состоит из фрагмента (2 байта) правого аудиоканала и фрагмента левого. Числа означают порядок размещения фрагментов. Фрагменты, номера которых выделены, отмечены меткой *стирания*. После Δ-восстановления ненадежные фрагменты, показанные на рисунке, оцениваются с помощью линейной интерполяции первого порядка между ближайшими соседними фрагментами из разных мест диска.

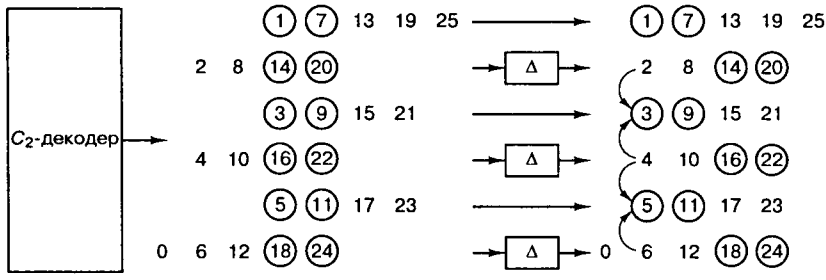


Рис. 8.19. Эффект чередования (время показано справа налево)

В проигрывателях компакт-дисков при появлении пакетов ошибок, превышающих 48 кадров и дающих в итоге 2 или более последовательных ненадежных фрагментов, применяется иной уровень защиты от ошибок. В этом случае система *подавляется* (звук приглушается), что незаметно для человеческого слуха, если время подавления не превышает нескольких миллисекунд. Для более подробного ознакомления со схемой кодирования CIRC в системе CD-DA см. [11–15].

8.4. Турбокоды

Схема каскадного кодирования впервые была предложена Форни [16] как метод получения высокоэффективного кода посредством комбинации двух или более *компонуемых кодов* (иногда называемых *составными*). В результате, такие коды могут корректировать ошибки в значительно более длинных кодах и имеют структуру, которая позволяет относительно легко осуществить декодирование средней сложности. Последовательные каскадные коды часто используются в системах с ограничением мощности, таких как космические зонды. Самая распространенная из этих схем содержит внешний код Рида-Соломона (выполняется первым, убирается последним), который следует за сверточным внутренним кодом (выполняется последним, убирается первым) [10]. Турбокод можно считать обновлением структуры каскадного кодирования с итеративным алгоритмом декодирования связанной кодовой последовательности. Поскольку такая схема имеет итеративную форму, на рис. 1.3 турбокодирование представлено как отдельная категория в структурированных последовательностях.

Турбокоды впервые были введены в 1993 году Берру, Главье и Цитимаджимой (Berrou, Glavieux, Thitimajshima) и опубликованы в [17, 18], где в описываемой схеме достигалась вероятность появления ошибок 10^{-5} при степени кодирования 1/2 и модуляции BPSK в канале с белым аддитивным гауссовым шумом (additive white Gaussian noise — AWGN) с E_b/N_0, равным 0,7 дБ. Коды образуются посредством компоновки двух или более составных кодов, являющихся разными вариантами чередования одной и той же информационной последовательности. Тогда как для сверточных кодов на финальном этапе декодер выдает жестко декодированные биты (или в более общем случае — декодированные символы), в каскадной схеме, такой как турбокод, для хорошей работы алгоритм декодирования не должен ограничивать себя, подавая на декодеры жесткую схему решений. Для лучшего использования информации, получаемой с каждого декодера, алгоритм декодирования должен применять, в первую очередь, мягкую схему декодирования, вместо жесткой. Для систем с двумя составными кодами концепция, лежащая в основе турбодекодирования, заключается в том, чтобы передать мягкую схему принятия решений с выхода одного декодера на вход другого и повторять эту процедуру до тех пор, пока не будут получены надежные решения.

8.4.1. Понятия турбокодирования

8.4.1.1. Функции правдоподобия

Математическое обоснование критерия проверки гипотез остается за теоремой Байеса, которая приводится в приложении Б. В области связи, где наибольший интерес представляют прикладные системы, включающие в себя каналы AWGN, наиболее распространенной формой теоремы Байеса является та, которая выражает апостериорную вероятность (a posteriori probability — APP) решения через случайную непрерывную переменную x как

$$P(d=i|x)=\frac{p(x|d=i)P(d=i)}{p(X)} \quad i=1,\ldots,M \tag{8.61}$$

и

$$p(x)=\sum_{i=1}^{M} p(x|d=i)P(d=i), \tag{8.62}$$

где $P(d=i|x)$ — это апостериорная вероятность, а $d=i$ представляет данные d, принадлежащие i-му классу сигналов из набора классов M. Ранее $p(x|d=i)$ представляло функцию плотности вероятности принимаемого непрерывного сигнала с шумом x, при $d=i$. Также $p(d=i)$, называемое априорной вероятностью, означает вероятность появления i-го класса сигналов. Обычно x представляет "наблюдаемую" случайную переменную или лежащую в основе критерия статистику, которая получается на выходе демодулятора или какого-либо иного устройства обработки сигналов. Поэтому $p(x)$ — это функция распределения вероятностей принятого сигнала x, дающая тестовую статистику в полном пространстве классов сигналов. В уравнении (8.61) при конкретном наблюдении $p(x)$ является коэффициентом масштабирования, поскольку он получается путем усреднения по всем классам пространства. Литера p нижнего регистра используется для обозначения функции распределения вероятностей непрерывной случайной переменной, а литера P верхнего регистра — для обозначения вероятности (априорной и апостериорной). Определение апостериорной вероятности принятого сигнала, из уравнения (8.61), можно представлять как результат эксперимента. Перед экспериментом, в общем, существует (или поддается оценке) априорная вероятность $P(d=i)$. В эксперименте для расчета апостериорной вероятности, $P(d=i|x)$, используется уравнение (8.61), и это можно считать "обновлением" имевшихся сведений, полученных при изучении принятого сигнала x.

8.4.1.2. Пример класса из двух сигналов

Пусть двоичные логические элементы 1 и 0 представляются электрическими напряжениями +1 и –1. Переменная d представляет бит переданных данных, который выглядит как уровень напряжения или логический элемент. Иногда более предпочтительным оказывается один из способов представления; читатель должен уметь различать это по контексту. Пусть двоичный 0 (или электрическое напряжение –1) будет нулевым элементом при сложении. На рис. 8.20 показана условная функция распределения вероятностей при передаче сигнала по каналу AWGN, представленная как функция правдоподобия. Функция, изображенная справа, $p(x|d=+1)$, представляет функцию распределения вероятностей случайной переменной x, которая передается при условии, что $d=+1$. Функция, изображенная слева, $p(x|d=-1)$, в свою очередь, представляет ту же функцию распределения вероятностей случайной переменной x, которая передается при условии, что $d=-1$. На оси абсцисс показан полный диапазон возможных значений тестовой статистики x, которая образуется в приемнике. На рис. 8.20 показано одно такое произвольное значение x_k, индекс которого представляет наблюдение, произведенное в k-й период времени. Прямая, опущенная в точку x_k, пересекает две кривые функций правдоподобия, что дает в итоге два значения правдоподобия $l_1=p(x_k|d_k=+1)$ и $l_2=p(x_k|d_k=-1)$. Хорошо известное правило принятия решения по жесткой схеме, называемое *принципом максимального правдоподобия*, определяет выбор данных $d_k=+1$ или $d_k=-1$, основываясь на большем из двух имеющихся значений l_1 или l_2. Для каждого бита данных в момент k решение гласит, что $d_k=+1$, если x_k попадает по правую сторону линии принятия решений, обозначаемой γ_0, в противном случае — $d_k=-1$.

Аналогичное правило принятия решения, известное как *максимум апостериорной вероятности* (maximum a posteriori — MAP), можно представить в виде *правила минимальной вероятности ошибки*, принимая во внимание априорную вероятность данных. В общем случае правило MAP выражается следующим образом.

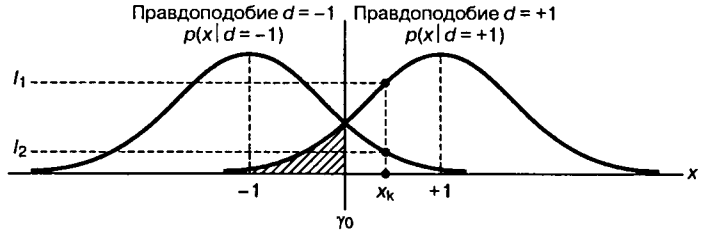


Рис. 8.20. Функции правдоподобия

$$P(d=+1\,|\,x) \underset{H_2}{\overset{H_1}{\gtrless}} P(d=-1\,|\,x) \tag{8.63}$$

Уравнение (8.63) утверждает, что выбирается одна из гипотез — H_1, $(d=+1)$, если апостериорная вероятность $P(d=+1|x)$ больше апостериорной вероятности $P(d=-1|x)$. В противном случае выбирается гипотеза H_2, $(d=-1)$. Воспользовавшись байесовской формой уравнения (8.61), можно заменить апостериорную вероятность в уравнении (8.63) эквивалентным выражением, что дает следующее.

$$p(x\,|\,d=+1)P(d=+1) \underset{H_2}{\overset{H_1}{\gtrless}} p(x\,|\,d=-1)P(d=-1) \tag{8.64}$$

Здесь функция распределения вероятности $p(x)$, имеющаяся в обеих частях неравенства, (8.61), была исключена. Уравнение (8.64), в целом представленное через дроби, дает так называемую *проверку отношения правдоподобий.*

$$\frac{p(x\,|\,d=+1)}{p(x\,|\,d=-1)} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{P(d=-1)}{P(d=+1)} \text{ или } \frac{p(x\,|\,d=+1)P(d=+1)}{p(x\,|\,d=-1)P(d=-1)} \underset{H_2}{\overset{H_1}{\gtrless}} 1 \tag{8.65}$$

8.4.1.3. Логарифмическое отношение правдоподобий

Если взять логарифм от соотношения правдоподобия, полученного в уравнениях (8.63)–(8.65), получится удобная во многих отношениях метрика, называемая логарифмическое отношение правдоподобия (log-likelihood ratio — LLR). Это вещественное представление мягкого решения вне декодера определяется выражением

$$L(d|x) = \lg\left[\frac{P(d=+1|x)}{P(d=-1|x)}\right] = \lg\left[\frac{p(x|d=+1)P(d=+1)}{p(x|d=-1)P(d=-1)}\right] \tag{8.66}$$

так, что

$$L(d|x) = \lg\left[\frac{p(x|d=+1)}{p(x|d=-1)}\right] + \lg\left[\frac{P(d=+1)}{P(d=-1)}\right] \tag{8.67}$$

или

$$L(d|x) = L(x|d) + L(d), \tag{8.68}$$

где $L(x|d)$ — это LLR тестовой статистики x, получаемой путем измерений x на выходе канала при чередовании условий, что может быть передан $d=+1$ или $d=-1$, а $L(d)$ — априорное LLR бита данных d. Для упрощения обозначений уравнение (8.68) можно переписать следующим образом.

$$L'(\hat{d}) = L_c(x) + L(d) \tag{8.69}$$

Здесь $L_c(x)$ означает, что данный член LLR получается в результате канальных измерений, произведенных в приемнике. Уравнения (8.61)—(8.69) получены только исходя из данных детектора. Далее введение декодера даст стандартные преимущества схемы принятия решений. Для систематических кодов было показано [17], что LLR (мягкий выход) вне декодера равняется следующему.

$$L(\hat{d}) = L'(\hat{d}) + L_e(\hat{d}) \tag{8.70}$$

Здесь $L'(\hat{d})$ — это LLR бита данных вне демодулятора (на входе декодера), а $L_e(\hat{d})$ называется *внешним* LLR и представляет внешнюю информацию, вытекающую из процесса декодирования. Выходная последовательность систематического декодера образована величинами, представляющими информационные биты или биты четности. Из уравнений (8.69) и (8.70) выходное LLR декодера теперь примет следующий вид.

$$L(\hat{d}) = L_c(x) + L(d) + L_e(\hat{d}) \tag{8.71}$$

Уравнение (8.71) показывает, что выходное LLR систематического декодера можно представить как состоящее из трех компонентов — канального измерения, априорного знания данных и внешнего LLR, относящегося только к декодеру. Чтобы получить финальное $L(\hat{d})$, нужно просуммировать отдельные вклады LLR, как показано в уравнении (8.71), поскольку все три компонента статистически независимы [17, 19]. Доказательство оставляем читателю в качестве самостоятельного упражнения (см. задачу 8.18.). Мягкий выход декодера $L(\hat{d})$ является вещественным числом, обеспечивающим в итоге как само принятие жесткого решения, так и его надежность. Знак $L(\hat{d})$ задает жесткое решение, т.е. при положительном знаке $L(\hat{d})$ решение — $d=+1$, а при отрицательном — $d=-1$. Величина $L(\hat{d})$ определяет надежность этого решения. Часто величина $L_e(\hat{d})$ вследствие декодирования имеет тот же знак, что и $L_c(x)+L(d)$, и поэтому повышает надежность $L(\hat{d})$.

8.4.1.4. Принципы итеративного (турбо) декодирования

В типичном приемнике демодулятор часто разрабатывается для выработки решений по мягкой схеме, которые затем будут переданы на декодер. В главе 7 повышение достоверности передачи в системе, по сравнению с жесткой схемой принятия решений, оценивается приблизительно в 2 дБ в канале AWGN. Такой декодер следует называть декодером с мягким входом и жестким выходом, поскольку процесс финального декодирования должен завершаться битами (жесткая схема). В турбокодах, где используется два или несколько составных кодов и декодирование подразумевает подключение выхода одного декодера ко входу дру-

гого для возможности поддержки итераций, декодер с жестким выходом нежелателен. Это связано с тем, что жесткая схема в декодере снизит производительность системы (по сравнению с мягкой схемой). Следовательно, для реализации турбодекодирования необходим *декодер с мягким входом и мягким выходом.* Во время первой итерации на таком декодере (с мягким входом и мягким выходом), показанном на рис. 8.21, данные считаются равновероятными, что дает начальное априорное значение LLR $L(d) = 0$ для третьего члена уравнения (8.67). Канальное значение LLR $L_c(x)$ получается путем взятия логарифма отношения величин l_1 и l_2 для определенных значений x (рис. 8.20) и является вторым членом уравнения (8.67). Выход декодера $L(\hat{d})$ на рис. 8.21 образуется из LLR детектора $L'(\hat{d})$ и внешнего LLR выхода $L_e(\hat{d})$ и представляет собой сведения, вытекающие из процесса декодирования. Как показано на рис. 8.21 для итеративного декодирования, внешнее правдоподобие подается обратно на вход (иного составного декодера) для обновления априорной вероятности информации следующей итерации.

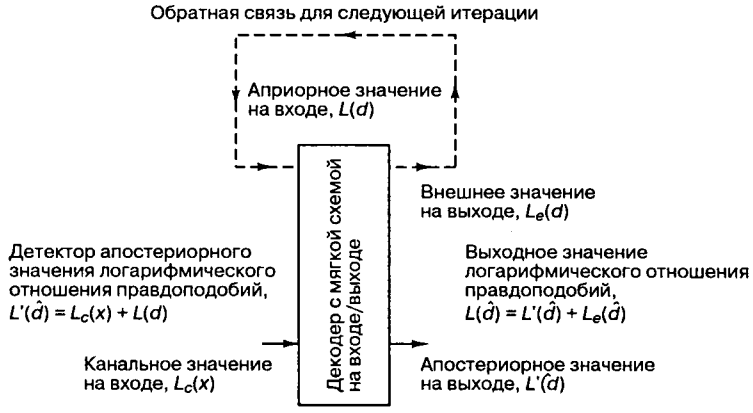


Рис. 8.21. Декодер с мягким входом и мягким выходом

8.4.2. Алгебра логарифма правдоподобия

Для более подробного объяснения итеративной обратной связи выходов мягких декодеров, вводится понятие алгебры логарифма правдоподобия [19]. Для статистически независимых данных d сумма двух логарифмических отношений правдоподобия (log-likelihood ratio — LLR) определяется следующим образом.

$$L(d_1) \boxplus L(d_2) \overset{\text{def}}{=} L(d_1 \oplus d_2) = \ln\left[\frac{e^{L(d_1)} + e^{L(d_2)}}{1 + e^{L(d_1)} e^{L(d_2)}}\right] \approx \tag{8.72}$$

$$\approx (-1) \times \operatorname{sgn}[L(d_1)] \times \operatorname{sgn}[L(d_2)] \times \min\left(|L(d_1)|, |L(d_2)|\right) \tag{8.73}$$

Здесь использован натуральный логарифм, а функция sgn (·) возвращает знак своего аргумента. В уравнении (8.72) имеется три операции сложения. Знак "⊕" применяется для обозначения суммы по модулю 2 данных, представленных двоичными цифрами. Знак ⊞ используется для обозначения суммы логарифмов правдоподобия или, что

то же самое, математической операции, описываемой уравнением (8.72). Сумма двух LLR обозначается оператором $\boxplus$, который определяется как LLR суммы по модулю 2 основных статистически независимых информационных битов. Вывод уравнения (8.72) показан в приложении 8А. Уравнение (8.73) является аппроксимацией уравнения (8.72), которая будет использована позднее в численном примере. Сложение LLR, определяемое уравнениями (8.72) и (8.73), дает один очень интересный результат в том случае, если один из LLR значительно превышает второй.

$$L(d) \boxplus \infty = -L(d)$$

и

$$L(d) \boxplus 0 = 0$$

Следует сказать, что алгебра логарифма правдоподобия, описанная здесь, немного отличается от той, которая используется в [19], из-за различного выбора нулевого элемента. В данном случае нулевым элементом двоичного набора (1, 0) выбран 0.

8.4.3. Пример композиционного кода

Рассмотрим двухмерный код (композиционный код), изображенный на рис. 8.22. Его структуру можно описать как массив данных, состоящий из k_1 строк и k_2 столбцов. В k_1 строках содержатся кодовые слова, образованные k_2 битами данных и $n_2 - k_2$ битами четности. Каждая из k_1 строк представляет собой кодовое слово кода (n_2, k_2). Аналогично k_2 столбцов содержат кодовые слова, образованные из k_1 бит данных и $n_1 - k_1$ бит четности. Таким образом, каждый из k_2 столбцов представляет собой кодовые слова кода (n_1, k_1). Различные участки структуры обозначены следующим образом: d — для данных, p_h — для горизонтальной четности (вдоль строк) и p_v — для вертикальной четности (вдоль столбцов). Фактически каждый блок битов данных размером $k_1 \times k_2$ кодирован двумя кодами — горизонтальным и вертикальным.

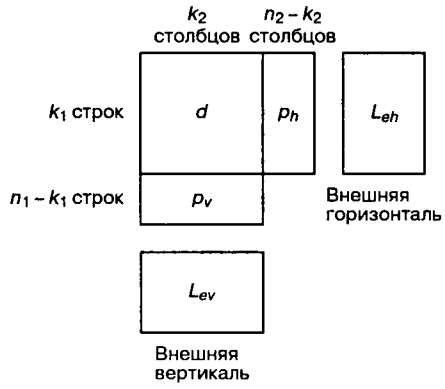


Рис. 8.22. Структура двухмерного композиционного кода

Еще на рис. 8.22 присутствуют блоки L_{eh} и L_{ev}, содержащие значения внешних LLR, полученные из горизонтального и вертикального кодов. Код с коррекцией ошибок дает некоторое улучшение достоверности передачи. Можно увидеть, что внешние LLR представляют собой меру этого улучшения. Заметьте, что такой композиционный

код является простым примером каскадного кода. Его структура описывается двумя отдельными этапами кодирования — горизонтальным и вертикальным.

Напомним, что решение при финальном декодировании каждого бита и его надежности зависит от значения $L(\hat{d})$, как показывает уравнение (8.71). Опираясь на это уравнение, можно описать алгоритм, дающий внешние LLR (горизонтальное и вертикальное) и финальное $L(\hat{d})$. Для композиционного кода алгоритм такого итеративного декодирования будет иметь следующий вид.

1. Устанавливается априорное LLR $L(d) = 0$ (если априорные вероятности битов данных не равны).
2. Декодируется горизонтальный код и, основываясь на уравнении (8.71), вычисляется горизонтальное LLR.

$$L_{eh}(\hat{d}) = L(\hat{d}) - L_c(x) - L(d)$$

3. На этапе 4 вертикального декодирования устанавливается $L(d) = L_{eh}(\hat{d})$.
4. Декодируется вертикальный код и, основываясь на уравнении (8.71), вычисляется вертикальное LLR.

$$L_{ev}(\hat{d}) = L(\hat{d}) - L_c(x) - L(d)$$

5. Для этапа 2 горизонтального декодирования устанавливается $L(d) = L_{eh}(\hat{d})$. Затем повторяются этапы 2–5.
6. После достаточного для получения надежного решения количества итераций (т.е. повторения этапов 2–5) следует перейти к этапу 7.
7. Мягким решением на выходе будет

$$L(\hat{d}) = L_c(x) + L_{eh}(\hat{d}) + L_{ev}(\hat{d}) \tag{8.74}$$

Далее следует пример, демонстрирующий применение этого алгоритма к очень простому композиционному коду.

8.4.3.1. Пример двухмерного кода с одним разрядом контроля четности

Пусть в кодере биты данных и биты контроля четности имеют значения, показанные на рис. 8.23, *а*. Связь между битами данных и битами контроля четности внутри конкретной строки (или столбца) выражается через двоичные цифры (1, 0) следующим образом.

$d_1 = 1$	$d_2 = 0$	$p_{12} = 1$
$d_3 = 0$	$d_4 = 1$	$p_{34} = 1$
$p_{13} = 1$	$p_{24} = 1$	

а) выходные двоичные цифры кодера

$L_c(x_1) = 1{,}5$	$L_c(x_2) = 0{,}1$	$L_c(x_{12}) = 2{,}5$
$L_c(x_3) = 0{,}2$	$L_c(x_4) = 0{,}3$	$L_c(x_{34}) = 2{,}0$
$L_c(x_{13}) = 6{,}0$	$L_c(x_{24}) = 1{,}0$	

б) логарифмическое отношение правдоподобий на входе декодера, $L_c(x)$

Рис. 8.23. Пример композиционного кода

$$d_i \oplus d_j = p_{ij} \tag{8.75}$$

и

$$d_i = d_j \oplus p_{ij} \quad i,j \in \{(1,2),\ (3,4),\ (1,3),\ (2,4)\} \tag{8.76}$$

Здесь символ "⊕" обозначает сумму по модулю 2. Переданные биты представлены последовательностью d_1, d_2, d_3, d_4, p_{12}, p_{34}, p_{13}, p_{24}. На входе приемника искаженные помехами биты представляются последовательностью $\{x_i\}$, $\{x_{ij}\}$. В данной ситуации для каждого принятого бита данных $x_i = d_i + n$, для каждого принятого бита контроля четности $x_{ij} = p_{ij} + n$, а n представляет собой распределение помех, которое статистически независимо от d_i и p_{ij}. Индексы i и j обозначают позицию в выходном массиве кодера, изображенном на рис. 8.23, *а*. Хотя зачастую удобнее использовать обозначение принятой последовательности в виде $\{x_k\}$, где k является временным индексом. Оба типа обозначений будут рассматриваться далее; i и j используются для позиционных отношений внутри композиционного кода, а k — для более общих аспектов временной зависимости сигнала. Какое из обозначений должно быть заметно по контексту? Если основываться на отношениях, установленных в уравнениях (8.67)–(8.69), и считать модель каналом AWGN с помехами, LLR для канальных измерений сигнала x_k, принятого в момент k, будет иметь следующий вид.

$$L_c(x_k) = \ln\left[\frac{p(x_k|d_k = +1)}{p(x_k|d_k = -1)}\right] = \tag{8.77,а}$$

$$= \ln\left(\frac{\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left[-\dfrac{1}{2}\left(\dfrac{x_k - 1}{\sigma}\right)^2\right]}{\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left[-\dfrac{1}{2}\left(\dfrac{x_k + 1}{\sigma}\right)^2\right]}\right) = \tag{8.77,б}$$

$$= -\frac{1}{2}\left(\frac{x_k - 1}{\sigma}\right)^2 + \frac{1}{2}\left(\frac{x_k + 1}{\sigma}\right)^2 = \frac{2}{\sigma^2}x_k \tag{8.77,в}$$

Здесь применяется натуральный логарифм. Если сделать предварительное допущение, что помеха имеет дисперсию σ_2, равную 1, то получим следующее.

$$L_c(x_k) = 2x_k \tag{8.78}$$

Рассмотрим пример, в котором информационная последовательность d_1, d_2, d_3, d_4 образована двоичными числами 1 0 0 1, как показано на рис. 8.23, *а*. Опираясь на уравнение (8.75), можно видеть, что контрольная последовательность p_{12}, p_{34}, p_{13}, p_{24} должна быть равна 1 1 1 1. Следовательно, переданная последовательность будет иметь следующий вид.

$$\{d_i\}, \{p_{ij}\} = 1\,0\,0\,1\,1\,1\,1\,1 \tag{8.79}$$

Если информационные биты выражаются через значения биполярного электрического напряжения +1 и –1, соответствующие логическим двоичным уровням 1 и 0, то переданная последовательность будет следующей.

$$\{d_i\}, \{p_{ij}\} = +1, -1, -1, +1, +1, +1, +1, +1$$

Допустим теперь, что помехи преобразуют эту последовательность информации и контрольных данных в принятую последовательность

$$\{x_i\}, \{x_{ij}\} = 0{,}75, 0{,}05, 0{,}10, 0{,}15, 1{,}25, 1{,}0, 3{,}0, 0{,}5, \tag{8.80}$$

где компоненты $\{x_i\}$, $\{x_{ij}\}$ указывают переданную информацию и контрольные данные $\{d_i\}$, $\{p_{ij}\}$. Таким образом, следуя позиционному описанию, принятую последовательность можно записать следующим образом.

$$\{x_i\}, \{x_{ij}\} = x_1, x_2, x_3, x_4, x_{12}, x_{34}, x_{13}, x_{24}$$

Из уравнения (8.78) предполагаемые канальные измерения дают следующие значения LLR.

$$\{L_c(x_i)\}, \{L_c(x_{ij})\} = 1{,}5, 0{,}1, 0{,}20, 0{,}3, 2{,}5, 2{,}0, 6{,}0, 1{,}0 \tag{8.81}$$

Эти величины показаны на рис. 8.23, *б* как входные измерения декодера. Следует заметить, что (при равной априорной вероятности переданных данных) если принимаются жесткие решения на основе значений $\{x_k\}$ или $\{L_c(x_k)\}$, описанных ранее, то такой процесс должен в результате давать две ошибки, поскольку и d_2, и d_3 могут быть неправильно трактованы как двоичная 1.

8.4.3.2. Внешние правдоподобия

В случае композиционного кода, изображенного на рис. 8.23, при выражении мягкого выхода для принятого сигнала, соответствующего данным d_1, используется уравнение (8.71), так что

$$L(\hat{d}_1) = L_c(x_1) + L(d_1) + \{[L_c(x_2) + L(d_2)] \boxplus L_c(x_{12})\}, \tag{8.82}$$

где члены $\{[L_c(x_2) + L(d_2)] \boxplus L_c(x_{12})\}$ представляют внешнее LLR, распределенное кодом (т.е. прием соответствующих данных d_2 и их априорной вероятности совместно с приемом соответствующей четности p_{12}). В общем случае мягким выходом $L(\hat{d}_i)$ для принятого сигнала, соответствующего данным d_i, будет

$$L(\hat{d}_i) = L_c(x_i) + L(d_i) + \{[L_c(x_j) + L(d_j)] \boxplus L_c(x_{ij})\}, \tag{8.83}$$

где $L_c(x_i)$, $L_c(x_j)$ и $L_c(x_{ij})$ — канальное измерение LLR приема соответствующих d_i, d_j и p_{ij}. $L(d_i)$, $L(d_j)$ — LLR для априорных вероятностей d_i и d_j, $\{[L_c(x_j) + L(d_j)] \boxplus L_c(x_{ij})\}$ — внешнее распределение LLR для кода. Уравнения (8.82) и (8.83) становятся понятнее при рассмотрении рис. 8.23, *б*. В данной ситуации, если считать, что происходит равновероятная передача сигнала, мягкий выход $L(\hat{d}_1)$ представляется измерением LLR детектора $L_c(x_1) = 1{,}5$ для приема, соответствующего данным d_1, плюс внешнее LLR $[L_c(x_2) + L(d_2) \boxplus L_c(x_{12}) = 2{,}5]$, получаемое в результате того, что данные d_2 и четность p_{12} также дают сведения о данных d_1, как это показывают уравнения (8.75) и (8.76).

8.4.3.3. Вычисление внешних правдоподобий

Для случая, показанного на рис. 8.23, горизонтальная часть расчетов для получения $L_{eh}(\hat{d})$ и вертикальная часть расчетов для получения $L_{ev}(\hat{d})$ выглядят следующим образом.

$$L_{eh}(\hat{d}_1) = [L_c(x_2) + L(d_2)] \boxplus L_c(x_{12}) \tag{8.84,а}$$

$$L_{ev}(\hat{d}_1) = [L_c(x_3) + L(d_3)] \boxplus L_c(x_{13}) \tag{8.84,б}$$

$$L_{eh}(\hat{d}_2) = [L_c(x_1) + L(d_1)] \boxplus L_c(x_{12}) \tag{8.85,а}$$

$$L_{ev}(\hat{d}_2) = [L_c(x_4) + L(d_4)] \boxplus L_c(x_{24}) \tag{8.85,б}$$

$$L_{eh}(\hat{d}_3) = [L_c(x_4) + L(d_4)] \boxplus L_c(x_{34}) \tag{8.86,а}$$

$$L_{ev}(\hat{d}_3) = [L_c(x_1) + L(d_1)] \boxplus L_c(x_{13}) \tag{8.86,б}$$

$$L_{eh}(\hat{d}_4) = [L_c(x_3) + L(d_3)] \boxplus L_c(x_{34}) \tag{8.87,а}$$

$$L_{ev}(\hat{d}_4) = [L_c(x_2) + L(d_2)] \boxplus L_c(x_{24}) \tag{8.87,б}$$

Значения LLR, показанные на рис. 8.23, входят в выражение для $L_{eh}(\hat{d})$ в уравнениях (8.84)–(8.87). Подразумевая передачу сигналов равновероятной, а начальную установку значения $L(d)$ равной нулю, получаем следующее.

$$L_{eh}(\hat{d}_1) = [0{,}1 + 0] \boxplus 2{,}5 \approx -0{,}1 \text{ — новое } L(d_1) \tag{8.88}$$

$$L_{eh}(\hat{d}_2) = [1{,}5 + 0] \boxplus 2{,}5 \approx -1{,}5 \text{ — новое } L(d_1) \tag{8.89}$$

$$L_{eh}(\hat{d}_3) = [0{,}3 + 0] \boxplus 2{,}0 \approx -0{,}3 \text{ — новое } L(d_1) \tag{8.90}$$

$$L_{eh}(\hat{d}_4) = [0{,}2 + 0] \boxplus 2{,}0 \approx -0{,}2 \text{ — новое } L(d_1), \tag{8.91}$$

где сложения логарифма правдоподобия производятся, исходя из приближения, показанного в уравнении (8.73). Далее, продолжая первое вертикальное вычисление, используются выражения для $L_{ev}(\hat{d})$ из уравнений (8.84)–(8.87). Теперь значение $L(d)$ можно обновить, исходя из нового значения $L(d)$, полученного из первого вертикального вычисления, показанного в уравнениях (8.88)–(8.91).

$$L_{ev}(\hat{d}_1) = [0{,}2 - 0{,}3] \boxplus 6{,}0 \approx 0{,}1 \text{ — новое } L(d_1) \tag{8.92}$$

$$L_{ev}(\hat{d}_2) = [0{,}3 - 0{,}2] \boxplus 1{,}0 \approx -0{,}1 \text{ — новое } L(d_2) \tag{8.93}$$

$$L_{ev}(\hat{d}_3) = [1{,}5 - 0{,}1] \boxplus 6{,}0 \approx -1{,}4 \text{ — новое } L(d_3) \tag{8.94}$$

$$L_{ev}(\hat{d}_4) = [0{,}1 - 1{,}5] \boxplus 1{,}0 \approx 1{,}0 \text{ — новое } L(d_4) \tag{8.95}$$

Результаты первой полной итерации двух этапов декодирования (горизонтального и вертикального) будут следующими.

Исходные измерения $L_c(x_k)$

1,5	0,1
0,2	0,3

–0,1	–1,5
–0,3	–0,2

$L_{ev}(\hat{d})$ после первого горизонтального декодирования

0,1	–0,1
–1,4	1,0

$L_{ev}(\hat{d})$ после первого вертикального декодирования

Каждый этап декодирования улучшает исходные LLR, которые основываются только на канальных измерениях. Это видно из расчетов выходного LLR декодера с помощью уравнения (8.74). Исходное LLR и внешние горизонтальные LLR вместе дают следующее улучшение (внешний вертикальный член еще не рассматривался).

Улучшение LLR из-за $L_{eh}(\hat{d})$

1,4	–1,4
–0,1	0,1

Исходное LLR совместно с горизонтальным и вертикальным внешним LLR дает следующее улучшение.

Улучшение LLR из-за $L_{eh}(\hat{d}) + L_{ev}(\hat{d})$

1,5	–1,5
–1,5	1,1

В данном случае можно видеть, что сведений, полученных лишь из горизонтального декодирования, достаточно для получения правильного жесткого решения вне декодера, но с низкой степенью доверия к битам данных d_3 и d_4. После включения внешних вертикальных LLR в декодер новые значения LLR появляются на более высоком уровне надежности и доверия. Пусть будет произведена еще одна вертикальная и одна горизонтальная итерация декодирования, чтобы определить наличие или отсутствие существенных изменений в результатах. Снова на помощь приходят отношения из уравнений (8.84)–(8.87), и далее следует горизонтальное вычисление для получения $L_{eh}(\hat{d})$ с новым $L(d)$ из первого вертикального расчета, показанного в уравнениях (8.92)–(8.95), так что получаем следующее.

$$L_{eh}(\hat{d}_1) = [0{,}1 - 0{,}1] \boxplus 2{,}5 \approx 0 \quad \text{— новое } L(d_1) \tag{8.96}$$

$$L_{eh}(\hat{d}_2) = [1{,}5 - 0{,}1] \boxplus 2{,}5 \approx -1{,}6 \quad \text{— новое } L(d_2) \tag{8.97}$$

$$L_{eh}(\hat{d}_3) = [0{,}3 - 1{,}0] \boxplus 2{,}0 \approx -1{,}3 \quad \text{— новое } L(d_3) \tag{8.98}$$

$$L_{eh}(\hat{d}_4) = [0{,}2 - 1{,}4] \boxplus 2{,}0 \approx 1{,}2 \text{ — новое } L(d_4) \tag{8.99}$$

Затем необходимо выполнить второе вертикальное вычисление для получения $L_{ev}(\hat{d})$ с новым $L(d)$, полученным из второго горизонтального расчета, показанного в уравнениях (8.96)–(8.99), что приводит к следующему.

$$L_{ev}(\hat{d}_1) = [0{,}2 - 1{,}3] \boxplus 6{,}0 \approx 1{,}1 \text{ — новое } L(d_1) \tag{8.100}$$

$$L_{ev}(\hat{d}_2) = [0{,}3 + 1{,}2] \boxplus 1{,}0 \approx -1{,}0 \text{ — новое } L(d_2) \tag{8.101}$$

$$L_{ev}(\hat{d}_3) = [1{,}5 + 0] \boxplus 6{,}0 \approx -1{,}5 \text{ — новое } L(d_3) \tag{8.102}$$

$$L_{ev}(\hat{d}_4) = [0{,}1 - 1{,}6] \boxplus 1{,}0 \approx 1{,}0 \text{ — новое } L(d_4) \tag{8.103}$$

Вторая итерация вертикального и горизонтального декодирования, дающая упомянутые выше величины, отражается на мягких выходных LLR, которые снова рассчитываются из уравнения (8.74), переписанного следующим образом.

$$L(\hat{d}) = L_c(x) + L_{eh}(\hat{d}) + L_{ev}(\hat{d}) \tag{8.104}$$

Горизонтальные и вертикальные LLR из уравнений (8.96)–(8.103) и итоговое LLR декодера показаны ниже. В данном примере вторые итерации, горизонтальная и вертикальная, что в целом дает всего четыре итерации, показывают скромный прирост, по сравнению с одной вертикальной и горизонтальной итерацией. Результаты показывают, что доверительные значения сохраняются для каждого из четырех данных.

Исходные измерения $L_c(x)$

1,5	0,1
0,2	0,3

0	–1,6
–1,3	1,2

$L_{ev}(\hat{d})$ после второго вертикального декодирования

1,1	–1,0
–1,5	1,0

$L_{eh}(\hat{d})$ после второго горизонтального декодирования

Мягкий выход равен $L(\hat{d}) = L_c(x) + L_{eh}(\hat{d}) + L_{ev}(\hat{d})$, который после всех четырех итераций дает следующие значения $L(\hat{d})$.

2,6	–2,5
–2,6	2,5

В результате видно, что получены правильные решения по каждому биту данных и уровень доверия к этим решениям высок. Итеративное декодирование турбокодов напоминает процесс решения кроссвордов. Первый проход по кроссворду, вероятно, содержит несколько ошибок. Некоторые слова нуждаются в подгонке, но когда буквы в нужных строках и столбцах не подходят, нужно вернуться и исправить слова, вписанные после первого прохода.

8.4.4. Кодирование с помощью рекурсивного систематического кода

Ранее были описаны основные идеи сочетаний, итераций и мягкого декодирования на примере простого композиционного кода. Затем эти идеи применялись при реализации турбокодов, которые образуются в результате параллельных сочетаний сверточных кодов [17, 20].

Далее наступает очередь обзора простых двоичных сверточных кодеров со степенью кодирования 1/2, длиной кодового ограничения K и памятью порядка $K-1$. На вход кодера в момент k, подается бит d_k, и соответствующим кодовым словом будет битовая пара (u_k, v_k), где

$$u_k = \sum_{i=0}^{K-1} g_{1i}\, d_{k-i} \text{ по модулю 2, } g_{1i} = 0, 1 \tag{8.105}$$

и

$$v_k = \sum_{i=0}^{K-1} g_{2i}\, d_{k-i} \text{ по модулю 2, } g_{2i} = 0, 1 \tag{8.106}$$

$\mathbf{G}_1 = \{g_{1i}\}$ и $\mathbf{G}_2 = \{g_{2i}\}$ — генераторы кода, а d_k представлен как двоичная цифра. Этот кодер можно представить как линейную систему с дискретной конечной импульсной характеристикой (finite impulse response — FIR), порождающую хорошо знакомый несистематический сверточный (nonsystematic convolutional — NSC) код, разновидность которого показана на рис. 8.24. Соответствующую решетчатую структуру можно увидеть на рис. 7.7. В данном случае длина кодового ограничения равна $K = 3$ и используются два генератора кода — $\mathbf{G}_1 = \{111\}$ и $\mathbf{G}_2 = \{101\}$. Хорошо известно, что при больших значениях E_b/N_0 достоверность передачи с кодом NSC выше, чем у систематического кода с той же памятью. При малых значениях E_b/N_0 существует обходной путь [17]. В качестве составляющих компонентов для турбокода был предложен класс сверточных кодов с бесконечной импульсной характеристикой [17]. Такие же компоненты используются в рекурсивных систематических сверточных (recursive systematic convolutional — RSC) кодах, поскольку в них предварительно кодированные биты данных постоянно должны подаваться обратно на вход кодера. При высоких степенях кодирования коды RSC дают значительно более высокие результаты, чем самые лучшие коды NSC, при любых значениях E_b/N_0. Двоичный код RSC со степенью кодирования 1/2 получается из кода NSC с помощью контура обратной связи и установки одного из двух выходов (u_k или v_k) равным d_k. На рис. 8.25, *а* показан пример такого RSC-кода с $K = 3$, где a_k получается из рекурсивной процедуры

$$a_k = d_k + \sum_{i=0}^{K-1} g_i' a_{k-i} \text{ по модулю 2}, \qquad (8.107)$$

а g_i' равно g_{1i}, если $u_k = d_k$, и g_{2i} — если $v_k = d_k$. На рис. 8.25, *б* изображена решетчатая структура RSC-кода, представленного на рис. 8.25, *а*.

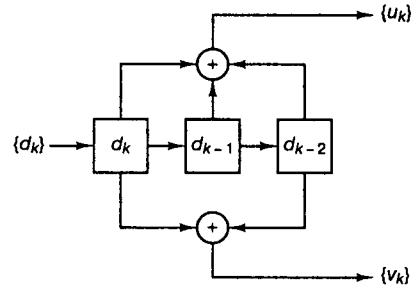


Рис. 8.24. Несистематический сверточный код (nonsystematic convolutional — NSC)

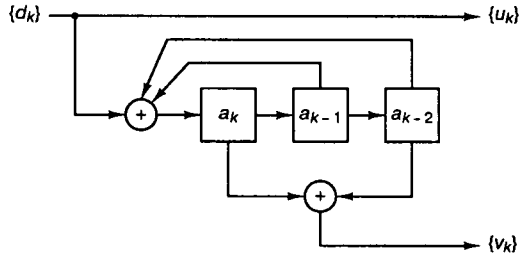


Рис. 8.25а. Рекурсивный систематический сверточный код (recursive systematic convolutional — RSC)

Рис. 8.25б. Решетчатая структура RSC-кода, представленного на рис. 8.25, а

Считается, что входной бит с одинаковой вероятностью может принимать как значение 1, так и 0. Кроме того, $\{a_k\}$ показывает те же статистические вероятности, что и $\{d_k\}$ [17]. Просвет одинаков у RSC-кода (рис. 8.25, *а*) и NSC-кода (рис. 8.24). Точно так же совпадает их решетчатая структура по отношению к переходам между состояниями и соответствующим входным битам. Впрочем, у RSC- и NSC-кодов две выход-

ные последовательности $\{u_k\}$ и $\{v_k\}$ не соответствуют той же входной последовательности $\{d_k\}$. Можно сказать, что при тех же генераторах кода распределение весовых коэффициентов кодовых слов RSC-кодера не изменяется, по сравнению с распределением весовых коэффициентов кодовых слов NSC-кодера. Единственное различие состоит в отображении между входной и выходной последовательностями данных.

Пример 8.5. Рекурсивные кодеры и их решетчатые диаграммы

а) Используя RSC-кодер (рис. 8.25, *а*), проверьте справедливость участка решетчатой структуры (диаграммы), изображенного на рис. 8.25, *б*.

б) Для кодера, указанного в п. а, начиная с последовательности данных $\{d_k\} = 1\ 1\ 1\ 0$, поэтапно покажите процедуру кодирования до нахождения выходного кодового слова.

Решение

а) Для кодеров NSC содержимое регистра и переходы между состояниями отслеживаются непосредственно. Но если кодер является рекурсивным, следует быть очень аккуратным. В табл. 8.5 содержится 8 строк, соответствующих 8 возможным переходам в данной системе, образованной из 4-х состояний. Первые четыре строки представляют переходы, когда входной информационный бит d_k является двоичным нулем, а последние четыре — переходы, в которых d_k является единицей. В данном случае процедуру кодирования с помощью табл. 8.5 и рис. 8.25 можно поэтапно описать следующим образом.

1. В момент введения произвольного входного бита, k, состояние перед переходом (начальное) определяется содержимым двух крайних разрядов регистра, а именно — a_{k-1} и a_{k-2}.
2. В любой строке таблицы (переход на решетке) поиск содержимого разряда a_k выполняется сложением (по модулю 2) битов d_k, a_{k-1} и a_{k-2} в этой строке.
3. Выходная кодовая последовательность битов, $u_k v_k$, для каждого возможного начального состояния (т.е. $a = 00$, $b = 10$, $c = 01$ и $d = 11$) находится путем прибавления (по модулю 2) a_k и a_{k-2} к $d_k = u_k$.

Таблица 8.5. Проверка участка решетки с рис. 8.25, *б*

Входной бит	Текущий бит	Начальное состояние		Кодовые биты		Конечное состояние	
$d_k = u_k$	a_k	a_{k-1}	a_{k-2}	u_k	v_k	a_k	a_{k-1}
	0	0	0	0	0	0	0
	1	1	0	0	1	1	1
0	1	0	1	0	0	1	0
	0	1	1	0	1	0	1
	1	0	0	1	1	1	0
	0	1	0	1	0	0	1
1	0	0	1	1	1	0	0
	1	1	1	1	0	1	1

Нетрудно убедиться, что элементы табл. 8.5 соответствуют участку решетки, изображенному на рис. 8.25, *б*. При использовании для реализации составных кодов регистров сдвига у турбокодеров проявляется интересное свойство, которое заключается в том, что два перехода, входящие в состояние, *не соответствуют* одному и тому же входному битовому значению (т.е. в данное состояние не входят две сплошные или две пунктирные линии). Это свойство проявляется, если полиномиальное описание обратной связи регистра сдвига имеет все порядки или одна из линий обратной связи выходит из разряда более высокого порядка, в данном случае a_{k-2}.

б) Существует два способа реализации кодирования входной информационной последовательности $\{d_k\} = 1\ 1\ 1\ 0$. Первый состоит в применении решетчатой диаграммы, а другой — в использовании цепи кодера. Воспользовавшись участком решетки, изображенным на рис. 8.25, *б*, мы выбираем переход по пунктирной линии (представляющий двоичную единицу) из состояния $a = 00$ (естественный выбор начального состояния) в следующее состояние $b = 10$ (которое подходит в качестве стартового для следующего входного бита). Следует заметить, что биты показаны на этом переходе как выходная кодовая последовательность 11. Эта процедура повторяется для каждого входного бита. Другой способ предполагает построение таблицы, такой как табл. 8.6, на основе цепи кодера, изображенной на рис. 8.25, *а*. Здесь время k показано от начала до конца всей процедуры (5 моментов времени и 4 временных интервала). Табл. 8.6 записывается в следующем порядке.

1. В произвольный момент времени бит данных d_k начинает преобразовываться в a_k путем суммирования его (по модулю 2) с битами a_{k-1} и a_{k-2} в той же строке.
2. Например, в момент времени $k = 2$ бит данных $d_k = 1$ преобразуется в $a_k = 0$ путем суммирования его с битами a_{k-1} и a_{k-2} в той же строке $k = 2$.
3. Итоговый выход $u_k v_k = 10$, определяемый логической схемой кодера, является кодовой битовой последовательностью, связанной со временем $k = 2$ (в действительности — интервалом между $k = 2$ и $k = 3$).
4. В момент $k = 2$ содержимое крайних правых разрядов $a_{k-1}\, a_{k-2}$ (10) представляет собой состояние системы в начале этого перехода.
5. Конечное состояние этого перехода представляется содержимым двух крайних левых регистров $a_k\, a_{k-1}$ в той же строке (01). Поскольку сдвиг битов происходит слева направо, это конечное состояние перехода в момент $k = 3$ будет представлено как стартовое в следующей строке.
6. Каждая строка описывается аналогично. Таким образом, в последнем столбце табл. 8.6 можно будет увидеть кодированную последовательность 1 1 1 0 1 1 0 0.

Таблица 8.6. Кодирование битовой последовательности с помощью кодера, изображенного на рис. 8.25, *а*

Время	Входной бит	Первый разряд	Состояние в момент k		Кодовые биты	
k	$d_k = u_k$	a_k	a_{k-1}	a_{k-2}	u_k	v_k
1	1	1	0	0	1	1
2	1	0	1	0	1	0
3	1	0	0	1	1	1
4	0	0	0	0	0	0
5			0	0		

8.4.4.1. Конкатенация кодов RSC

Рассмотрим параллельную конкатенацию двух RSC-кодеров, подобных изображенному на рис. 8.25. Хороший турбокод строится из составных кодов с небольшой длиной кодового ограничения ($K = 3$–5). В качестве примера такого турбокодера можно взять кодер, показанный на рис. 8.26, где переключатель v_k делает степень кодирования всего кода равной 1/2. Без переключателя степень кодирования кода будет равна 1/3. Ограничений на количество соединяемых кодеров нет. Составные коды должны иметь одинаковую длину кодового ограничения и степень кодирования. Целью

разработки турбокода является наилучший подбор составных кодов путем минимизации просвета кода [21]. При больших значениях E_b/N_0 это эквивалентно максимизации минимального весового коэффициента кодовых слов. Хотя при низких значениях E_b/N_0 (область, представляющая наибольший интерес) оптимизация распределения весовых коэффициентов кодовых слов является более важной, чем их максимизация или минимизация [20].

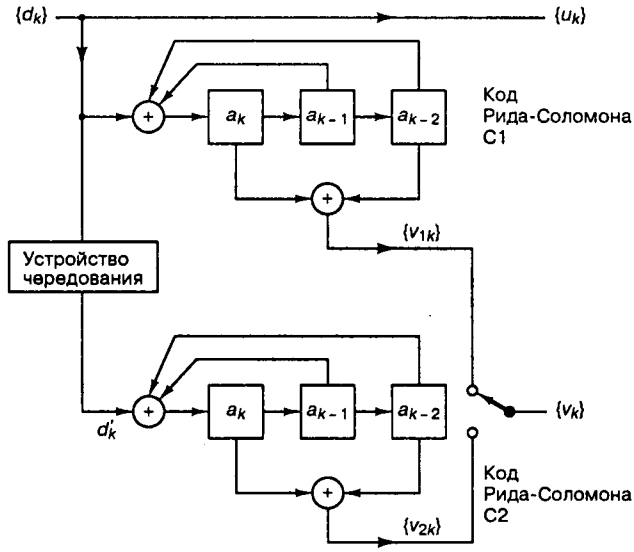


Рис. 8.26. Схема параллельного соединения двух RSC-кодеров

Турбокодер, изображенный на рис. 8.26, выдает кодовые слова из каждого из двух своих составных кодеров. Распределение весовых коэффициентов кодовых слов без такого параллельного соединения зависит от того, сколько кодовых слов из одного составного кодера комбинируется с кодовыми словами из другого составного кодера.

Интуитивно понятно, что следует избегать спаривания кодовых слов с малым весовым коэффициентом из одного кодера с кодовыми словами с малым весовым коэффициентом из другого кодера. Большого количества таких спариваний можно избежать, сконфигурировав надлежащим образом устройство чередования. Устройство, которое обрабатывает данные случайным образом, более эффективно, чем рассмотренное ранее блочное устройство чередования [22].

Если составной кодер не рекурсивный, входная последовательность с единичным весовым коэффициентом (0 0 ... 0 0 1 0 0 ... 0 0) всегда будет генерировать кодовое слово с малым весовым коэффициентом на входе второго кодера, при любой конструкции устройства чередования. Иначе говоря, устройство чередования не сможет повлиять на выходное распределение весовых коэффициентов кодовых слов, если составные коды не рекурсивные. Впрочем, если составные коды рекурсивные, входная последовательность с единичным весовым коэффициентом генерирует бесконечную импульсную характеристику (выход с бесконечным весовым коэффициентом). Следовательно, при рекурсивных кодах входная последовательность с единичным весовым коэффициентом не дает кодового слова с минимальным весовым коэффициентом вне кодера. Кодированный выходной весовой коэффициент сохраняется конечным только

при погашении решетки, процессе, принуждающем кодированную последовательность к переходу в конечное состояние таким образом, что кодер возвращается к нулевому состоянию. Фактически сверточный код преобразуется в блочный.

Для кодера, изображенного на рис. 8.26, кодовое слово с минимальным весовым коэффициентом для каждого составного кодера порождается входной последовательностью с весовым коэффициентом 3 (0 0 ... 0 0 1 1 1 0 0 0 ... 0 0) и тремя последовательными единицами. Другая последовательность, порождающая кодовые слова с малым весом, представлена последовательностью с весом 2 (0 0 ... 0 0 1 0 0 1 0 0 ... 0 0). Однако после перестановок, внесенных устройством чередования, любая из этих опасных структур имеет слабую вероятность появления на входе второго кодера, что делает маловероятной возможность комбинирования одного кодового слова с малым весом с другим кодовым словом с малым весом.

Важным аспектом компоновки турбокодов является их рекурсивность (систематический аспект незначителен). Это бесконечная импульсная характеристика, присущая кодам RSC, которая является защитой от генерации кодовых слов с малым весом, не поддающихся исправлению в устройстве чередования. Можно обсудить то, что производительность турбокодера сильно поддается влиянию со стороны кодовых слов с малым весом, производимых входной последовательностью с весом 2. В защиту этого можно сказать, что входную последовательность с весом 1 можно проигнорировать, поскольку она дает кодовые слова с большим весом из-за бесконечной импульсной характеристики кодера. Для входной последовательности, имеющей вес 3 и более, правильно сконфигурированное устройство чередования делает вероятность появления кодовых слов с малым весом на выходе относительно низкой [21–25].

8.4.5. Декодер с обратной связью

Использование алгоритма Витерби является оптимальным методом декодирования для минимизации вероятности появления ошибочной последовательности. К сожалению, этот алгоритм (с жесткой схемой на выходе) не подходит для генерации апостериорной вероятности (a posteriori probability — APP) или мягкой схемы на выходе для каждого декодированного бита. Подходящий для этой задачи алгоритм был предложен Балом и др. [26]. Алгоритм Бала был модифицирован Берру и др. [17] для использования в кодах RSC. Апостериорную вероятность того, что декодированный бит данных $d_k = i$, можно вывести из совместной вероятности $\lambda_k^{i,m}$, определяемой как

$$\lambda_k^{i,m} = P\{d_k = i, S_k = m | R_1^N\}, \tag{8.108}$$

где $S_k = m$ — состояние кодера в момент времени k, а R_1^N — принятая двоичная последовательность за время от $k = 1$ в течение некоторого времени N.

Таком образом, апостериорная вероятность того, что декодированный информационный бит $d_k = i$ представляется как двоичная цифра, получается путем суммирования совокупных вероятностей по всем состояниям.

$$P\{d_k = i | R_1^N\} = \sum_m \lambda_k^{i,m} \quad i = 0, 1 \tag{8.109}$$

Далее логарифмическое отношение правдоподобий (log-likelihood ratio — LLR) переписывается как логарифм отношения апостериорных вероятностей.

$$L(\hat{d}_k) = \lg\left[\frac{\sum_m \lambda_k^{1,m}}{\sum_m \lambda_k^{0,m}}\right] \tag{8.110}$$

Декодер осуществляет схему решений, известную как решающее правило *максимума апостериорной вероятности* (maximum a posteriori — MAP), путем сравнения $L(\hat{d}_k)$ с нулевым пороговым значением.

$$\begin{aligned} \hat{d}_k &= 1, \quad \text{если} \quad L(\hat{d}_k) > 0 \\ \hat{d}_k &= 0, \quad \text{если} \quad L(\hat{d}_k) < 0 \end{aligned} \tag{8.111}$$

Для систематического кода LLR $L(\hat{d}_k)$, связанное с каждым декодированным битом $\hat{d}_k$, можно описать как сумму LLR для $\hat{d}_k$ вне демодулятора и других LLR, порождаемых декодером (внешние сведения), как показано уравнениями (8.72) и (8.73). Рассмотрим обнаружение последовательности данных с помехами, исходящей из кодера, изображенного на рис. 8.26, с помощью декодера, представленного на рис. 8.27. Предполагается, что используется двоичная модуляция и дискретный гауссов канал без памяти. Вход декодера формируется набором R_k из двух случайных переменных x_k и y_k. Для битов d_k и v_k, которые в момент времени k представляются двоичными числами (1, 0), переход к принятым биполярным импульсам (+1, –1) можно записать следующим образом.

$$x_k = (2d_k - 1) + i_k \tag{8.112}$$

и

$$y_k = (2v_k - 1) + q_k \tag{8.113}$$

Здесь i_k и q_k являются двумя случайными статистически независимыми переменными с одинаковой дисперсией σ_2, определяющей распределение помех. Избыточная информация y_k разуплотняется и пересылается на декодер DEC1 как y_{1k}, если $v_k = v_{1k}$, и на декодер DEC2 как y_{2k}, если $v_k = v_{2k}$. Если избыточная информация начальным декодером не передается, то вход соответствующего декодера устанавливается на нуль. Следует отметить, что выход декодера DEC1 имеет структуру чередования, аналогичную структуре, использованной в передатчике между двумя составными кодерами. Это связано с тем, что информация, обрабатываемая декодером DEC1, является неизмененным выходом кодера C1 (искаженной канальным шумом). И наоборот, информация, обрабатываемая декодером DEC2, является искаженным выходом кодера C2, вход которого составляют как раз те данные, что поступают в C1, но обработаны устройством чередования. Декодер DEC2 пользуется выходом декодера DEC1, обеспечивая такое же временное упорядочение этого выхода, как и входа C2 (т.е. две последовательности в декодере DEC2 должны придерживаться позиционной структуры сигналов в каждой последовательности).

8.4.5.1. Декодирование при наличии контура обратной связи

Уравнение (8.71) можно переписать для мягкого выхода в момент времени k с нулевой начальной установкой априорного LLR $L(d_k)$. Это делается на основе предположения о равной вероятности информационных битов. Следовательно,

$$L(\hat{d}_k) = L_c(x_k) + L_e(\hat{d}_k) = \\ = \lg\left(\frac{p(x_k|d_k=1)}{p(x_k|d_k=0)}\right) + L_e(\hat{d}_k), \quad (8.114)$$

где $L(\hat{d}_k)$ — мягкий выход декодера, а $L_c(x_k)$ — LLR канального измерения, получаемый из отношения функций правдоподобия $p(x_k|d_k=i)$, связанных с моделью дискретного канала без памяти. $L_e(\hat{d}_k) = L(\hat{d}_k)\big|_{x_k=0}$ является функцией избыточной информации. Это внешние сведения, получаемые декодером и не зависящие от входных данных x_k декодера. В идеале $L_c(x_k)$ и $L_e(\hat{d}_k)$ искажаются некоррелированным шумом, а следовательно, $L_e(\hat{d}_k)$ может использоваться как новое наблюдение d_k другим декодером для образования итеративного процесса. Основным принципом передачи информации обратно на другой декодер является то, что декодер никогда не следует заполнять собственными данными (иначе искажения на входе и выходе будут сильно коррелировать).

Для гауссового канала в уравнении (8.114) при описании канального LLR $L_c(x_k)$ использовался натуральный логарифм, как и в уравнении (8.77). Уравнение (8.77,в) можно переписать следующим образом.

$$L_c(x_k) = -\frac{1}{2}\left(\frac{x_k-1}{\sigma}\right)^2 + \frac{1}{2}\left(\frac{x_k+1}{\sigma}\right)^2 = \frac{2}{\sigma^2}x_k \quad (8.115)$$

Оба декодера, DEC1 и DEC2, используют модифицированный алгоритм Бала [26]. Если данные $L_1(\hat{d}_k)$ и y_k, подаваемые на вход декодера DEC2 (рис. 8.27), являются статистически независимыми, то LLR $L_2(\hat{d}_k)$ на выходе декодера DEC2 можно переписать как

$$L_2(\hat{d}_k) = f[L_1(\hat{d}_k)] + L_{e2}(\hat{d}_k) \quad (8.116)$$

при

$$L_1(\hat{d}_k) = \frac{2}{\sigma_0^2}x_k + L_{e1}(\hat{d}_k), \quad (8.117)$$

где $f[\cdot]$ используется для выражения функциональной зависимости. Внешние сведения $L_{e2}(\hat{d}_k)$ вне декодера DEC2 являются функцией последовательности $\{L_1(\hat{d}_k)\}_{n\neq k}$. Поскольку $L_1(\hat{d}_k)$ зависит от наблюдения R_1^N, внешние сведения $L_{e2}(\hat{d}_k)$ коррелируют с наблюдениями x_k и y_{1k}. Тем не менее, чем больше значение $|n-k|$, тем меньше коррелируют $L_1(\hat{d}_k)$ и наблюдения x_k и y_{1k}. Вследствие чередования выходов декодеров DEC1 и DEC2, внешние сведения $L_{e2}(\hat{d}_k)$ слабо коррелируют с наблюдениями x_k и y_{1k}. Поэтому можно совместно использовать их для декодирования битов d_k [17]. На рис. 8.27 показана процедура подачи параметра $z_k = L_{e2}(\hat{d}_k)$ на де-

кодер DEC1 как эффект разнесения в итеративном процессе. Вообще, $L_{e2}(\hat{d}_k)$ имеет тот же знак, что и d_k. Следовательно, $L_{e2}(\hat{d}_k)$ может увеличить соответствующее LLR и, значит, повысить надежность каждого декодированного бита данных.

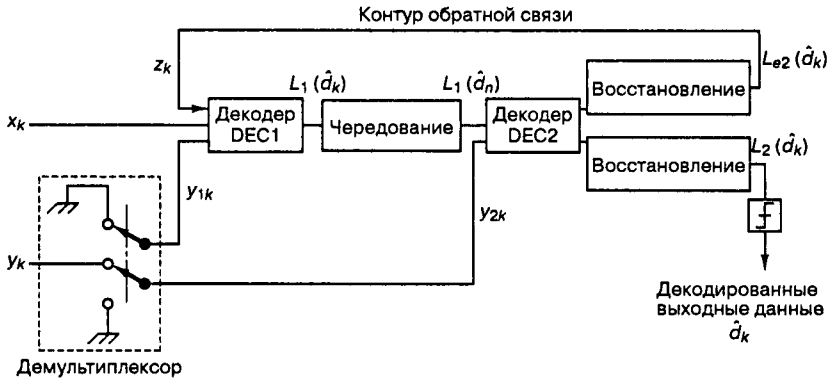


Рис. 8.27. Схема декодера с обратной связью

Подробное описание алгоритма вычисления LLR $L(\hat{d}_k)$ апостериорной вероятности каждого бита данных было представлено несколькими авторами [17, 18, 30]. В работах [27–31] были высказаны предположения относительно снижения конструктивной сложности алгоритмов. Приемлемый подход к представлению процесса, дающего значения апостериорной вероятности для каждого информационного бита, состоит в реализации оценки максимально правдоподобной последовательности, или алгоритма Витерби, и вычислении ее по двум направлениям блоков кодовых битов. Если осуществлять такой двунаправленный алгоритм Витерби по схеме раздвижных окон — получатся метрики, связанные с предшествующими и последующими состояниями. В результате получим апостериорную вероятность для каждого бита данных, имеющегося в блоке. Итак, декодирование турбокодов можно оценить как в два раза более сложное, чем декодирование одного из составных кодов с помощью алгоритма Витерби.

8.4.5.2. Достоверность передачи при турбокодировании

В [17] приведены результаты моделирования методом Монте-Карло кодера со степенью кодирования 1/2, $K = 5$, построенного на генераторах $\mathbf{G}_1 = \{1\ 1\ 1\ 1\ 1\}$ и $\mathbf{G}_2 = \{1\ 0\ 0\ 0\ 1\}$, при параллельном соединении и использовании устройства чередования с массивом 256×256. Был использован модифицированный алгоритм Бала и блок, длиной 65536 бит. После 18 итераций декодирования вероятность появления ошибки в бите P_B была меньше 10^{-5} при $E_b/N_0 = 0{,}7$ дБ. Характер снижения вероятности появления ошибки при увеличении числа итераций можно увидеть на рис. 8.28. Заметьте, что достигается предел Шеннона –1,6 дБ. Требуемая ширина полосы пропускания приближается к бесконечности, а емкость (степень кодирования кода) приближается к нулю. Поэтому предел Шеннона является интересной границей с теоретической точки зрения, но не является практической целью. Для двоичной модуляции несколько авторов использовали в качестве *практического* предела Шеннона значения $P_B = 10^{-5}$ и $E_b/N_0 = 0{,}2$ дБ для кода со степенью кодирования 1/2. Таким образом, при параллельном соединении сверточных кодов RSC и декодировании с обратной свя-

зью, достоверность передачи турбокода при $P_B = 10^{-5}$ находится в 0,5 дБ от (практического) предела Шеннона. Существует класс кодов, в которых, вместо параллельного, используется последовательное соединение чередуемых компонентов. Предполагается, что последовательное соединение кодов может дать характеристики [28], превышающие аналоги при параллельном соединении.

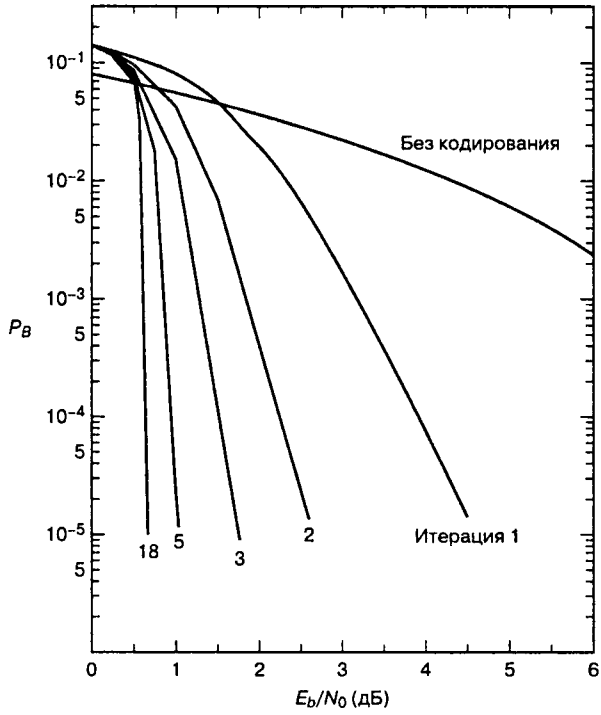


Рис. 8.28. Вероятность появления битовой ошибки как функция E_b/N_0 и количества итераций. (Источник: Berrou C., Glavieux A. and Thitimajshima P. "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo Codes". IEEE Proc. of Int'l Conf. on Communications, Geneva, Switzerland, May, 1993 (ICC '93), pp. 1064–1070.)

8.4.6. Алгоритм MAP

Процесс декодирования турбокода начинается с формирования *апостериорных вероятностей* (a posteriori probability — APP) для всех информационных битов, которые затем используются для изменения значений информационных битов в соответствии с *принципом максимума апостериорной* (maximum a posteriori — MAP) вероятности информационного бита. В ходе приема искаженной последовательности кодированных битов осуществляется схема принятия решений, основанная на значениях апостериорных вероятностей, и алгоритм MAP для определения наиболее вероятного информационного бита, который должен быть передан за время прохождения бита. Здесь имеется отличие от алгоритма Витерби, в котором апостериорная вероятность для каждого бита данных не существует. Вместо этого в алгоритме Витерби находится наиболее вероятная последовательность, которая могла быть передана. Но в реализации

обоих алгоритмов, впрочем, имеется сходство (см. раздел 8.4.6.3). Если декодированное P_B мало, существует незначительное различие в производительности между алгоритмами MAP и Витерби с мягким выходом (soft-output Viterbi algorithm — SOVA). Более того, при высоких значениях P_B и низких значениях E_b/N_0 алгоритм MAP превосходит алгоритм SOVA на 0,5 дБ и более [30, 31]. Это может оказаться очень важным для турбокодов, поскольку первая итерация декодирования может давать довольно высокую вероятность ошибки. Алгоритм MAP основывается на той же идее, что и алгоритм Витерби, — обработка блоков кодовых битов в двух направлениях. Как только такое двунаправленное вычисление даст состояние и метрики ветвей блока, можно начинать расчет апостериорной вероятности и MAP для каждого бита данных в блоке. Здесь предлагается алгоритм MAP декодирования для систематических сверточных кодов; полагается, что используется канал AWGN, как указано Питробоном [30]. Расчет начинается с отношения значений апостериорных вероятностей, известных как отношения правдоподобий $\Lambda(\hat{d}_k)$, или их логарифмов, $L(\hat{d}_k)$, называемых логарифмическими отношениями правдоподобий (log-likelihood ratio — LLR), как было показано в уравнении (8.110).

$$\Lambda(\hat{d}_k) = \frac{\sum_m \lambda_k^{1,m}}{\sum_m \lambda_k^{0,m}} \tag{8.118,а}$$

и

$$L(\hat{d}_k) = \lg\left[\frac{\sum_m \lambda_k^{1,m}}{\sum_m \lambda_k^{0,m}}\right] \tag{8.118,б}$$

Здесь $\lambda_k^{i,m}$ (совокупная вероятность того, что $d_k = i$ и $S_k = m$, при условии, что принята кодовая последовательность R_1^N, получаемая с момента $k = 1$ в течение некоторого времени N) определяется уравнением (8.108) и повторно приводится ниже.

$$\lambda_k^{i,m} = P\{d_k = i, S_k = m | R_1^N\}, \tag{8.119}$$

где R_1^N представляет искаженную последовательность кодированных битов, передаваемую по каналу, демодулированную и поданную на декодер согласно мягкой схеме решений. В действительности, алгоритм MAP требует, чтобы последовательность на выходе демодулятора подавалась на декодер по одному блоку из N бит за такт. Пусть R_1^N имеет следующий вид.

$$R_1^N = \{R_1^{k-1}, R_k, R_{k+1}^N\} \tag{8.120}$$

Чтобы упростить применение теоремы Байеса, уравнение (8.119) переписывается с использованием букв A, B, C и D. Таким образом, уравнение (8.119) примет следующий вид.

$$\lambda_k^{i,m} = P(\underbrace{d_k = i, S_k = m}_{A} | \underbrace{R_1^{k-1}}_{B}, \underbrace{R_k}_{C}, \underbrace{R_{k+1}^N}_{D}) \tag{8.121}$$

Вспомним, что теорема Байеса гласит следующее.

$$\begin{aligned} P(A|B,C,D) &= \frac{P(A,B,C,D)}{P(B,C,D)} = \frac{P(B|A,C,D)P(A,C,D)}{P(B,C,D)} = \\ &= \frac{P(B|A,C,D)P(D|A,C)P(A,C)}{P(B,C,D)} \end{aligned} \tag{8.122}$$

Отсюда, в приложении теоремы к уравнению (8.121), получается следующее.

$$\begin{aligned} \lambda_k^{i,m} = P(R_1^{k-1}|d_k = i, S_k = m, R_k^N) P(R_{k+1}^N|d_k = i, S_k = m, R_k) \times \\ \times P(d_k = i, S_k = m, R_k)/P(R_1^N), \end{aligned} \tag{8.123}$$

причем $R_k^N = \{R_k, R_{k+1}^N\}$. Уравнение (8.123) можно переписать, выделяя вероятностный член, вносящий вклад $\lambda_k^{i,m}$. В следующем разделе три множителя в правой части уравнения (8.123) будут определены и описаны как прямая метрика состояния, обратная метрика состояния и метрика ветви.

8.4.6.1. Метрики состояний и метрика ветви

Первый множитель в правой части уравнения (8.123) является прямой метрикой состояния для момента k и состояния m и обозначается α_k^m. Таким образом, для $i = 1, 0$

$$P(R_1^{k-1} | \overbrace{d_k = i}^{\text{Несущественно}}, S_k = m, \overbrace{R_k^N}^{\text{Несущественно}}) = P(R_1^{k-1}(S_k = m)) \stackrel{\text{def}}{=} \alpha_k^m \tag{8.124}$$

Следует отметить, что $d_k = i$ и R_k^N обозначены как несущественные, поскольку предположение о том, что $S_k = m$, подразумевает, что на события до момента k не влияют измерения после момента k. Другими словами, будущее не оказывает влияния на прошлое; таким образом, $P(R_1^{k-1})$ не зависит от того, что $d_k = i$ и последовательность равна R_k^N. В то же время, поскольку кодер обладает памятью, состояние кодера $S_k = m$ основывается на прошлом, а значит, этот член является значимым и его следует оставить в выражении. Очевидно, что форма уравнения (8.124) является понятной, поскольку представляет прямую метрику состояния α_k^m для момента k как вероятность того, что прошлая последовательность зависит только от теперешнего состояния, вызванного этой последовательностью и ничем более. В этом сверточном кодере нетрудно узнать уже упоминавшийся в главе 7 Марковский процесс.

Точно так же второй сомножитель в правой части уравнения (8.123) представляет собой обратную метрику состояния β_k^m для момента времени k и состояния m, определяемую следующим выражением.

$$P(R_{k+1}^N|d_k = i, S_k = m, R_k) = P(R_{k+1}^N|S_k = f(i,m)) \stackrel{\text{def}}{=} \beta_{k+1}^{f(i,m)} \tag{8.125}$$

Здесь $f(i, m)$ — это следующее состояние, определяемое входом i и состоянием m, а $\beta_{k+1}^{f(i,m)}$ — обратная метрика состояния в момент $k+1$ и состояние $f(i, m)$. Ясно, что уравнение (8.125) удовлетворяется, поскольку обратная метрика состояния β_{k+1}^{m} в будущий момент времени $k+1$ представлена как вероятность будущей последовательности, которая зависит от состояния (в будущий момент $k+1$), которое, в свою очередь, является функцией входного бита и состояния (в текущий момент k). Это уже знакомое основное определение конечного автомата (см. раздел 7.2.2).

Третий сомножитель в правой части уравнения (8.123) представляет собой метрику ветви (в состоянии m, в момент времени k), которая обозначается $\delta_k^{i,m}$. Таким образом, можно записать следующее.

$$P(d_k = i, S_k = m, R_k) \stackrel{\text{def}}{=} \delta_k^{i,m} \tag{8.126}$$

Подстановка уравнений (8.124)–(8.126) в уравнение (8.123) дает следующее, более компактное выражение для совокупной вероятности.

$$\lambda_k^{i,m} = \frac{\alpha_k^m \delta_k^{i,m} \beta_{k+1}^{f(i,m)}}{P(R_1^N)} \tag{8.127}$$

Используя уравнение (8.127), формулу (8.118) можно представить следующим образом.

$$\Lambda(\hat{d}_k) = \frac{\sum_m \alpha_k^m \delta_k^{1,m} \beta_{k+1}^{f(1,m)}}{\sum_m \alpha_k^m \delta_k^{0,m} \beta_{k+1}^{f(0,m)}} \tag{8.128,а}$$

и

$$L(\hat{d}_k) = \log\left[\frac{\sum_m \alpha_k^m \delta_k^{1,m} \beta_{k+1}^{f(1,m)}}{\sum_m \alpha_k^m \delta_k^{0,m} \beta_{k+1}^{f(0,m)}}\right] \tag{8.128,б}$$

Здесь $\Lambda(\hat{d}_k)$ — это отношение правдоподобий k-го бита данных; $L(\hat{d}_k)$, логарифм $\Lambda(\hat{d}_k)$, является логарифмическим отношением правдоподобий для k-го бита данных, где, в общем случае, логарифм берется по основанию e.

8.4.6.2. Расчет прямой метрики состояния

Исходя из уравнения (8.124), α_k^m можно представить как сумму всех возможных переходов из момента $k-1$.

$$\alpha_k^m = \sum_{m'} \sum_{j=0}^{1} P(d_{k-1} = j, S_{k-1} = m', R_1^{k-1} | S_k = m) \tag{8.129}$$

R_1^{k-1} можно переписать как $\{R_1^{k-2} R_{k-1}\}$ и, согласно теореме Байеса,

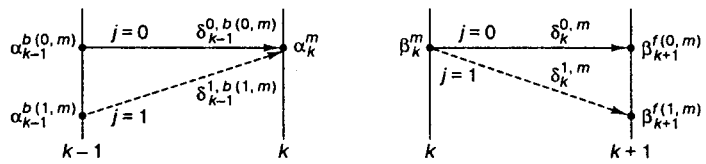
$$\alpha_k^m = \sum_{m'} \sum_{j=0}^{1} P(R_1^{k-2} | S_k = m, d_{k-1} = j, S_{k-1} = m', R_{k-1}) \times \tag{8.130,а}$$
$$\times P(d_{k-1} = j, S_{k-1} = m', R_{k-1} | S_k = m) =$$

$$= \sum_{j=0}^{1} P(R_1^{k-2}, S_{k-1} = b(j,m)) P(d_{k-1} = j, S_{k-1} = b(j,m), R_{k-1}), \tag{8.130,б}$$

где $b(j, m)$ — это состояние по предыдущей ветви, соответствующей входу j, исходящее обратно по времени из состояния m. Уравнение (8.130,б) может заменить уравнение (8.130,а), поскольку сведения о состоянии m' и входе j в момент времени $k-1$ полностью определяют путь в состояние $S_k = m$. Воспользовавшись уравнениями (8.124) и (8.126) для упрощения обозначений в уравнении (8.130), можно получить следующее.

$$\alpha_k^m = \sum_{j=0}^{1} \alpha_{k-1}^{b(j,m)} \delta_{k-1}^{j,b(j,m)} \tag{8.131}$$

Уравнение (8.131) означает, что новая прямая метрика состояния m в момент k получается из суммирования двух взвешенных метрик состояний в момент $k-1$. Взвешивание включает метрики ветвей, связанные с переходами, соответствующими информационным битам 1 и 0. На рис. 8.29, *а* показано применение двух разных типов обозначений для параметра α. Запись $\alpha_{k-1}^{b(j,m)}$ используется для обозначения прямой метрики состояния в момент времени $k-1$, если имеется два возможных предыдущих состояния (зависящих от того, равно ли j единице или нулю). А запись α_k^m применяется для обозначения прямой метрики состояния в момент k, если имеется два возможных перехода из предыдущего момента, которые оканчиваются в том же состоянии m в момент k.



а) Прямая метрика состояния

$$\alpha_k^m = \alpha_{k-1}^{b(0,m)} \delta_{k-1}^{0,b(0,m)} + \alpha_{k-1}^{b(1,m)} \delta_{k-1}^{1,b(1,m)},$$

где $b(j, m)$ — прошлое состояние, соответствующее входному j

б) Обратная метрика состояния

$$\beta_k^m = \beta_{k+1}^{f(0,m)} \delta_k^{0,m} + \beta_{k+1}^{f(1,m)} \delta_k^{1,m},$$

где $f(j, m)$ — следующее состояние, определяемое входным j и состоянием m

Метрика ветви

$$\delta_k^{i,m} = \pi_k^i \exp\left(x_k u_k^i + y_k v_k^{i,m}\right)$$

Рис. 8.29. Графическое представление расчета α_k^m и β_k^m. (Источник: Pietrobon S. S. "Implementation and Performance of a Turbo/Map Decoder". Int'l. J. of Satellite Communications, vol. 16, Jan.-Feb., 1998, pp. 23–46.)

8.4.6.3. Расчет обратной метрики состояния

Возвращаясь к уравнению (8.125), где $\beta_{k+1}^{f(i,m)} = P[R_{k+1}^N | S_{k+1} = f(i,m)]$, имеем следующее.

$$\beta_k^m = P(R_k^N | S_k = m) = P(R_k, R_{k+1}^N | S_k = m) \tag{8.132}$$

β_k^m можно представить как сумму вероятностей всех возможных переходов в момент $k+1$.

$$\beta_k^m = \sum_{m'} \sum_{j=0}^{1} P(d_k = j, S_{k+1} = m', R_k, R_{k+1}^N | S_k = m) \tag{8.133}$$

Применяя теорему Байеса, получим следующее.

$$\begin{aligned} \beta_k^m = \sum_{m'} \sum_{j=0}^{1} & P(R_{k+1}^N | S_k = m, d_k = j, S_{k+1} = m', R_k) \times \\ & \times P(d_k = j, S_{k+1} = m', R_k | S_k = m) \end{aligned} \tag{8.134}$$

В первом члене правой части уравнения (8.134) $S_k = m$ и $d_k = j$ полностью определяют путь, ведущий в $S_{k+1} = f(j, m)$; следующее состояние будет иметь вход j и состояние m. Таким образом, эти условия позволяют заменить $S_{k+1} = m'$ на $S_k = m$ во втором члене уравнения (8.134), что дает следующее.

$$\begin{aligned} \beta_k^m &= \sum_{j=0}^{1} P(R_{k+1}^N | S_{k+1} = f(j,m)) P(d_k = j, S_k = m, R_k) = \\ &= \sum_{j=0}^{1} \delta_k^{j,m} \beta_{k+1}^{f(j,m)} \end{aligned} \tag{8.135}$$

Уравнение (8.135) показывает, что новая обратная метрика состояния m в момент k, получается путем суммирования двух взвешенных метрик состояния в момент $k+1$. Взвешивание включает метрики ветвей, связанные с переходами, соответствующими информационным битам 1 и 0. На рис. 8.29, *б* показано применение двух разных типов обозначений для параметра β. Первый тип, запись $\beta_{k+1}^{f(j,m)}$, используется для обозначения обратной метрики состояния в момент времени $k+1$, если имеется два возможных предыдущих состояния (зависящих от того, равно ли j единице или нулю). Второй тип, β_k^m, применяется для обозначения обратной метрики состояния в момент k, если имеется два возможных перехода, поступающих в момент $k+1$, которые выходят из того же состояния m в момент времени k. На рис. 8.29 приведены пояснения к вычислениям прямой и обратной метрик.

Алгоритм декодирования MAP подобен алгоритму декодирования Витерби (см. раздел 7.3). В алгоритме Витерби метрика ветви прибавляется к метрике состояния. Затем сравнивается и выбирается минимальное расстояние (максимально правдоподобное) для получения следующей метрики состояния. Этот процесс называется сложение, сравнение и выбор (Add-Compare-Select — ACS). В алгоритме MAP выполняется умножение (в лога-

рифмическом представлении — сложение) метрик состояния и метрик ветвей. Затем, вместо сравнения, осуществляется их суммирование для вычисления следующей прямой (обратной) метрики состояния, как это видно из рис. 8.29. Различия воспринимаются на уровне интуиции. В алгоритме Витерби осуществляется поиск наиболее вероятной последовательности (пути); следовательно, выполняется постоянное сравнение и отбор, для того чтобы отыскать наилучший путь. В алгоритме MAP выполняется поиск правдоподобного или логарифмически правдоподобного числа (в мягкой схеме); следовательно, за период времени процесс использует все метрики из всех возможных переходов, чтобы получить полную статистическую картину информационных битов в данном периоде времени.

8.4.6.4. Расчет метрики ветви

Сначала обратимся к уравнению (8.126).

$$\begin{aligned}\beta_k^{i,m} &= P(d_k = i, S_k = m, R_k) = \\ &= P(R_k | d_k = i, S_k = m) P(S_k = m | d_k = i) P(d_k = i)\end{aligned} \tag{8.136}$$

Здесь R_k представляет собой последовательность $\{x_k, y_k\}$, x_k — это принятые биты данных с шумом, а y_k — принятые контрольные биты с шумом. Поскольку помехи влияют на информационные биты и биты контроля четности независимо, текущее состояние не зависит от текущего входа и, следовательно, может быть одним из 2^{υ} состояний, где υ — это число элементов памяти в сверточной кодовой системе. Иными словами, длина кодового ограничения этого кода, K, равняется $\upsilon + 1$. Значит,

$$P(S = m | d_k = i) = \frac{1}{2^{\upsilon}}$$

и

$$\delta_k^{i,m} = P(x_k | d_k = i, S_k = m) P(y_k | d_k = i, S_k = m) \frac{\pi_k^i}{2^{\upsilon}}, \tag{8.137}$$

где π_k^i обозначает $P(d_k = i)$, априорную вероятность d_k.

Из уравнения (1.25,г) в главе 1, вероятность $P(X_k = x_k)$ того, что случайная переменная X_k примет значение x_k, связана с функцией плотности вероятности $p_{x_k}(x_k)$ следующим образом.

$$P(X_k = x_k) = p_{x_k}(x_k)\, dx_k \tag{8.138}$$

Для упрощения обозначений случайная переменная X_k, принимающая значение x_k, часто будет называться "случайной переменной x_k", которая будет представлять значения x_k и y_k в уравнении (8.137). Таким образом, для канала AWGN, в котором шум имеет нулевое среднее и дисперсию σ^2, при замене вероятностного члена в уравнении (8.137) его эквивалентом (функцией плотности вероятности) используется уравнение (8.138), что дает следующее.

$$\delta_k^{i,m} = \frac{\pi_k^i}{2^{\upsilon}\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k - u_k^i}{\sigma}\right)^2\right] dx_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_k - v_k^{i,m}}{\sigma}\right)^2\right] dy_k \tag{8.139}$$

Здесь u_k и v_k представляют переданные биты данных и биты контроля четности (в биполярной форме), а dx_k и dy_k являются дифференциалами x_k и y_k и далее будут включаться в постоянную A_k. Следует заметить, что параметр u_k^i представляет данные, не зависящие от состояния m, поскольку код имеет память. Для того чтобы привести выражение к более простому виду, нужно исключить все члены в числителе и знаменателе и использовать сокращения; в результате получим следующее.

$$\delta_k^{i,m} = A_k \pi_k^i \exp\left[\frac{1}{\sigma^2}\left(x_k u_k^i + y_k v_k^{i,m}\right)\right] \tag{8.140}$$

Если подставить уравнение (8.140) в уравнение (8.128,а), получим следующее.

$$\Lambda(\hat{d}_k) = \pi_k \exp\left(\frac{2x_k}{\sigma^2}\right) \frac{\sum\limits_m \alpha_k^m \exp\left(\frac{y_k v_k^{1,m}}{\sigma^2}\right)\beta_{k+1}^{f(1,m)}}{\sum\limits_m \alpha_k^m \exp\left(\frac{y_k v_k^{0,m}}{\sigma^2}\right)\beta_{k+1}^{f(0,m)}} = \tag{8.141,а}$$

$$= \pi_k \exp\left(\frac{2x_k}{\sigma^2}\right)\pi_k^e \tag{8.141,б}$$

и

$$L(\hat{d}_k) = L(d_k) + L_c(x_k) + L_e(\hat{d}_k) \tag{8.141,в}$$

Здесь $\pi_k = \pi_k^i / \pi_k^0$ является входным отношением априорных вероятностей (априорное правдоподобие), а π_k^e — внешним выходным правдоподобием, каждое в момент времени k. В уравнении (8.141,б) член π_k^e можно считать фактором коррекции (вследствие кодирования), который меняет входные априорные сведения о битах данных. В турбокоде такие корректировочные члены проходят из одного декодера в другой, чтобы улучшить отношение правдоподобий для каждого информационного бита и, таким образом, минимизировать вероятность появления ошибок декодирования. Следовательно, процесс декодирования влечет за собой использование уравнения (8.141,б) для получения за несколько итераций $\Lambda(\hat{d}_k)$. Внешнее правдоподобие π_k^e, получаемое из конкретной итерации, заменяет априорное правдоподобие π_{k+1} на следующую итерацию. Взятие логарифма от $\Lambda(\hat{d}_k)$ в уравнении (8.141,б) дает уравнение (8.141,в), которое показывает те же результаты, что и уравнение (8.71). Они заключаются в том, что итоговые данные $L(\hat{d}_k)$ (согласно мягкой схеме принятия решений) образуются тремя членами LLR — априорным LLR, LLR канального измерения и внешним LLR.

Алгоритм MAP можно реализовать через отношение правдоподобий $\Lambda(\hat{d}_k)$, как показывает уравнение (8.128,а) или (8.141,в); конструкция станет менее громоздкой за счет устранения операций умножения.

8.4.7. Пример декодирования по алгоритму MAP

На рис. 8.30 изображен пример декодирования по алгоритму MAP. На рис. 8.30, *а* представлен систематический сверточный кодер с длиной кодового ограничения $K = 3$ и степенью кодирования 1/2. Входные данные — последовательность $\mathbf{d} = \{1, 0, 0\}$, соответствующая временам $k = 1, 2, 3$. Выходная кодированная битовая последовательность образуется путем последовательного взятия одного бита из последовательности $\mathbf{u} = \{1, 0, 0\}$ вслед за битом контроля четности из последовательности $\mathbf{v} = \{1, 0, 1\}$. В каждом случае крайний слева бит является самым первым. Таким образом, выходной последовательностью будет 1 1 0 0 0 1 или ее биполярное представление — +1 +1 −1 −1 −1 +1. На рис. 8.30, *б* видны результаты искажения последовательностей **u** и **v** векторами помех $\mathbf{n}_x$ и $\mathbf{n}_y$, так что теперь они обозначаются как $\mathbf{x} = \mathbf{u} + \mathbf{n}_x$ и $\mathbf{y} = \mathbf{v} + \mathbf{n}_y$. Как показано на рис. 8.30, *б*, входные данные демодулятора, поступающие на декодер в моменты $k = 1, 2, 3$, имеют значения 1,5; 0,8; 0,5; 0,2; –0,6; 1,2. Также показаны априорные вероятности того, что принятые биты данных будут равны 1 или 0, что обозначается как π^1 и π^0. Предполагается, что эти вероятности будут одинаковы для всех k моментов времени. В этом примере уже имеется вся необходимая информация для расчета метрик ветвей и метрик состояний и ввода их значений в решетчатую диаграмму декодера, изображенную на рис. 8.30, *в*. На решетчатой диаграмме каждый переход, возникающий между временами k и $k+1$, соответствует информационному биту d_k, который появляется на входе кодера в момент начала перехода k. В момент времени k кодер находится в некотором состоянии m, а в момент $k+1$ он переходит в новое состояние (возможно, такое же). Если использовать такую решетчатую диаграмму для отображения последовательности кодовых битов (представляющих N бит данных), последовательность будет описываться N временами переходов и $N+1$ состояниями.

8.4.7.1. Расчет метрик ветвей

Начнем с уравнения (8.140) при $\pi_k^i = 0{,}5$ (в данном примере информационные биты считаются равновероятными для любых времен). Для простоты предполагается, что $A_k = 1$ для всех моментов и $\sigma^2 = 1$. Таким образом, $\delta_k^{i,m}$ примет следующий вид.

$$\delta_k^{i,m} = 0{,}5\exp(x_k u_k^i + y_k v_k^{i,m}) \tag{8.142}$$

На что похожа основная функция приемника, определяемая уравнением (8.142)? Выражение напоминает корреляционный процесс. В декодере в каждый момент k принимается пара данных (x_k, относящееся к битам данных, и y_k, относящееся к контрольным битам). Метрика ветви рассчитывается путем умножения принятого x_k на каждый первообразный сигнал u_k и принятого y_k на каждый первообразный сигнал v_k. Для каждого перехода по решетке величина метрики ветви будет функцией того, насколько согласуются пара данных, принятых с помехами, и кодовые значения битов этого перехода по решетке. При $k = 1$ для вычисления восьми метрик ветвей (переходов из состояний m для всех значений данных i) применяется уравнение (8.142). Для простоты, состояния на решетке обозначены следующим образом: $a = 00$, $b = 10$, $c = 01$, $d = 11$. Заметьте, что кодированные битовые значения, u_k, v_k, каждого перехода по решетке указаны над самими переходами, как можно видеть на

рис. 8.30, *в* (только для $k = 1$), и их можно получить обычным образом, используя структуру кодера (см. раздел 7.2.4.). Для переходов по решетке на рис. 8.30, *в* оговаривается, что пунктирные и сплошные линии обозначают информационные биты 1 и 0. Расчеты дают такие значения.

$$\delta_{k=1}^{1,m=a} = \delta_{k=1}^{1,m=b} = 0{,}5\exp[(1{,}5)(1) + (0{,}8)(1)] = 5{,}0$$

$$\delta_{k=1}^{0,m=a} = \delta_{k=1}^{0,m=b} = 0{,}5\exp[(1{,}5)(-1) + (0{,}8)(-1)] = 0{,}05$$

$$\delta_{k=1}^{1,m=c} = \delta_{k=1}^{1,m=d} = 0{,}5\exp[(1{,}5)(1) + (0{,}8)(-1)] = 1{,}0$$

$$\delta_{k=1}^{0,m=c} = \delta_{k=1}^{0,m=d} = 0{,}5\exp[(1{,}5)(-1) + (0{,}8)(1)] = 0{,}25$$

Затем эти расчеты повторяются, с помощью уравнения (8.142), для восьми метрик ветвей в момент $k = 2$.

$$\delta_{k=2}^{1,m=a} = \delta_{k=2}^{1,m=b} = 0{,}5\exp[(0{,}5)(1) + (0{,}2)(1)] = 1{,}0$$

$$\delta_{k=2}^{0,m=a} = \delta_{k=2}^{0,m=b} = 0{,}5\exp[(0{,}5)(-1) + (0{,}2)(-1)] = 0{,}25$$

$$\delta_{k=2}^{1,m=c} = \delta_{k=2}^{1,m=d} = 0{,}5\exp[(0{,}5)(1) + (0{,}2)(-1)] = 0{,}67$$

$$\delta_{k=2}^{0,m=c} = \delta_{k=2}^{0,m=d} = 0{,}5\exp[(0{,}5)(-1) + (0{,}2)(1)] = 0{,}37$$

Снова расчеты повторяются для значений восьми метрик ветвей уже в момент $k = 3$.

$$\delta_{k=3}^{1,m=a} = \delta_{k=3}^{1,m=b} = 0{,}5\exp[(-0{,}6)(1) + (1{,}2)(1)] = 0{,}91$$

$$\delta_{k=3}^{0,m=a} = \delta_{k=3}^{0,m=b} = 0{,}5\exp[(-0{,}6)(-1) + (1{,}2)(-1)] = 0{,}27$$

$$\delta_{k=3}^{1,m=c} = \delta_{k=3}^{1,m=d} = 0{,}5\exp[(-0{,}6)(1) + (1{,}2)(-1)] = 0{,}08$$

$$\delta_{k=3}^{0,m=c} = \delta_{k=3}^{0,m=d} = 0{,}5\exp[(-0{,}6)(-1) + (1{,}2)(1)] = 3{,}0$$

8.4.7.2. Расчет метрик состояний

Как только при всех k рассчитаны восемь значений $\delta_k^{i,m}$, можно вычислить прямые метрики состояний α_k^m, воспользовавшись рис. 8.29, 8.30, *в* и уравнением (8.131), которое повторно приводится ниже.

$$\alpha_{k+1}^m = \sum_{j=0}^{1} \delta_k^{j,b(j,m)} \alpha_k^{b(j,m)}$$

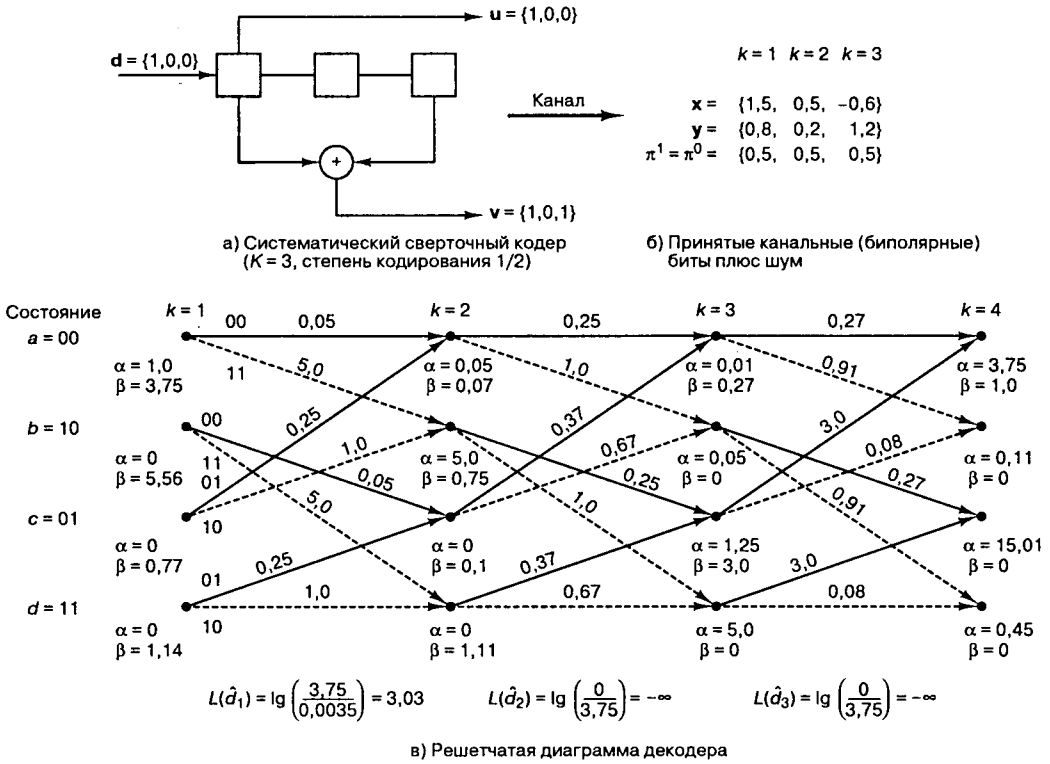
Допустим, что начальным состоянием кодера будет $a = 00$. Тогда,

$$\alpha_{k=1}^{m=a} = 0 \text{ и } \alpha_{k=1}^{m=b} = \alpha_{k=1}^{m=c} = \alpha_{k=1}^{m=d} = 0$$

$$\alpha_{k=2}^{m=a} = (0{,}05)(1{,}0) + (0{,}25)(0) = 0{,}05$$

$$\alpha_{k=2}^{m=b} = (5{,}0)(1{,}0) + (1{,}0)(0) = 5{,}0 \qquad \text{и т.д.}$$

$$\alpha_{k=2}^{m=c} = \alpha_{k=2}^{m=d} = 0$$



а) Систематический сверточный кодер ($K = 3$, степень кодирования 1/2)

б) Принятые канальные (биполярные) биты плюс шум

в) Решетчатая диаграмма декодера

Рис. 8.30. Пример декодирования по алгоритму MAP ($K = 3$, степень кодирования 1/2, систематическое)

как показано с помощью решетчатой диаграммы на рис. 8.30, *в*. Аналогично можно вычислить обратные метрики состояний β_k^m, воспользовавшись рис. 8.29, 8.30, *в* и уравнением (8.135), которое повторно приводится ниже.

$$\beta_k^m = \sum_{j=0}^{1} \delta_k^{j,m} \beta_{k+1}^{f(j,m)}$$

Последовательность данных и код в этом примере умышленно были изменены так, что финальным состоянием решетки в момент времени $k = 4$ является $a = 00$. В противном случае нужно использовать остаточные биты для принудительного изменения конечного состояния системы в такое известное состояние. Таким образом, в этом примере, проиллюстрированном на рис. 8.30, исходя из того, что конечное состояние — $a = 00$, можно рассчитать обратные метрики состояний.

$$\beta_{k=4}^{m=a} = 1{,}0 \text{ и } \beta_{k=4}^{m=b} = \beta_{k=4}^{m=c} = \beta_{k=4}^{m=d} = 0$$

$$\beta_{k=3}^{m=a} = (0{,}27)(1{,}0) + (0{,}91)(0) = 0{,}27$$

$$\beta_{k=3}^{m=b} = \beta_{k=3}^{m=d} = 0$$

$$\beta_{k=3}^{m=c} = (3{,}0)(1{,}0) + (0{,}08)(0) = 3{,}0$$

Все значения обратных метрик состояний показаны на решетке (рис. 8.30, *в*).

8.4.7.3. Расчет логарифмического отношения правдоподобий

Теперь у нас есть рассчитанные метрики β, α и δ для кодированной битовой последовательности рассматриваемого примера. В процессе турбодекодирования для нахождения решения согласно мягкой схеме, $\Lambda(\hat{d}_k)$ или $L(\hat{d}_k)$, для каждого бита данных можно воспользоваться уравнением (8.128) или (8.141). При использовании турбокодов этот процесс повторяется несколько раз, чтобы достичь необходимой надежности решений. В целом все заканчивается применением параметра внешнего правдоподобия из уравнения (8.141,б) для вычисления и повторного расчета в несколько итераций отношения правдоподобий $\Lambda(\hat{d}_k)$. Внешнее правдоподобие π_k^e последней итерации заменяет в следующей итерации априорное правдоподобие π_{k+1}.

В нашем примере будут использованы метрики, рассчитанные ранее (с одним прохождением через декодер). Для вычисления LLR каждого информационного бита в последовательности $\{d_k\}$ берется уравнение (8.128,б). Затем, с помощью правила принятия решений из уравнения (8.111), итоговые данные, представленные согласно мягкой схеме решений, преобразуются в решения в жесткой схеме. Для $k = 1$, опуская некоторые нулевые слагаемые, получаем следующее.

$$L(\hat{d}_k) = \lg\left(\frac{1{,}0\times 5{,}0\times 0{,}75}{1{,}0\times 0{,}05\times 0{,}07}\right) = \lg\left(\frac{3{,}75}{0{,}0035}\right) = 3{,}03$$

Для $k = 2$, опуская некоторые нулевые слагаемые, получаем следующее.

$$L(\hat{d}_k) = \lg\left(\frac{(0{,}05\times 1{,}0\times 0)+(5{,}0\times 1{,}0\times 0)}{(0{,}05\times 0{,}25\times 0{,}27)+(5{,}0\times 0{,}25\times 3{,}0)}\right) = \lg\left(\frac{0}{3{,}75}\right) = -\infty$$

Для $k = 3$

$$L(\hat{d}_k) = \lg\left[\frac{(0{,}01\times 0{,}91\times 0)+(0{,}05\times 0{,}91\times 0)}{(0{,}01\times 0{,}27\times 1{,}0)+(0{,}05\times 0{,}27\times 0)}\right.$$
$$\left.\frac{+(1{,}25\times 0{,}08\times 0)+(5{,}0\times 0{,}08\times 0)}{+(1{,}25\times 3{,}0\times 1{,}0)+(5{,}0\times 3{,}0\times 0)}\right] =$$
$$= \lg\left(\frac{0}{3{,}75}\right) = -\infty$$

С помощью уравнения (8.111) для выражения финального решения относительно битов в моменты $k = 1, 2, 3$, последовательность декодируется как {1 0 0}. Итак, получен абсолютно точный результат, совпадающий с теми данными, которые были введены в декодер.

8.4.7.4. Реализация конечного автомата с помощью регистра сдвига

В этой книге используются регистры сдвига с прямой и обратной связью, представленные по большей части как разряды памяти и соединительные линии. Важно обратить внимание на то, что часто оказывается удобным представлять кодер (конкретнее, рекурсивный кодер) на регистрах сдвига несколько иным образом. Некоторые авторы для обозначения временных задержек (как правило, длиной в 1 бит) используют блоки,

помеченные буквами D или T. Соединения вне блоков, передающие напряжение или логические уровни, представляют память кодера между тактами. Два формата — блоки памяти и блоки задержек — никоим образом не меняют характеристик или функционирования описанного выше процесса. Для некоторых конечных автоматов с множеством рекурсивных соединений при отслеживании сигналов более удобным может оказаться применение формата блоков задержек. В задачах 8.23 и 8.24 используются кодеры, изображенные на рис. 38.2 и 38.3. При использовании формата разрядов памяти текущее состояние системы определяется содержимым крайних правых $K-1$ разрядов. Аналогично для формата блоков задержек текущее состояние определяется уровнями сигналов в крайних правых $K-1$ узлах (соединения вне блоков задержек). Для обоих форматов связь между памятью v и длиной кодового ограничения K одинакова, т.е. $v=K-1$. Таким образом, на рис. 38.2 три блока задержек означают, что $v=3$ и $K=4$. Аналогично на рис. 38.3 два блока задержек означают, что $v=2$, а $K=3$.

8.5. Резюме

В этой главе мы рассмотрели коды Рида-Соломона, важный класс недвоичных блочных кодов, специально применяемых для коррекции пакетных ошибок. Коды Рида-Соломона особенно привлекательны, поскольку эффективность кода растет с его длиной. При большой длине блока коды можно сконфигурировать таким образом, что время декодирования будет значительно меньше, чем у других кодов с той же длиной блока. Это связано с тем, что декодер работает с целыми символами, а не битами. Следовательно, для 8-битовых символов арифметические операции будут выполняться на уровне байтов. По сравнению с двоичными кодами той же длины это повышает не только сложность логики, но и производительность.

Далее была описана методика, называемая чередованием, которая без потерь в качестве позволяет использовать большинство блочных и сверточных схем кодирования в каналах с импульсными помехами или периодическим замиранием. В качестве примера была приведена система цифровой аудиозаписи на компакт-дисках, иллюстрирующая, какую важную роль играют кодирование Рида-Соломона и чередование в устранении эффектов импульсных помех.

Мы описали каскадные коды и принципы турбокодирования, основная конфигурация которых — это соединение двух или более составных кодов. Здесь также были рассмотрены фундаментальные статистические меры, такие как апостериорная вероятность и правдоподобие, которые затем использовались для описания достоверности передачи декодера с мягким входом и мягким выходом. Кроме того, было показано, как повышается достоверность передачи при включении каскадного декодера с мягким выходом в итеративный процесс декодирования. Затем эти идеи были использованы при параллельном соединении рекурсивных систематических сверточных (recursive systematic convolutional — RSC) кодов, в результате чего было получено объяснение, почему в турбокодах такие коды более предпочтительны в качестве компонентов. В общих чертах здесь описан декодер с обратной связью и представлены его отличительные особенности. Далее была разработана математика декодера, основанного на принципе максимума апостериорной вероятности (maximum a posteriori — MAP), и приведен численный пример (пересечение решетчатой диаграммы в двух направлениях), в котором в итоге были получены выходные данные, оформленные согласно мягкой схеме принятия решений.

Приложение 8A. Сложение логарифмических отношений правдоподобий

Ниже приводятся алгебраические подробности, используемые при выводе уравнения (8.72).

$$L(d_1) \boxplus L(d_2) \stackrel{\text{def}}{=} L(d_1 \oplus d_2) = \ln\left(\frac{e^{L(d_1)} + e^{L(d_2)}}{1 + e^{L(d_1)} e^{L(d_2)}}\right) \tag{8A.1}$$

Начнем с записи отношения правдоподобия апостериорной вероятности того, что информационный бит равен +1, к апостериорной вероятности того, что он равен –1. Поскольку логарифм отношения правдоподобий, обозначаемый $L(d)$, берется по основанию e, это можно записать следующим образом.

$$L(d) = \ln\left[\frac{P(d=+1)}{P(d=-1)}\right] = \ln\left[\frac{P(d=+1)}{1-P(d=+1)}\right], \tag{8A.2}$$

так что

$$e^{L(d)} = \left[\frac{P(d=+1)}{1-P(d=+1)}\right]. \tag{8A.3}$$

Выражая $P(d+1)$, получаем следующее.

$$e^{L(d)} - e^{L(d)} \times P(d=+1) = P(d=+1) \tag{8A.4}$$

$$e^{L(d)} = P(d=+1) \times [1 + e^{L(d)}] \tag{8A.5}$$

и

$$P(d=+1) = \frac{e^{L(d)}}{1+e^{L(d)}} \tag{8A.6}$$

Из уравнения (8A.6) видно, что

$$P(d=-1) = 1 - P(d=+1) = 1 - \frac{e^{L(d)}}{1+e^{L(d)}} = \frac{1}{1+e^{L(d)}}. \tag{8A.7}$$

Пусть d_1 и d_2 — два статистически независимых бита данных, задаваемых уровнями напряжения +1 и –1, соответствующими логическим уровням 1 и 0.

При таком формате сложение (по модулю 2) d_1 и d_2 дает –1, если d_1 и d_2 имеют одинаковое значение (оба равны +1 или –1, одновременно), и +1, если d_1 и d_2 имеют разные значения. Тогда

$$\begin{aligned} L(d_1 \oplus d_2) &= \ln\left[\frac{P(d_1 \oplus d_2 = 1)}{P(d_1 \oplus d_2 = -1)}\right] = \\ &= \ln\left[\frac{P(d_1=+1) \times P(d_2=-1) + [1-P(d_1=+1)][1-P(d_2=-1)]}{P(d_1=+1) \times P(d_2=+1) + [1-P(d_1=+1)][1-P(d_2=+1)]}\right] \end{aligned} \tag{8A.8}$$

Воспользовавшись уравнениями (8A.6) и (8A.7) для замены вероятностного члена в уравнении (8A.8), получаем следующее.

$$L(d_1 \oplus d_2) = \ln\left[\frac{\left(\frac{e^{L(d_1)}}{1+e^{L(d_1)}}\right)\left(\frac{1}{1+e^{L(d_2)}}\right)+\left(\frac{1}{1+e^{L(d_1)}}\right)\left(\frac{e^{L(d_2)}}{1+e^{L(d_2)}}\right)}{\left(\frac{e^{L(d_1))}}{1+e^{L(d_1)}}\right)\left(\frac{e^{L(d_2)}}{1+e^{L(d_2)}}\right)+\left(\frac{1}{1+e^{L(d_1)}}\right)\left(\frac{1}{1+e^{L(d_2)}}\right)}\right] = \tag{8A.9}$$

$$= \ln\left[\frac{\left(\frac{e^{L(d_1)}+e^{L(d_2)}}{[1+e^{L(d_1)}][1+e^{L(d_2)}]}\right)}{\left(\frac{e^{L(d_1)}e^{L(d_2)}+1}{[1+e^{L(d_1)}][1+e^{L(d_2)}]}\right)}\right] = \tag{8A.10}$$

$$= \ln\left[\frac{e^{L(d_1)}+e^{L(d_2)}}{1+e^{L(d_1)}e^{L(d_2)}}\right] \tag{8A.11}$$

Литература

1. Gallager R. G. *Information Theory and Reliable Communication.* John Wiley and Sons, New York, 1968.
2. Odenwalder J. P. *Error Control Coding Handbook.* Linkabit Corporation, San Diego, CA, July, 15, 1976.
3. Derlekamp E. R., Peile R. E. and Pope S. P. *The Application of Error Control to Communicftions.* IEEE Communication Magazine, vol. 25, n. 4, April, 1987, pp. 44–57.
4. Hagenauer J. and Lutz E. *Forward Error Correction Coding for Fading Compensation in Mobile Satellite Channels.* IEEE J. on Selected Areas in Comm., vol. SAC-5, n. 2, February, 1987, pp. 215–225.
5. Blahut R. E. *Theory and Practice of Error Control Codes.* Addison-Wesley Publishing Co., Reading, Massachusetts, 1983.
6. *Reed-Solomon Codes and Their Applications,* ed. Wicker S. B. and Bhargava V. K. IEEE Press, Piscataway, New Jersey, 1983.
7. Ramsey J. L. *Realization of Optimum Interleavers.* IEEE Trans. Inform. Theory, vol. IT-16, n. 3, May, 1970, pp.338–345.
8. Forney G. D. *Burst-Correcting Codes for the Classic Bursty Channel.* IEEE Trans. Commun. Technol., vol. COM-19, October, 1971, pp. 772–781.
9. Clark G. C. Jr. and Cain J. B. *Error-Corection Coding for Digiral Communications.* Plenum Press, New York, 1981.
10. J. H. Yuen, et. al. *Modulation and Coding for Satellite and Space Communications.* Proc. IEEE, vol. 78., n. 7, July, 1990, pp. 1250–1265.
11. Peek J. B. H. *Communications Aspects of the Compact Disc Digital Audio System.* IEEE Communication Magazine, vol. 23, n. 2, February, 1985, pp.7–20.
12. Berkhout P. J. and Eggermont L. D. J. *Digital Audio Systems.* IEEE ASSP Magazine, October, 1985, pp. 45–67.
13. Driessen L. M. H. E. and Vries L. B. *Performance Calculations of the Compact Disc Error Correcting Code on Memoryless Channel.* Fourth Int'l. Conf. Video and Data Recording, Southampton, England, April 20–23, 1982, IERE Conference Proc #54, pp. 385–395.
14. Hoeve H., Timmermans J. and Vries L. B. *Error Correction in the Compact Disc System.* Philips Tech. Rev., vol. 40, n. 6, 1982, pp. 166–172.
15. Pohlmann K. C. *The Compact Disk Handbook.* A-R Editions Inc., Madison, Wisconsin, 1992.
16. Forney G. D. Jr. *Concatenated Codes.* Cambridge, Massachusetts: M. I. T. Press, 1966.
17. Berrou C., Glavieux A. and Thitimajshima P. *Near Shannon Limit Error-Correcting Coding and Decoding: Turbo Codes.* IEEE Prceedings of the Int. Conf. on Communications, Geneva, Switzerland, May, 1993 (ICC'93), pp.1064–1070.

18. Berrou C., Glavieux A. *Near Optimum Error Correcting Coding and Decoding: Turbo-Codes.* IEEE Trans. On Communications, vol. 44, n. 10, October, 1996, pp. 1261–1271.
19. Hagenauer J. *Iterative Decoding of Binary Block and Convolutional Codes.* IEEE Trans. On Information Theory, vol. 42, n. 2, March, 1996, pp. 429–445.
20. Divsalar D. and Pollara F. *On the Design of Turbo Codes.* TDA Progress Report 42–123, Jet Propulsion Laboratory, Pasadena, California, November, 15, 1995, pp. 99–121.
21. Divsalar D. and McElience R. J. *Effective Free Distance of Turbo Codes.* Electronic Letters, vol. 23, n. 5, February, 29, 1996, pp. 445–446.
22. Dolinar S. and Divsalar D. *Weight Distributions for Turbo Codes Using Random and Nonrandom Permutations.* TDA Progress Report 42–122, Jet Propulsion Laboratory, Pasadena, California, August, 15, 1995, pp. 56–65.
23. Divsalar D. and Pollara F. *Turbo Codes for Deep-Space Communications.* TDA Progress Report 42–120, Jet Propulsion Laboratory, Pasadena, California, February, 15, 1995, pp. 29–39.
24. Divsalar D. and Pollara F. *Multiple Turbo Codes for Deep-Space Communications.* TDA Progress Report 42–121, Jet Propulsion Laboratory, Pasadena, California, May, 15, 1995, pp. 66–77.
25. Divsalar D. and Pollara F. *Turbo Codes for PCS Applications.* Proc. ICC'95, Seattle, Washington, June, 18–22, 1995.
26. Bahl. L. R., Cocke J., Jelinek F. and Raviv J. *Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate.* Trans. Inform. Theory, vol. IT-20, March, 1974, pp. 248–287.
27. Benedetto S. et. al. *Soft Output Decoding Algorithm in Iterative Decoding of Turbo Codes.* TDA Progress Report 42–124, Jet Propulsion Laboratory, Pasadena, California, February, 15, 1996, pp. 63–87.
28. Benedetto S. et. al. *A Soft-Input Soft-Output Maximum A Posteriori (MAP) Module to Decode Parallel and Serial Concatenated Codes.* TDA Progress Report 42–127, Jet Propulsion Laboratory, Pasadena, California, November, 15, 1996, pp. 63–87.
29. Benedetto S. et. al. *A Soft-Input Soft-Output APP Module for Iterative Decoding of Concatenated Codes.* IEEE Communication Letters, vlo. 1, n. 1, January, 1997, pp. 22–24.
30. Pietrobon S. *Implementation and Performance of a Turbo/MAP Decoder.* Int'l. J. Satellite Commun., vol. 15, January–February, 1998, pp. 23–46.
31. Robertson P., Villebrun E. and Hoeher P. *A Comparison of Optimal and Sub-Optimal MAP Decoding Algorithms Operating in the Log Domain.* Proc. of ICC'95, Seattle, Washington, June, 1995, pp. 1009–1013.

Задачи

8.1. Определите, какой из следующих полиномов будет примитивным. *Подсказка*: самый простой способ состоит в применении LFSR; аналогично способу, показанному на рис. 8.8.

а) $1 + X^2 + X^3$

б) $1 + X + X^2 + X^3$

в) $1 + X^2 + X^4$

г) $1 + X^3 + X^4$

д) $1 + X + X^2 + X^3 + X^4$

е) $1 + X + X^5$

ж) $1 + X^2 + X^5$

з) $1 + X^3 + X^5$

и) $1 + X^4 + X^5$

8.2. а) Какова способность к коррекции ошибок у кода (7, 3)? Сколько битов в символе?

б) Определите количество строк и столбцов нормальной матрицы (см. раздел 6.6), необходимой для представления кода (7, 3), описанного в п. а.

в) Подтвердите, что при использовании размерности нормальной матрицы из п. б получается способность к коррекции ошибок, найденная в п. а.

г) Является ли код (7, 3) совершенным? Если нет, то какую остаточную способность к коррекции ошибок он имеет?

8.3. а) Определите множество элементов $\{0, \alpha^0, \alpha^1, \alpha^2, \ldots, \alpha^{2^m-2}\}$ через образующие элементы конечного поля GF(2^m), при $m = 4$.

б) Для конечного поля, определенного в п. а, составьте таблицу сложения, аналогичную табл. 8.2.

в) Постройте таблицу умножения, подобную табл. 8.3.

г) Найдите полиномиальный генератор для кода (31, 27).

д) Кодируйте сообщение {96 нулей, затем 1100100011111} (крайний правый бит является первым) систематическим кодом (7, 3). Почему, по вашему мнению, сообщение построено с таким большим количеством нулей в начале?

8.4. С помощью полиномиального генератора для кода (7, 3), кодируйте сообщение 010110111 (крайний правый бит является первым) в систематической форме. Для нахождения полинома контроля четности используйте полиномиальное деление. Представьте итоговое кодовое слово в двоичной и полиномиальной форме.

8.5. а) Применяйте регистр LFSR для кодирования сообщения {6, 5, 1} (крайний правый бит является первым) с помощью кода (7, 3) в систематической форме. Представьте итоговое кодовое слово в двоичной форме.

б) Проверьте результат кодирования в п. а путем вычисления полинома кодового слова со значениями корней полиномиального генератора кода (7, 3).

8.6. а) Пусть кодовое слово, найденное в задаче 8.5, искажается в ходе передачи, в результате чего крайние правые 6 бит были инвертированы. Найдите значения всех синдромов путем вычисления полинома поврежденного кодового слова со значениями корней полиномиального генератора, $\mathbf{g}(X)$.

б) Проверьте, что значения синдромов, вычисленные в п. а, можно найти путем вычисления полинома ошибок, $\mathbf{e}(X)$, со значениями корней генератора $\mathbf{g}(X)$.

8.7. а) Воспользуйтесь авторегрессионной моделью из уравнения (8.40) вместе с искаженным кодовым словом из задачи 8.6 для нахождения месторасположения каждой символьной ошибки.

б) Найдите значение каждой символьной ошибки.

в) Воспользуйтесь сведениями, полученными в пп. а и б, чтобы исправить искаженное кодовое слово.

8.8. Последовательность 1011011000101100 подается на вход блочного устройства чередования размером 4 × 4. Какой будет выходная последовательность? Та же последовательность передана на сверточное устройство чередования, изображенное на рис. 8.13. Какой будет выходная последовательность в этом случае?

8.9. Для каждого из следующих условий разработайте устройство чередования для системы связи, действующей в канале с импульсными помехами со скоростью передачи 19200 кодовых символов/с.

а) Как правило, пакет шума длится 250 мс. Системным кодом является БЧХ (127, 36) при $d_{min} = 31$. Прямая задержка не превышает 5 с.

б) Как правило, пакет шума длится 20 мс. Системным кодом является сверточный код (127, 36) с обратной связью при степени кодирования 1/2, способный корректировать в среднем 3 символа в последовательности из 21 символа. Прямая задержка не превышает 160 мс.

8.10. а) Рассчитайте вероятность появления байтовой (символьной) ошибки после декодирования данных, находящихся на компакт-диске, как было описано в разделе 8.3. Считается, что вероятность передачи канального символа с ошибкой для компакт-диска составляет 10^{-3}. Предполагается также, что внешний и внутренний декодеры сконфигурированы для кор-

рекции всех 2-символьных ошибок и процесс чередования исключает корреляции ошибок между собой.

б) Повторите расчеты п. а для компакт-диска, для которого вероятность ошибочной передачи канального символа равна 10^{-2}.

8.11. Система, в которой реализована модуляция BPSK, принимает равновероятные биполярные символы (+1 или –1) с шумом AWGN. Дисперсия шума считается равной единице. В момент k значение принятого сигнала x_k равняется 0,11.

а) Вычислите два значения правдоподобия для этого принятого сигнала.

б) Каким будет максимальное апостериорное решение, +1 или –1?

в) Априорная вероятность того, что переданный символ равен +1, равна 0,3. Каким будет максимальное апостериорное решение, +1 или –1?

г) Считая априорные вероятности равными полученным в п. в, рассчитайте логарифмическое отношение правдоподобий $L(d_k|x_k)$.

8.12. Рассмотрим пример двухмерного кода с контролем четности, описанного в разделе 8.4.3. Как указано, переданные символы представляются в виде последовательности d_1, d_2, d_3, d_4, p_{12}, p_{34}, p_{13}, p_{24}, которая определяет степень кодирования кода равной 1/2. Конкретная схема, требующая более высоких скоростей, выдает кодовую последовательность этого кода с исключениями через один бит, что дает в итоге степень кодирования, равную 2/3. Переданный выход теперь определяется последовательностью $\{d_i\}$, $\{p_{ij}\}$ = +1 –1 –1 +1 +1 +1, где i и j являются индексами месторасположения. Помехи преобразуют эту последовательность данных и контрольных разрядов в принятую последовательность $\{x_k\}$ = 0,75, 0,05, 0,10, 0,15, 1,25, 3,0, где k —временной индекс. Вычислите значения мягкого выхода для битов данных после двух горизонтальных и двух вертикальных итераций декодирования. Дисперсия считается равной единице.

8.13. Рассмотрим параллельное соединение двух составных RSC-кодеров, как показано на рис. 8.26. Устройство чередования блоков размером 10 отображает последовательность входных битов $\{d_k\}$ в последовательность $\{d'_k\}$, где влияние устройства чередования задается равным [6, 3, 8, 9, 5, 7, 1, 4, 10, 2], т.е. первый бит входного блока данных отображается на позицию 6, второй бит — на позицию 3 и т.д. Входная последовательность равна (0, 1, 1, 0, 0, 1, 0, 1, 1, 0). Предполагается, что составной кодер начинает из нулевого состояния и к нему принудительно прибавляется бит погашения, необходимый для перевода кодера обратно в нулевое состояние.

а) Рассчитайте 10-битовую контрольную последовательность $\{v_{1k}\}$.

б) Рассчитайте 10-битовую контрольную последовательность $\{v_{2k}\}$.

в) Переключатель осуществляет исключение из последовательности $\{v_k\}$ так, что последовательность $\{v_k\}$ становится равной v_{1k}, $v_{2(k+1)}$, $v_{1(k+2)}$, $v_{2(k+3)}$, ..., а степень кодирования кода равна 1/2. Рассчитайте весовой коэффициент выходного кодового слова.

г) Если декодирование осуществляется на основе алгоритма MAP, какие изменения, по-вашему, нужно внести в метрики состояний и метрики ветвей, если кодер не погашен?

8.14. а) Для нерекурсивного кодера, изображенного на рис. 38.1, вычислите минимальное расстояние по всему коду.

б) Для рекурсивного кодера, изображенного на рис. 8.26, вычислите минимальное расстояние по всему коду. Считайте, что исключений битов нет, а значит, степень кодирования кода равна 1/3.

в) Для кодера, показанного на рис. 8.26, обсудите изменения в весовом коэффициенте выходной последовательности, если вход каждого составного кодера определяется последовательностью с весом 2 (00...00100100...00) (считать, что исключений нет).

г) Повторите п. в для последовательности с весом 2 (00...0010100...00).

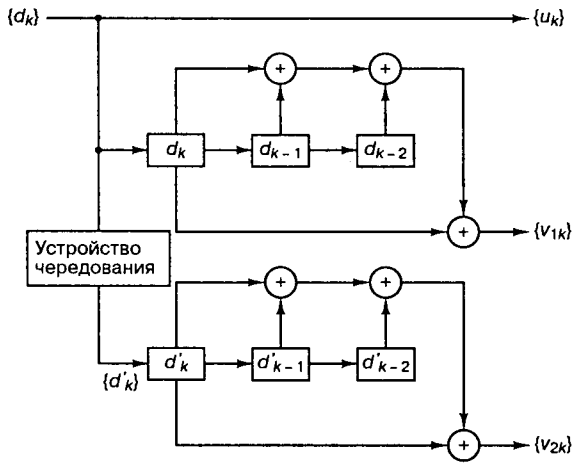


Рис. 38.1. Схема кодера с нерекурсивными составными кодами

8.15. Рассмотрим кодер, представленный на рис.8.25, *а*. Пусть он используется в качестве составного кода в турбокоде. Его решетчатая структура из 4 состояний изображена на рис.8.25, *б*. Степень кодирования кода равна 1/2. Ветвь, помеченная как u, v, представляет выходное ответвляющееся слово (кодовые биты) для той ветви, где u является битами данных (систематический код), а v — контрольными битами. Биты данных и контроля четности передаются за каждый такт k. Сигналы, принятые из демодулятора, имеют искаженные помехами значения u, v: 1,9; 0,7 — в момент $k = 1$ и $-0,4$; 0,8 — в момент $k = 2$. Предполагается, что априорные вероятности битов 1 и 0 равновероятны и что кодер начинает из нулевого состояния в начальный момент $k = 1$. Также считается, что дисперсия помех равна 1,3. Напомним, что последовательность N бит характеризуется N интервалами переходов и $N + 1$ состояниями (от начального до конечного). Следовательно, в данном случае биты стартуют в моменты $k = 1, 2$, и интерес представляют метрики в моменты $k = 1, 2, 3$.

а) Рассчитайте метрики ветвей для моментов $k = 1$ и $k = 2$, которые нужны для применения алгоритма MAP.

б) Вычислите прямые метрики состояний для моментов $k = 1, 2$ и 3.

в) Ниже для каждого правильного состояния в табл. 38.1 даются значения обратных метрик в моменты $k = 2$ и $k = 3$. Основываясь на данных таблицы и значениях, полученных в пп. а и б, вычислите значение логарифмического отношения правдоподобий, соответствующего битам данных в моменты $k = 1$ и $k = 2$. С помощью правила принятия решений MAP найдите наиболее вероятную последовательность информационных битов, которая могла быть передана.

Таблица 38.1

β_k^m	$k = 2$	$k = 3$
$m = a$	4,6	2,1
$m = b$	2,4	11,5
$m = c$	5,7	3,4
$m = d$	4,3	0,9

8.16. Пусть принятая последовательность, полученная в задаче 8.15 для кода со степенью кодирования 2/3, создается путем исключений из кода со степенью кодирования 1/2 (задаваемого решеткой на рис. 8.25, *б*). Исключение происходит таким образом, что передается только каждый второй контрольный бит. Таким образом, принятая последовательность из четырех сигналов представляет собой информационный бит, контрольный бит, информационный бит, информационный бит. Вычислите метрики ветвей и прямые метрики состояний для моментов времени $k = 1$ и $k = 2$, которые необходимы для алгоритма MAP.

8.17. Решетка для кода из четырех состояний используется как составной код в турбокоде, как показано на рис. 8.25, *б*. Степень кодирования равна 1/2, а ветвь, обозначенная как u, v, представляет собой выход, ответвляющееся слово (кодированные биты) для этой ветви, где u — это информационные биты, а v — биты четности. Из демодулятора принимается блок из $N = 1024$ фрагментов. Пусть первый сигнал из демодулятора поступает в момент $k = 1$ и в каждый последующий момент k поступают зашумленные биты данных и контроля четности. В момент времени $k = 1023$ принятые сигналы имеют зашумленные значения u, v, равные 1,3; –0,8, а в момент $k = 1024$ значения равны –1,4; –0,9. Предполагается, что априорные вероятности того, что биты данных равны 1 или 0, равны и что конечное состояние кодера будет $a = 00$ в завершающий момент $k = 1025$. Также считается, что дисперсия помех равна 2,5.

а) Рассчитайте метрики ветвей для моментов $k = 1023$ и $k = 1024$.

б) Рассчитайте обратные метрики состояний для моментов $k = 1023$, 1024 и 1025.

в) Ниже в табл. 38.2 даются значения прямых метрик состояний в моменты $k = 1023$ и $k = 1024$ для каждого правильного состояния. Основываясь на таблице и информации из пп. а и б, вычислите значения отношений правдоподобий, связанных с информационными битами в моменты времени $k = 1023$ и $k = 1024$. Используя правило принятия решений алгоритма MAP, найдите наиболее вероятную последовательность битов данных, которая могла быть передана.

Таблица 38.2

α_k^m	$k = 1023$	$k = 1024$
$m = a$	6,6	12,1
$m = b$	7,0	1,5
$m = c$	4,2	13,4
$m = d$	4,0	5,9

8.18. Имеется два статистически независимых наблюдения зашумленного сигнала, x_1 и x_2. Проверьте, что логарифмическое отношение правдоподобий (log-likelihood ratio — LLR) $L(d|x_1, x_2)$ можно выразить через индивидуальные LLR как

$$L(d|x_1, x_2) = L(x_1|d) + L(x_1|d) + L(d),$$

где $L(d)$ является априорным LLR основного бита данных d.

8.19. а) Используя теорему Байеса, подробно распишите этапы преобразования α_k^m, приведенной в уравнениях (8.129) и (8.130,б). *Подсказка*: воспользуйтесь упрощенной системой обозначений, применяемой в уравнениях (8.121) и (8.122).

б) Объясните, каким образом суммирование по состояниям m' в уравнении (8.130,а) дает в итоге уравнение (8.130,б).

в) Повторите п. а и покажите, как уравнение (8.133) переходит в уравнение (8.135). Также объясните, как суммирование по состояниям m' в будущий момент $k + 1$ дает уравнение (8.135).

8.20. Исходя из уравнения (8.139) для метрики ветви $\delta_k^{i,m}$, покажите, каким образом получается уравнение (8.140), и укажите, какие члены следует считать постоянными A_k в уравнении (8.140). Почему члены A_k не появляются в уравнении (8.140,а)?

8.21. Устройство чередования на рис. 8.27 (аналогичное устройству в соответствующем кодере) гарантирует, что выходная последовательность декодера DEC1 упорядочена во времени так же, как и последовательность $\{y_{2k}\}$. Можно ли реализовать это более простым способом? Что можно сказать о применении устройства восстановления в нижнем ряду? Не будет ли это более простым способом? Если это осуществить, тогда можно будет убрать два прежних устройства восстановления. Объясните, почему это не сработает.

8.22. При декодировании согласно алгоритму Витерби, используется устройство сложения, сравнения и выбора (add-compare-select — ACS). А при реализации алгоритма максимума апостериорной (maximum a posteriori — MAP) вероятности в турбодекодировании не применяется идея сравнения и выбора переходов. Вместо этого в алгоритме MAP рассматриваются все метрики ветвей и состояний для данного интервала времени. Объясните это принципиальное различие между двумя алгоритмами.

8.23. На рис. 38.2 показан рекурсивный систематический сверточный (recursive systematic convolutional — RSC) кодер со степенью кодирования 1/2, $K = 4$. Заметьте, что на рисунке используется формат памяти в виде 1-битовых блоков задержек (см. раздел 8.4.7.4). Следовательно, текущее состояние цепи можно описать с помощью уровней сигналов в точках a_{k-1}, a_{k-2}, a_{k-3}, аналогично способу описания состояния в формате разрядов памяти. Составьте таблицу, аналогичную табл. 8.5, которая будет определять все возможные переходы в цепи, и с ее помощью изобразите участок решетки.

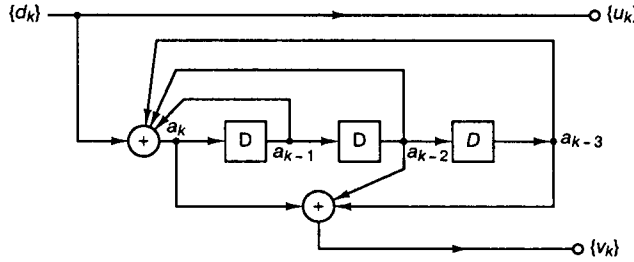


Рис. 38.2. Рекурсивный систематический сверточный (RSC) кодер со степенью кодирования 1/2, K = 4

8.24. На рис. 38.3 показан рекурсивный систематический сверточный (recursive systematic convolutional — RSC) кодер со степенью кодирования 2/3, $K = 3$. Заметьте, что на рисунке используется формат памяти в виде 1-битовых блоков задержек (см. раздел 8.4.7.4). Составьте таблицу, аналогичную табл. 8.5, которая будет определять все возможные переходы в цепи, и с ее помощью изобразите участок решетки. С помощью таблицы, подобной табл. 8.6, найдите выходное кодовое слово для информационной последовательности 1 1 0 0 1 1 0 0 1 1. В течение каждого такта данные поступают в цепь в виде пары $\{d_{1k}, d_{2k}\}$, а каждое выходное ответвляющееся слово $\{d_{1k}, d_{2k}, v_k\}$ образуется из пары битов данных и одного контрольного бита, v_k.

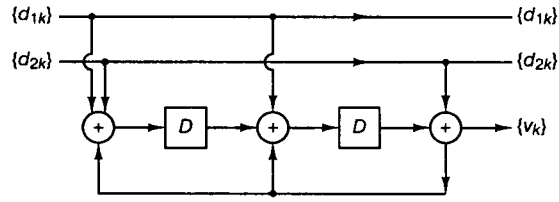


Рис. 38.3. Рекурсивный систематический сверточный (RSC) кодер со степенью кодирования 2/3, K = 3

8.25. Рассмотрим турбокод, состоящий из двух сверточных кодов, которые, в свою очередь, состоят из четырех состояний. Оба составных кода описываются решеткой, которая изображена на рис. 8.25, *б*. Степень кодирования кода равна 1/2, а длина блока — 12. Второй кодер оставлен не погашенным. Метрики ветвей, прямые метрики состояний и обратные метрики состояний для информационных битов, связанных с кодером в конечном состоянии, описываются матрицей, изображенной ниже. Принятый 12-сигнальный вектор образован из сигнала данных, сигнала контроля четности, сигнала данных, сигнала контроля четности и т.д. и имеет следующие значения.

1,2 1,3 −1,2 0,6 −0,4 1,9 −0,7 −1,9 −2,2 0,2 −0,1 0,6

Матрица ветвей $\delta_k^{i,m}$

$$\delta_k^{i,m} = \begin{bmatrix} \delta_1^{0,a} & \delta_2^{0,a} & \cdots & \delta_6^{0,a} \\ \delta_1^{1,a} & \ddots & \cdots & \delta_6^{1,a} \\ \delta_1^{0,b} & \ddots & \ddots & \vdots \\ \delta_1^{1,b} & \ddots & \ddots & \vdots \\ \delta_1^{0,c} & \ddots & \ddots & \vdots \\ \delta_1^{1,c} & \ddots & \ddots & \vdots \\ \delta_1^{0,d} & \ddots & \ddots & \vdots \\ \delta_1^{1,d} & \cdots & \cdots & \delta_6^{1,d} \end{bmatrix} = \begin{bmatrix} 1{,}00 & 1{,}00 & 1{,}00 & 1{,}00 & 1{,}00 & 1{,}00 \\ 3{,}49 & 0{,}74 & 2{,}12 & 0{,}27 & 0{,}37 & 1{,}28 \\ 1{,}00 & 1{,}00 & 1{,}00 & 1{,}00 & 1{,}00 & 1{,}00 \\ 3{,}49 & 0{,}74 & 2{,}12 & 0{,}27 & 0{,}37 & 1{,}28 \\ 1{,}92 & 1{,}35 & 2{,}59 & 0{,}39 & 1{,}11 & 1{,}35 \\ 1{,}82 & 0{,}55 & 0{,}82 & 0{,}70 & 0{,}33 & 0{,}95 \\ 1{,}92 & 1{,}38 & 2{,}59 & 0{,}39 & 1{,}11 & 1{,}35 \\ 1{,}82 & 0{,}55 & 0{,}82 & 0{,}70 & 0{,}33 & 0{,}95 \end{bmatrix}$$

Альфа-матрица (α_k^m)

$$\alpha_k^m = \begin{bmatrix} \alpha_1^a & \alpha_2^a & \cdots & \alpha_7^a \\ \alpha_1^b & \ddots & \cdots & \alpha_7^b \\ \alpha_1^c & \cdots & \ddots & \vdots \\ \alpha_1^d & \cdots & \cdots & \alpha_7^d \end{bmatrix} = \begin{bmatrix} 1{,}00 & 1{,}00 & 1{,}00 & 5{,}05 & 8{,}54 & 10{,}41 & 24{,}45 \\ 0{,}00 & 0{,}00 & 1{,}92 & 12{,}79 & 5{,}07 & 10{,}93 & 31{,}48 \\ 0{,}00 & 3{,}49 & 0{,}74 & 4{,}03 & 14{,}16 & 8{,}22 & 24{,}30 \\ 0{,}00 & 0{,}00 & 4{,}71 & 5{,}77 & 5{,}63 & 17{,}53 & 27{,}76 \end{bmatrix}$$

Бета-матрица (β_k^m)

$$\beta_k^m = \begin{bmatrix} \beta_1^a & \beta_2^a & \cdots & \beta_7^a \\ \beta_1^b & \ddots & \cdots & \beta_7^b \\ \beta_1^c & \cdots & \ddots & \vdots \\ \beta_1^d & \cdots & \cdots & \beta_7^d \end{bmatrix} = \begin{bmatrix} 24{,}45 & 5{,}44 & 2{,}83 & 1{,}12 & 1{,}00 & 1{,}00 & 1{,}00 \\ 24{,}43 & 5{,}62 & 3{,}17 & 0{,}70 & 0{,}37 & 1{,}28 & 0{,}00 \\ 21{,}32 & 5{,}45 & 3{,}53 & 0{,}81 & 0{,}43 & 0{,}00 & 0{,}00 \\ 21{,}31 & 5{,}79 & 2{,}75 & 1{,}14 & 1{,}42 & 0{,}00 & 0{,}00 \end{bmatrix}$$

Вычислите логарифмическое отношение правдоподобий для каждого из шести информационных битов $\{d_k\}$. С помощью правила принятия решений алгоритма MAP найдите наиболее вероятную последовательность информационных битов, которая могла быть передана.

Вопросы для самопроверки

8.1. Объясните высокую эффективность кодов Рида-Соломона при наличии *импульсных помех* (см. раздел 8.1.2.).

8.2. Объясните, почему кривые на рис. 8.6 показывают снижение достоверности передачи при низких значениях степеней кодирования (см раздел 8.1.3.).

8.3. Среди всех способов определения примитивности полинома наиболее простой — использование линейного регистра сдвига с обратной связью (linear feedback shift register — LFSR). Объясните эту процедуру (см. пример 8.2.).

8.4. Объясните, каким образом можно получить синдром, вычисляя принятый полином со всеми значениями корней полиномиального генератора кода (см. раздел 8.1.6.1).

8.5. Какое ключевое преобразование осуществляет система чередования/восстановления над импульсными помехами (см. раздел 8.2.1.)?

8.6. Почему предел Шеннона, равный –1,6 дБ, не представляет интереса при разработке реальных систем (см. раздел 8.4.5.2.)?

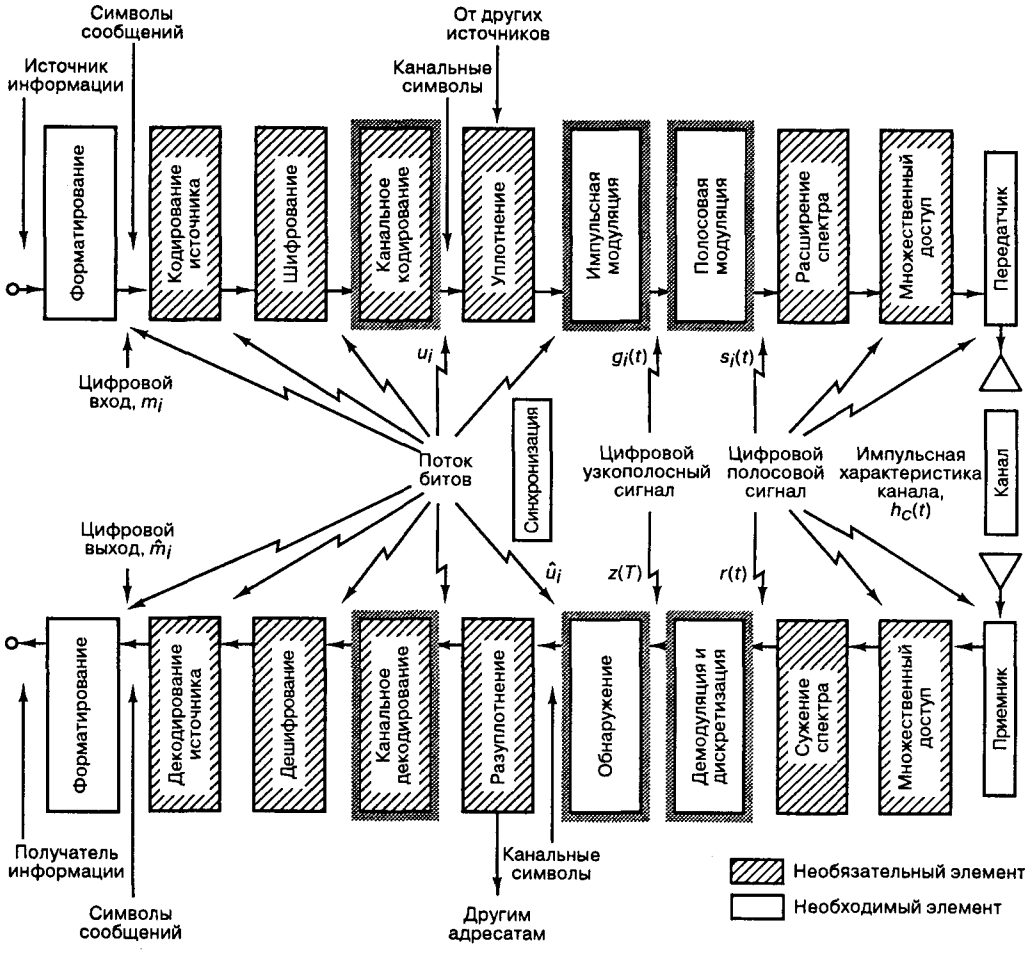
8.7. Почему алгоритм декодирования Витерби не дает апостериорных вероятностей (см. раздел 8.4.6.)?

8.8. Каково более описательное название алгоритма Витерби (см. раздел 8.4.6.)?

8.9. Опишите сходство и отличие в реализации декодирования на основе алгоритмов Витерби и максимума апостериорной вероятности (MAP) (см. раздел 8.4.6.).

ГЛАВА 9

Компромиссы при использовании модуляции и кодирования



9.1. Цели разработчика систем связи

Системные компромиссы — это неотъемлемая часть всех разработок цифровых систем связи. Разработчик должен стремиться к 1) увеличению скорости передачи бит R до максимально возможной; 2) минимизации вероятности появления битовой ошибки P_B; 3) минимизации потребляемой мощности, или, что то же самое, минимизации требуемого отношения энергии одного бита к спектральной плотности мощности шума E_b/N_0; 4) минимизации ширины полосы пропускания W; 5) максимизации эффективности использования системы, т.е. к обеспечению надежного обслуживания для максимального числа пользователей с минимальными задержками и максимальной устойчивостью к возникновению конфликтов; и 6) минимизации конструктивной сложности системы, вычислительной нагрузки и стоимости системы. Конечно, разработчик системы может попытаться удовлетворить все требования одновременно. Однако очевидно, что требования 1 и 2 противоречат требованиям 3 и 4; они предусматривают одновременное увеличение скорости R и минимизацию P_B, E_b/N_0, W. Существует несколько сдерживающих факторов и теоретических ограничений, которые неизбежно влекут за собой компромиссы в любых системных требованиях.

Минимальная теоретически требуемая ширина полосы частот по Найквисту
Теорема о пропускной способности Шеннона-Хартли (и предел Шеннона)
Государственное регулирование (например, распределение частот)
Технологические ограничения (например, современные комплектующие)
Другие системные требования (например, орбиты спутников)

Некоторые реализуемые компромиссы между кодированием и модуляцией можно лучше показать через изменение положения рабочей точки на одной из двух плоскостей — характеристике вероятности появления ошибки и характеристике эффективности использования полосы частот; обе описываются в следующих разделах.

9.2. Характеристика вероятности появления ошибки

На рис. 9.1 показаны семейства кривых зависимости P_B от E_b/N_0 для когерентного обнаружения ортогональных (рис. 9.1, *а*) и многофазных сигналов (рис. 9.1, *б*). Для представления каждой k-битовой последовательности модулятор использует один из $M = 2^k$ сигналов, где M — размер набора символов. На рис. 9.1, *а* изображено потенциальное снижение частоты появления ошибок с повышением k (или M) при передаче ортогональных сигналов. Для ортогональных наборов сигналов, таких как сигналы в ортогональной частотной манипуляции (frequency shift keying — FSK), увеличение размера набора символов может дать снижение P_B, или требуемого E_b/N_0, за счет увеличения полосы пропускания. На рис. 9.1, *б* показано повышение частоты появления ошибок с увеличением k (или M) при передаче неортогональных сигналов. Для наборов неортогональных сигналов, таких как сигналы в многофазной манипуляции (multiple phase shift keying — MPSK), расширение набора символов может снизить требования к полосе пропускания за счет повышения P_B, или требуемого значения E_b/N_0. Далее эти семейства кривых (рис. 9.1, *а* или *б*) будут называться *кривыми характеристик вероятности появления ошибок*, а плоскость, в которой они лежат, — *плоскостью вероятности появления ошибок*. Такие характеристики показывают, где может располагаться рабочая точка для конкретных схем модуляции и кодирования. Для системы с данной скоростью передачи

информации каждую кривую на плоскости можно связать с различными фиксированными значениями минимально необходимой полосы пропускания; а значит, некое множество кривых можно представить как множество *кривых равной полосы пропускания*. При передвижении по кривой в направлении возрастания ординаты, ширина полосы пропускания, необходимая для передачи, увеличивается; и напротив, если перемещаться в обратном направлении, то требуемая полоса пропускания уменьшится. После выбора схемы модуляции и кодирования, а также номинального значения E_b/N_0 функционирование системы характеризуется конкретной точкой на плоскости вероятности появления ошибок. Возможные компромиссы можно рассматривать как изменение рабочей точки на одной из кривых или как переход с рабочей точки одной кривой семейства в рабочую точку другой. Эти компромиссы изображены на рис. 9.1 *а* и *б* как смещения рабочей точки системы в направлении, указанном стрелками. Перемещение рабочей точки вдоль линии 1 между точками *a* и *b* можно считать компромиссом между P_B и характеристикой E_b/N_0 (при фиксированном значении W). Аналогично сдвиг вдоль линии 2, между точками *c* и *d*, является поиском компромисса между P_B и W (при фиксированном значении E_b/N_0). И наконец, перемещение вдоль линии 3, между точками *e* и *f*, представляет собой поиск компромисса между W и E_b/N_0 (при фиксированном значении P_B). Сдвиг вдоль линии 1 — это снижение или повышение номинального значения E_b/N_0. Этого можно достичь, например, путем повышения мощности передатчика; это означает, что компромисс можно осуществить просто "поворотом регулятора" даже после завершения конфигурации системы. В то же время другие компромиссы (сдвиги вдоль линий 2 или 3) включают изменения в схеме модуляции или кодирования, а значит, их следует осуществлять на этапе разработки системы. Изменять тип модуляции и кодирования в системе программным путем можно будет с помощью *программных средств связи* [1].

9.3. Минимальная ширина полосы пропускания по Найквисту

В любой реализуемой системе, выполняющей неидеальную фильтрацию, будет межсимвольная интерференция (intersymbol interference, ISI) — хвост одного импульса распространяется на соседние символы и мешает процессу обнаружения. Найквист [2] показал, что теоретическая минимальная ширина полосы пропускания (ширина полосы частот по Найквисту), требуемая для немодулированной передачи R_s символов за секунду без межсимвольной интерференции, составляет $R_s/2$ Гц. Это основное теоретическое ограничение, вынуждающее разработчика настолько аккуратно использовать полосу частот, насколько это возможно (см. раздел 3.3). На практике минимальная ширина полосы частот по Найквисту увеличивается на 10–40% вследствие ограничений реальных фильтров. Таким образом, *реальная* пропускная способность цифровых систем связи снижается с идеальных 2 символа/с/Гц до 1,8–1,4 символа/с/Гц. Из набора M символов, система модуляции или кодирования присваивает каждому символу k-битовое значение, где $M = 2^k$. Таким образом, число битов на символ можно представить как $k = \log_2 M$, и, следовательно, скорость передачи данных, или скорость передачи битов R, должна быть в k раз больше скорости передачи символов R_s, как видно из следующего основного соотношения.

$$R = kR_s \quad \text{или} \quad R_s = \frac{R}{k} = \frac{R}{\log_2 M} \tag{9.1}$$

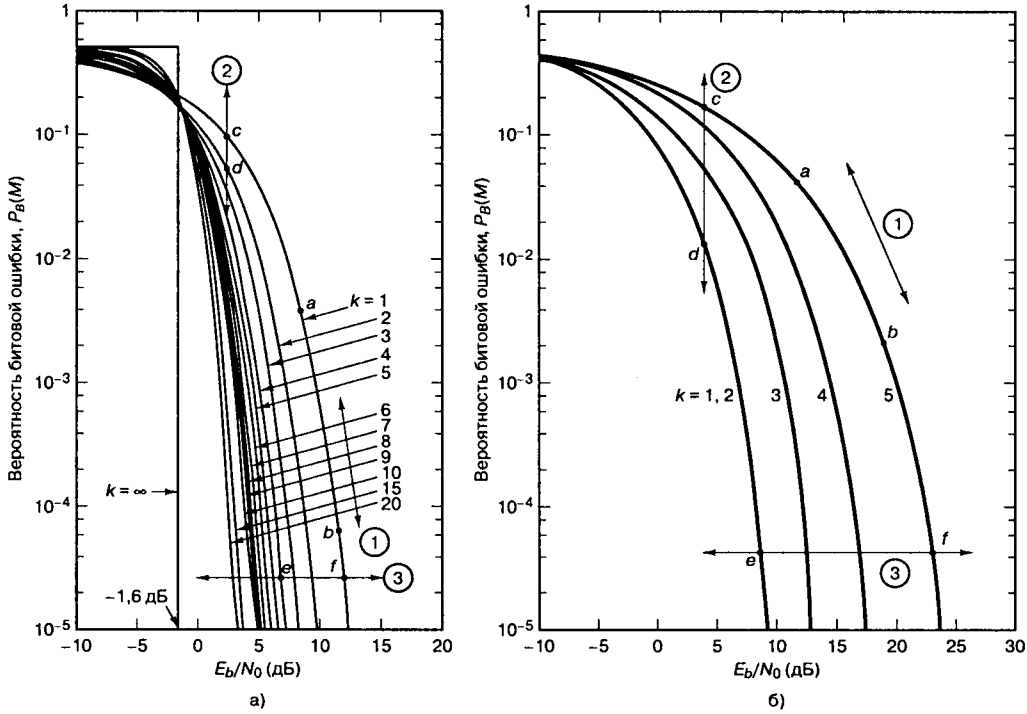


Рис. 9.1. Зависимость вероятности появления битовой ошибки от E_b/N_0 при когерентном обнаружении M-арных сигналов: а) ортогональные сигналы; б) многофазные сигналы

Для системы с фиксированной скоростью передачи символов из выражения (9.1) видно, что с ростом k увеличивается и скорость передачи битов R. При использовании схемы MPSK с увеличением k повышается эффективность использования полосы частот R/W, измеряемая в бит/с/Гц. Например, сдвиг вдоль линии 3, из точки e в точку f, как видно на рис. 9.1, *б*, представляет собой повышение E_b/N_0 за счет снижения требований к полосе пропускания. Другими словами, при той же полосе пропускания сигналы, модулированные MPSK, можно передавать с повышенной скоростью передачи данных, а значит с увеличенным R/W.

Пример 9.1. Классификация схем цифровой модуляции

В некотором смысле все схемы цифровой модуляции относятся к одному из двух классов с противоположными характеристиками. Первый класс — это передача ортогональных сигналов; достоверность такой передачи описывается кривыми на рис. 9.1, *а*. Ко второму классу относится передача неортогональных сигналов (набор векторов сигналов можно отобразить на плоскости). На рис. 9.1, *б* представлен пример передачи неортогональных сигналов — модуляция MPSK. Вообще, любая фазовая/амплитудная модуляция (например, QAM) относится ко второму классу. Используя рис. 9.1, ответьте на следующие вопросы.

а) Как в случае M-арной передачи сигналов будет меняться достоверность передачи (увеличиваться или снижаться) при повышении M?

б) Возможности выбора в цифровой связи почти всегда сопряжены с компромиссами. Если достоверность передачи растет, то за счет чего?

в) Если растет вероятность появления ошибки, то какую выгоду можно получить из этого?

Решение

а) Из рис. 9.1 можно видеть, что повышение или снижение достоверности передачи зависит от рассматриваемого класса передачи сигналов. Рассмотрим передачу ортогональных сигналов (рис. 9.1, *а*), где достоверность передачи растет с увеличением k или M. Напомним, что существует лишь два способа сравнения подобных характеристик достоверности передачи. Можно провести вертикальную линию при некотором фиксированном значении E_b/N_0 и увидеть, что при уменьшении k или M P_B снижается. Или наоборот, можно провести горизонтальную линию, фиксирующую некоторое значение P_B, и с ростом k или M отметить снижение требований к E_b/N_0. Точно так же на рис. 9.1, *б* можно видеть, что при неортогональной передаче, такой как модуляция MPSK, характеристики ведут себя абсолютно иначе. Достоверность передачи снижается при увеличении k или M.

б) Чем мы платим за передачу ортогональных сигналов, при которой достоверность передачи повышается с ростом k или M? Наиболее распространенным вариантом передачи ортогональных сигналов является схема MFSK, где $k = 1$ и $M = 2$, а набор тонов состоит из двух сигналов. Если $k = 2$ и $M = 4$, в наборе уже содержится четыре тона. При $k = 3$ и $M = 8$ будет уже восемь сигналов и т.д. При использовании схемы MFSK за время передачи символа отсылается только один тон, но доступная полоса пропускания — это весь набор тонов. Следовательно, при увеличении k или M за повышение достоверности передачи придется платить расширением требуемой полосы пропускания.

в) При передаче неортогональных сигналов (схема MPSK или QAM) с ростом k или M достоверность передачи падает. Логично предположить, что компромисс повлечет за собой снижение требований к полосе пропускания. Рассмотрим следующий пример. Пусть требуется скорость передачи данных $R = 9600$ бит/с, а в качестве модуляции используется 8-уровневая схема PSK. Тогда с помощью уравнения (9.1) скорость передачи символов находится следующим образом.

$$R_s = \frac{R}{\log_2 M} = \frac{9600 \text{ бит/с}}{3 \text{ бит/символ}} = 3200 \text{ символов/с}$$

Если для передачи воспользоваться 16-уровневой схемой PSK, то скорость передачи символов будет равна следующему.

$$R_s = \frac{9600 \text{ бит/с}}{4 \text{ бит/символ}} = 2400 \text{ символов/с}$$

Если применить 32-уровневую схему PSK, скорость передачи символов будет равна

$$R_s = \frac{9600 \text{ бит/с}}{5 \text{ бит/символ}} = 1920 \text{ символов/с}.$$

Что происходит, когда на рис. 9.1, *б* рабочая точка сдвигается вдоль горизонтальной линии с кривой с $k = 3$ на кривую с $k = 4$ и далее на кривую с $k = 5$? При данной скорости передачи данных и вероятности появления ошибки каждый такой сдвиг позволяет осуществлять передачу на все более низких скоростях. Всякий раз, когда говорится "более низкая скорость передачи сигнала", это эквивалентно сообщению, что имеется возможность уменьшить ширину полосы пропускания. Аналогично любое повышение скорости передачи сигналов соответствует увеличению ширины полосы пропускания.

9.4. Теорема Шеннона-Хартли о пропускной способности канала

Шеннон [3] показал, что пропускная способность канала C с аддитивным белым гауссовым шумом (additive white Gaussian noise — AWGN) является функцией средней мощности принятого сигнала S, средней мощности шума N и ширины полосы пропускания W. Выражение для пропускной способности (теорема Шеннона-Хартли) можно записать следующим образом.

$$C = W\log_2\left(1+\frac{S}{N}\right) \tag{9.2}$$

Если W измеряется в герцах, а логарифм берется по основанию 2, то пропускная способность будет иметь размерность бит/с. Теоретически (при использовании достаточно сложной схемы кодирования) информацию по каналу можно передавать с любой скоростью R ($R \le C$) со *сколь угодно малой* вероятностью возникновения ошибки. Если же $R > C$, то кода, на основе которого можно добиться сколь угодно малой вероятности возникновения ошибки, не существует. В работе Шеннона показано, что величины S, N и W *устанавливают пределы скорости передачи, а не вероятности появления ошибки.* Шеннон [4] использовал уравнение (9.2) для графического представления доступных пределов производительности прикладных систем. Этот график, показанный на рис. 9.2, представляет нормированную пропускную способность канала C/W в бит/с/Гц как функцию отношения сигнал/шум (signal-to-noise ratio — SNR) в канале. График, представленный на рис. 9.3, изображает зависимость нормированной полосы пропускания канала W/C в бит/с/Гц от отношения сигнал/шум канала. Иногда рис. 9.3 используется как иллюстрация компромисса между мощностью и полосой пропускания, присущего идеальному каналу. Однако это не совсем компромисс [5], поскольку мощность обнаруженного шума пропорциональна полосе пропускания.

$$N = N_0 W \tag{9.3}$$

Подставив выражение (9.3) в уравнение (9.2) и немного преобразовав последнее, получаем следующее.

$$\frac{C}{W} = \log_2\left(1+\frac{S}{N_0 W}\right) \tag{9.4}$$

Если битовая скорость передачи равна пропускной способности канала ($R = C$), то с помощью тождества (3.30) можно записать следующее.

$$\frac{S}{N_0 C} = \frac{E_b}{N_0} \tag{9.5}$$

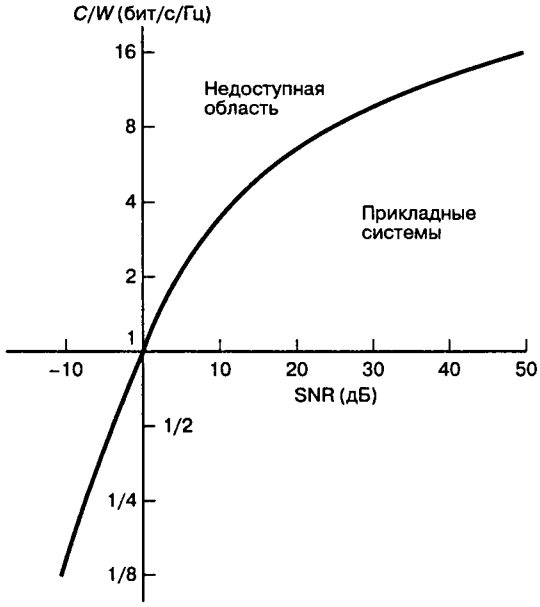


Рис. 9.2. Зависимость нормированной пропускной способности канала от SNR канала

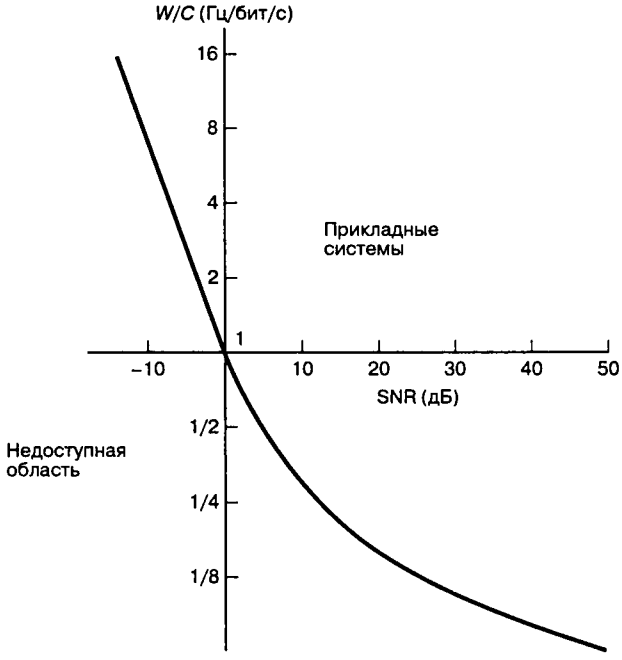


Рис. 9.3. Зависимость нормированной полосы пропускания канала от SNR канала

Таким образом, уравнение (9.4) можно модифицировать следующим образом.

$$\frac{C}{W} = \log_2\left[1 + \frac{E_b}{N_0}\left(\frac{C}{W}\right)\right] \tag{9.6,а}$$

$$2^{C/W} = 1 + \frac{E_b}{N_0}\left(\frac{C}{W}\right) \tag{9.6,б}$$

$$\frac{E_b}{N_0} = \frac{W}{C}(2^{C/W} - 1) \tag{9.6,в}$$

На рис. 9.4 представлен график зависимости W/C от E_b/N_0, описываемой формулой (9.6,в); асимптотическое поведение этой кривой при $C/W \to 0$ (или $W/C \to \infty$) рассматривается в следующем разделе.

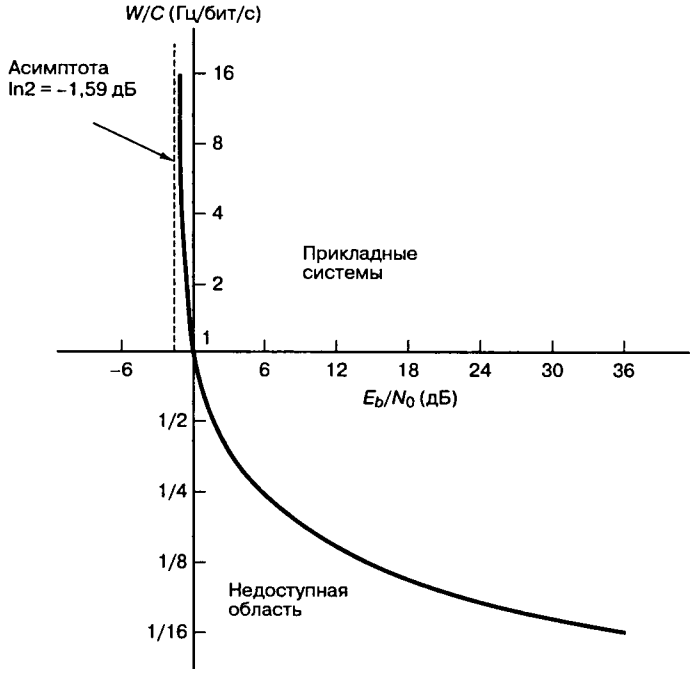


Рис. 9.4. Зависимость нормированной полосы пропускания канала от E_b/N_0

9.4.1. Предел Шеннона

Существует нижнее предельное значение E_b/N_0, при котором ни при какой скорости передачи нельзя осуществить безошибочную передачу информации. С помощью соотношения

$$\lim_{x \to 0}(1+x)^{1/x} = e$$

можно рассчитать граничное значение E_b/N_0.

Пусть

$$x = \frac{E_b}{N_0}\left(\frac{C}{W}\right)$$

Тогда, из уравнения (9.6,а),

$$\frac{C}{W} = x\log_2(1+x)^{1/x}$$

и

$$1 = \frac{E_b}{N_0}\log_2(1+x)^{1/x}$$

В пределе, при $C/W \to 0$, получаем

$$\frac{E_b}{N_0} = \frac{1}{\log_2 e} = 0{,}693 \tag{9.7}$$

или, в децибелах,

$$\frac{E_b}{N_0} = -1{,}6 \text{ дБ}$$

Это значение E_b/N_0 называется *пределом Шеннона* (Shannon limit). На рис. 9.1, *а* предел Шеннона — это кривая зависимости P_B от E_b/N_0 при $k \to \infty$. При $E_b/N_0 = -1{,}6$ данная кривая скачкообразно изменяет свое значение с $P_B = 1/2$ на $P_B = 0$. В действительности достичь предела Шеннона невозможно, поскольку k возрастает неограниченно, а с ростом k возрастают требования к полосе пропускания и повышается сложность реализации системы. Работа Шеннона — это теоретическое доказательство существования кодов, которые могут улучшить P_B или снизить требуемое значение E_b/N_0 от уровней некодированных двоичных схем модуляции до уровней, приближающихся к предельной кривой. При вероятности появления битовой ошибки 10^{-5} двоичная фазовая манипуляция (binary phase-shift-keying — BPSK) требует значения E_b/N_0, равного 9,6 дБ (оптимум некодированной двоичной модуляции). Следовательно, в данном случае в работе Шеннона указано, что теоретически, за счет использования кодирования, производительность можно повысить на 11,2 дБ по сравнению с некодированной двоичной модуляцией. В настоящее время большую часть такого улучшения (почти 10 дБ) можно получить с помощью турбокодов (см. раздел 8.4). Оптимальную разработку системы можно наилучшим образом представить как поиск рациональных компромиссов среди различных ограничений и взаимно противоречивых требований. Компромиссы модуляции и кодирования, т.е. выбор конкретных схем модуляции и кодирования для наилучшего использования переданной мощности и ширины полосы, являются очень важными, поскольку имеется много причин для снижения мощности, а также существует необходимость экономии спектра радиочастот.

9.4.2. Энтропия

Для разработки системы связи с определенной способностью к обработке сообщений нужна метрика измерения объема передаваемой информации. Шеннон [3] ввел такую метрику H, называемую энтропией источника сообщений (имеющего n возможных

выходных значений). *Энтропия* (entropy) определяется как среднее количество информации, приходящееся на один выход источника, и выражается следующим образом.

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \quad \text{бит/выход источника} \tag{9.8}$$

Здесь p_i — вероятность i-го выходного значения и $\Sigma p_i = 1$. Если сообщение двоичное или источник имеет только два возможных выходных значения с вероятностями p и $q = (1 - p)$, выражение для энтропии примет следующий вид.

$$H = -(p \log_2 p + q \log_2 q) \tag{9.9}$$

Зависимость энтропии от p показана на рис. 9.5.

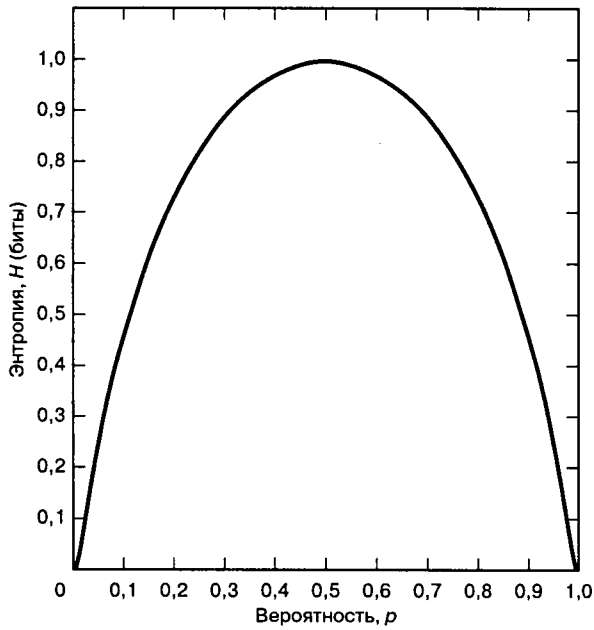


Рис. 9.5. Зависимость энтропии от вероятности (два события)

Величина H имеет ряд особенностей.

1. Если логарифм в уравнении (9.8) берется по основанию 2, единица измерения H — среднее число бит на событие. Здесь единица измерения *бит* — это мера *количества информации*, и ее не следует путать с термином "бит", означающим "двоичная цифра" (binary digit — bit).
2. Сам термин "энтропия" имеет несколько неопределенный смысл, что вызвано наличием нескольких формулировок в статистической механике. Для информационного источника с двумя равновероятными состояниями (например, выбрасывание монеты правильной формы) из рис. 9.5 видно, что неопределенность исхода и, следовательно, среднее количество информации максимальны. Как

только вероятности уходят от равновероятного состояния, среднее количество информации снижается. В пределе, когда одна из вероятностей обращается в нуль, H также обращается в нуль. Результат известен до того, как произойдет событие, так что исход не несет в себе дополнительной информации.

3. Для иллюстрации связи между количеством информации и априорной вероятностью (если априорная вероятность сообщения на приемнике является нулем или единицей, сообщение можно не посылать) рассмотрим следующий пример. После девятимесячной беременности женщина оказывается в родильной палате. Муж с волнением ждет в приемной. Через некоторое время к нему подходит врач и говорит: "Примите мои поздравления, вы стали отцом". Какую информацию отец получил от врача *после медицинского исхода*? Почти никакой; отец практически достоверно знал, что ребенок должен родиться. Если бы врач сказал, "вы стали отцом мальчика" или "вы стали отцом девочки", он передал бы 1 бит информации, поскольку существует 50% вероятность того, что ребенок окажется девочкой или мальчиком.

Пример 9.2. Среднее количество информации в английском языке

а) Найдите среднее количество информации в бит/знак для английского языка, считая, что каждая из 26 букв алфавита появляется с равной вероятностью. Пробелы и знаки пунктуации не учитываются.

б) Поскольку буквы в английском языке (или каком-либо ином) появляются с различной частотой, ответ на п. а — это верхняя граница среднего количества информации на знак. Повторите п. а, считая, что буквы алфавита появляются со следующими вероятностями.

$p = 0{,}10$: для букв a, e, o, t

$p = 0{,}07$: для букв h, i, n, r, s

$p = 0{,}02$: для букв c, d, f, l, m, p, u, y

$p = 0{,}01$: для букв b, g, j, k, q, v, w, x, z

Решение

а) $H = -\sum_{i=1}^{26} \frac{1}{26} \log_2\left(\frac{1}{26}\right) = 4{,}7$ бит/знак

б) $H = -(4 \times 0{,}1 \log_2 0{,}1 + 5 \times 0{,}07 \log_2 0{,}07 + 8 \times 0{,}02 \log_2 0{,}02 + 9 \times 0{,}01 \log_2 0{,}01) = 4{,}17$ бит/знак

Если 26 букв алфавита нужно выразить в некоторой двоичной схеме кодирования, то для каждой буквы требуется пять двоичных цифр. Пример 9.2 показывает, что должен существовать способ кодирования английского текста *в среднем* меньшим числом двоичных цифр для одной буквы (среднее количество информации, содержащееся в каждом знаке, меньше 5 бит). Подробнее тема кодирования источника будет рассмотрена в главе 13.

9.4.3. Неоднозначность и эффективная скорость передачи информации

Пусть по двоичному симметричному каналу (определенному в разделе 6.3.1) со скоростью 1000 двоичных символов/с происходит передача информации, а априорная вероятность передачи нуля или единицы одинакова. Допустим также, что помехи в канале

настолько значительны, что, независимо от переданного символа, вероятность приема единицы равна 1/2 (то же самое — для нуля). В таком случае половина принятых символов должна *случайно* оказаться правильной, и может создаться впечатление, что система обеспечивает скорость 500 бит/с, хотя на самом деле никакой информации не передается. Одинаково "хороший" прием дает и использование "информации", поступившей из канала, и генерация этой "информации" методом подбрасывания правильной монеты. Утраченной является информация о корректности переданных символов. Для оценки неопределенности в принятом сигнале Шеннон [3] использует поправочный коэффициент, который называет *неоднозначностью* (equivocation). Неоднозначность определяется как *условная энтропия* сообщения X, обусловленная данным сообщением Y, или

$$\begin{aligned} H(X|Y) &= -\sum_{X,Y} P(X|Y)\log_2 P(X|Y) = \\ &= -\sum_{Y} P(Y)\sum_{X} P(X|Y)\log_2 P(X|Y) \end{aligned}, \tag{9.10}$$

где X — сообщение, переданное источником, Y — принятый сигнал, $P(X, Y)$ — совместная вероятность X и Y, а $P(X|Y)$ — условная вероятность X при приеме Y. Неоднозначность можно представить как неуверенность в передаче X при условии принятия Y. Для *канала без ошибок* $H(X|Y) = 0$, поскольку принятие сообщения Y абсолютно точно определяет X. В то же время для канала с ненулевой вероятностью возникновения символьной ошибки $H(X|Y) > 0$, поскольку канал вносит некоторую неопределенность. Рассмотрим двоичную последовательность X, для которой априорные вероятности источника $P(X = 1) = P(X = 0) = 1/2$ и где, в среднем, в принятую последовательность из 100 бит канал вносит одну ошибку ($P_B = 0{,}01$). Исходя из уравнения (9.10), неоднозначность $H(X|Y)$ можно записать следующим образом.

$$\begin{aligned} H(X|Y) &= -[(1 - P_B)\log_2(1 - P_B) + P_B \log_2 P_B] = \\ &= -(0{,}99 \log_2 0{,}99 + 0{,}01 \log_2 0{,}01) = \\ &= 0{,}081 \text{ бит/полученный символ} \end{aligned}$$

Таким образом, в каждый принятый символ канал вносит 0,081 бит неопределенности.

Шеннон показал, что среднее эффективное количество информации H_{eff} в приемнике получается путем вычитания неоднозначности из энтропии источника. Следовательно,

$$H_{\text{eff}} = H(X) - H(X|Y) \tag{9.11}$$

Для системы, передающей равновероятные двоичные символы, энтропия $H(X)$ равна 1 бит/символ. Если символы принимаются с $P_B = 0{,}01$, неоднозначность, как показано выше, равна 0,081 бит/(принятый символ). Тогда, используя уравнение (9.11), можем записать эффективную энтропию H_{eff} принятого сигнала.

$$H_{\text{eff}} = 1 - 0{,}081 = 0{,}919 \text{ бит/полученный символ}$$

Иными словами, если, например, за секунду передается $R = 1000$ двоичных символов, то R_{eff} можно выразить следующим образом.

$$R_{\text{eff}} = RH_{\text{eff}} = 1000 \text{ символов/с} \times 0{,}919 \text{ бит/символ} = 919 \text{ бит/с} \tag{9.12}$$

Отметим, что в предельном случае $P_B = 0{,}5$

$$H(X|Y) = -(0{,}5 \log_2 0{,}5 + 0{,}5 \log_2 0{,}5) = 1 \text{ бит/символ}$$

Используя формулы (9.12) и (9.11) при $R = 1000$ символов/с, получаем

$$R_{\text{eff}} = 1000 \text{ символов/с } (1 - 1) = 0 \text{ бит/с},$$

что и следовало ожидать.

Пример 9.3. Кажущееся противоречие с пределом Шеннона

График зависимости P_B от E_b/N_0 обычно показывает плавный рост P_B при увеличении E_b/N_0. Например, кривые вероятности появления битовых ошибок на рис. 9.1 показывают, что в пределе при E_b/N_0, стремящемся к нулю, P_B *стремится* к 0,5. Таким образом, кажется, что всегда (при сколь угодно малом значении E_b/N_0) имеется ненулевая скорость передачи информации. На первый взгляд *это не согласуется* с величиной предела Шеннона $E_b/N_0 = -1{,}6$ дБ, ниже которого невозможна безошибочная передача информации или ниже которого даже бесконечная полоса пропускания дает конечную скорость передачи информации (см. рис. 9.4).

а) Предложите способ разрешения кажущегося противоречия.

б) Покажите, каким образом коррекция неоднозначности по Шеннону может помочь разрешить данное противоречие для двоичной системы с модуляцией PSK, если энтропия источника равна 1 бит/символ. Предположим, что рабочая точка на рис. 9.1, *б* соответствует $E_b/N_0 = 0{,}1$ (−10 дБ).

Решение

а) Величина E_b, традиционно используемая при расчетах каналов в прикладных системах, — это энергия принятого сигнала, приходящаяся на *переданный символ*. Однако E_b в уравнении (9.6) — это энергия сигнала, приходящаяся на один бит *принятой информации*. Для разрешения описанного выше кажущегося противоречия следует учитывать потери информации, вызываемые помехами канала.

б) На основе уравнения (4.79) для BPSK можно записать

$$P_B = Q(\sqrt{2E_b/N_0}) = Q(0{,}447),$$

где Q определено в формуле (3.43) и представлено в табличной форме в приложении Б (табл. Б.1). Из таблицы находим, что $P_B = 0{,}33$. Далее находим неоднозначность и эффективную энтропию.

$$\begin{aligned} H(X|Y) &= -[(1 - P_B) \log_2 (1 - P_B) + P_B \log_2 P_B] = \\ &= -(0{,}67 \log_2 0{,}67 + 0{,}33 \log_2 0{,}33) = \\ &= 0{,}915 \text{ бит/символ} \end{aligned}$$

$$\begin{aligned} H_{\text{eff}} &= H(X) - H(X|Y) = \\ &= 1 - 0{,}915 = \\ &= 0{,}085 \text{ бит/символ} \end{aligned}$$

Следовательно,

$$\left(\frac{E_b}{N_0}\right)_{\text{eff}} = \frac{(E_b/N_0) \text{ джоуль на символ/ватт на символ}}{H_{\text{eff}} \text{ бит/символ}} =$$

$$= \frac{0{,}1}{0{,}085} = 1{,}176 \frac{\text{джоуль на бит}}{\text{ватт/Гц}} =$$

$$= 0{,}7 \text{ дБ}$$

Таким образом, эффективное значение E_b/N_0 равно 0,7 дБ на принятый информационный бит, что значительно больше предела Шеннона −1,6 дБ.

9.5. Плоскость "полоса-эффективность"

С помощью уравнения (9.6) можно составить график зависимости нормированной полосы пропускания канала W/C (в Гц/бит/с) от E_b/N_0, как показано на рис. 9.4. Здесь в качестве независимой переменной взято E_b/N_0 и можно видеть *компромисс между активной мощностью и полосой пропускания*, так сказать, в деле. Можно показать [5], что качественно спроектированные системы должны стремиться к работе в области излома кривой компромисса между полосой пропускания и мощностью для идеального ($R = C$) канала. Характеристики реальных систем часто отличаются от идеальных не более чем на 10 дБ. Наличие излома означает, что в системах, в которых предпринимается попытка уменьшить занимаемую полосу пропускания канала или снизить требуемую мощность, приходится все больше повышать значение другого параметра (что является не очень желательным). Например, возвращаясь к рис. 9.4, можно сказать, что идеальная система, работающая при $E_b/N_0 = 1{,}8$ дБ и использующая полосу частот с нормированной шириной 0,5 Гц/бит/с, для уменьшения используемой полосы частот до 0,1 Гц/бит/с должна поднять E_b/N_0 до 20 дБ. Подобное будет происходить и при попытке компромисса в обратную сторону.

С помощью уравнения (9.6,в) можно также получить зависимость C/W от E_b/N_0. Она показана на графике зависимости R/W от E_b/N_0 (рис. 9.6). Обозначим эту плоскость как плоскость "полоса-эффективность". Ордината R/W — это мера объема данных, которые можно передать через единицу полосы частот за данное время; следовательно, она отображает эффективность использования ресурса полосы пропускания. Независимая переменная E_b/N_0 измеряется в децибелах. На рис. 9.6 кривая $R = C$ — это граница, разделяющая область реальных прикладных систем связи и область, в которой такие системы связи теоретически невозможны. Подобно изображенной на рис. 9.2, характеристика эффективности полосы пропускания на рис. 9.6 устанавливает предельные параметры, которые достижимы для прикладных систем. Поскольку в качестве независимой переменной более предпочтительно E_b/N_0, чем SNR, рис. 9.6 удобнее рис. 9.2 с точки зрения сравнения компромиссов кодирования и модуляции в цифровой связи. Отметим, что на рис. 9.6 проиллюстрирована зависимость эффективности использования полосы частот от E_b/N_0 для систем с одной несущей. Для систем с множественными несущими эффективность использования полосы частот зависит от разнесения несущих (и типа модуляции). В этом случае компромисс — это насколько разнесены несущие (что приводит к повышению эффективности использования полосы частот) без возникновения неприемлемых помех соседних каналов (adjacent channel interference — ACI).

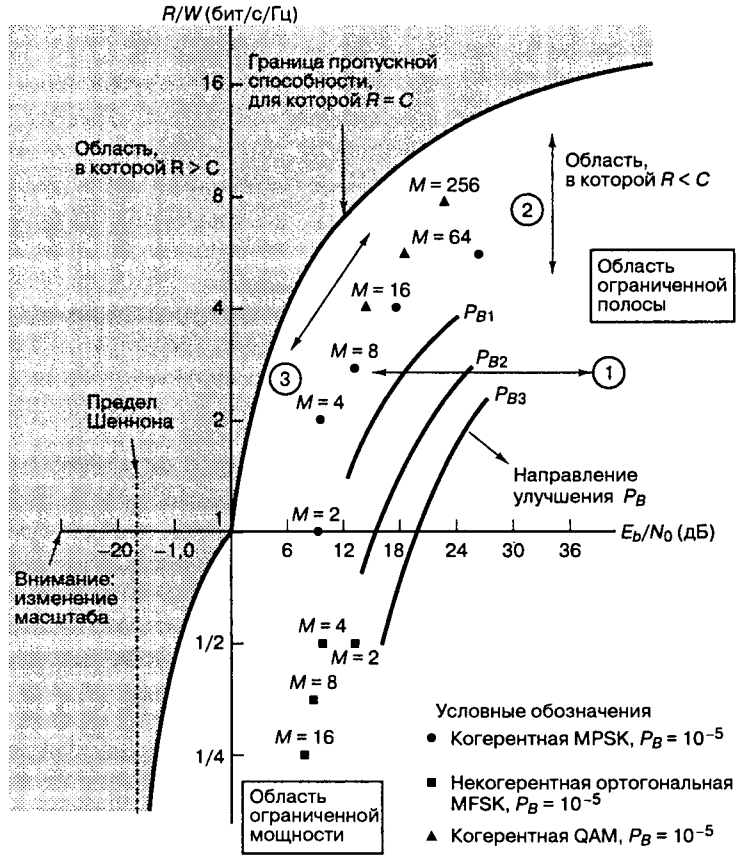


Рис. 9.6. Плоскость "полоса-эффективность"

9.5.1. Эффективность использования полосы при выборе схем MPSK и MFSK

На рис. 9.6 показаны рабочие точки для когерентной модуляции MPSK при вероятности битовой ошибки 10^{-5}. Предполагается, что до модуляции осуществляется фильтрация по Найквисту (идеальная прямоугольная), так что минимальная двойная полоса пропускания на промежуточной частоте (intermediate frequency — IF) $W_{IF} = 1/T$, где T — длительность символа. Таким образом, из уравнения (9.1) эффективность использования полосы частот $R/W = \log_2 M$, где M — размер набора символов. Для реальных каналов и сигналов производительность следует понизить, чтобы учесть увеличение полосы пропускания, требуемое для создания реализуемых фильтров. Отметим, что при модуляции MPSK R/W растет с увеличением M. Кроме того, положение рабочих точек MPSK указывает, что для модуляции BPSK ($M = 2$) и квадратичной PSK, или QPSK ($M = 4$), требуются одинаковые значения E_b/N_0. Иными словами, при том же значении E_b/N_0 эффективность использования полосы частот для схемы QPSK равна 2 бит/с/Гц, в отличие от 1 бит/с/Гц для схемы BPSK. Эта уникальная особенность является следствием того, что QPSK представляет собой эффективную комбинацию двух сигналов в модуляции BPSK, которые передаются на ортогональных компонентах несущей.

На рис. 9.6 также изображены рабочие точки некогерентной ортогональной модуляции MFSK при вероятности появления битовой ошибки 10^{-5}. Предполагается, что полоса передачи равна $W_{IF} = M/T$. Следовательно (исходя из уравнения (9.1)), эффективность использования полосы частот равна $R/W = (\log_2 M)/M$. Отметим, что при модуляции MFSK R/W снижается с увеличением M. Также следует отметить, что положение рабочих точек MFSK указывает, что модуляция BFSK ($M = 2$) и квадратичная FSK (M = 4) имеют одинаковую эффективность использования полосы частот, хотя первая требует большего значения E_b/N_0 для той же вероятности появления ошибки. Эффективность использования полосы частот изменяется с коэффициентом модуляции (разнесение частот в герцах, деленное на скорость передачи битов). Предполагается, что для каждого сигнала, модулированного MFSK, требуется одинаковое приращение полосы пропускания, а значит, при $M = 2$ эффективность использования полосы составляет 1 бит/с/2 Гц или 1/2, а при $M = 4$ R/W — 2 бит/с/4 Гц, или 1/2. Таким образом, двоичная и 4-уровневая ортогональная FSK характеризуются одинаковыми значениями R/W.

На рис. 9.6 также показаны рабочие точки для когерентной квадратурной амплитудной модуляции (quadrature amplitude modulation — QAM). Видно, что на фоне остальных модуляций QAM наиболее эффективно использует полосу частот; к этому типу модуляции мы еще обратимся в разделе 9.8.3.

9.5.2. Аналогия между графиками эффективности использования полосы частот и вероятности появления ошибки

График эффективности использования полосы на рис. 9.6 аналогичен графику вероятности ошибки на рис. 9.1. Предел Шеннона (рис. 9.1) является аналогом предельной пропускной способности (рис. 9.6). Кривые на рис. 9.1 называются кривыми равной полосы пропускания. На рис. 9.6 можно аналогично описать кривые равной вероятности для различных схем кодирования и модуляции. Кривые, обозначенные как P_{B1}, P_{B2} и P_{B3}, являются гипотетическими конструкциями для некоторых произвольных схем модуляции и кодирования; кривая P_{B1} представляет собой наибольшую из трех вероятность появления ошибки, а кривая P_{B3} — наименьшую. Также на рисунке указано направление снижения P_B.

Ранее, при изучении графика вероятности появления ошибки, рассматривались возможные компромиссы между P_B, E_b/N_0 и W. Аналогичные компромиссы можно рассмотреть и на графике эффективности использования полосы частот. Возможные компромиссы отображены на рис. 9.6 как сдвиги рабочей точки в направлениях, указанных стрелками. Сдвиг рабочей точки вдоль линии 1 можно рассматривать как поиск компромиссов между P_B и E_b/N_0 при фиксированном значении R/W. Точно так же сдвиг вдоль линии 2 — это поиск компромиссов между P_B и W (или R/W) при фиксированном значении E_b/N_0. И наконец, сдвиг вдоль линии 3 показывает поиск компромиссов между W (или R/W) и E_b/N_0 при постоянном значении P_B. На рис. 9.6 (как и на рис. 9.1) сдвиг вдоль линии 1 может быть вызван повышением или снижением номинального E_b/N_0. Сдвиги вдоль линии 2 или 3 требуют изменений схемы модуляции или кодирования.

Два основных ресурса связи — это переданная мощность и ширина полосы пропускания. Для разных систем связи один из этих ресурсов дороже другого, и следовательно, большую часть систем можно классифицировать как системы ограниченной мощности или ограниченной полосы пропускания. В *системах ограниченной мощности*

для экономии энергии за счет полосы пропускания можно использовать схемы кодирования, эффективно использующие мощность, тогда как в *системах ограниченной полосы* можно применять методы эффективной (с точки зрения используемого спектра) модуляции для экономии полосы частот за счет увеличения расхода энергии.

9.6. Компромиссы при использовании модуляции и кодирования

На рис. 9.7 проводится аналогия между двумя графиками рабочих характеристик, вероятности появления ошибок (рис. 9.1) и эффективности использования полосы частот (рис. 9.6). Рис. 9.7, *а* и *б* изображены в тех же координатах, что рис. 9.1 и 9.6. Вследствие выбора соответствующего масштаба они имеют симметричный вид. В обоих случаях стрелки и обозначения показывают основное следствие сдвига рабочей точки в направлении, указанном стрелкой (собственно сдвиг — это подбор схем кодирования и модуляции). Обозначения, соотнесенные с каждой стрелкой, означают следующее: "Выигрыш (*B*) по *X* за счет (*C*) *Y* при фиксированном (*Ф*) *Z*". Предметом компромиссов являются параметры P_B, *W*, *R/W* и *P* (мощность или *S/N*). Как сдвиг рабочей точки в сторону предела Шеннона (рис. 9.7, *а*) может дать снижение P_B или требуемой мощности передатчика (за счет полосы пропускания), так и сдвиг в сторону предельной пропускной способности канала (рис. 9.7, *б*) может повысить эффективность использования полосы частот за счет повышения требуемой мощности или увеличения P_B.

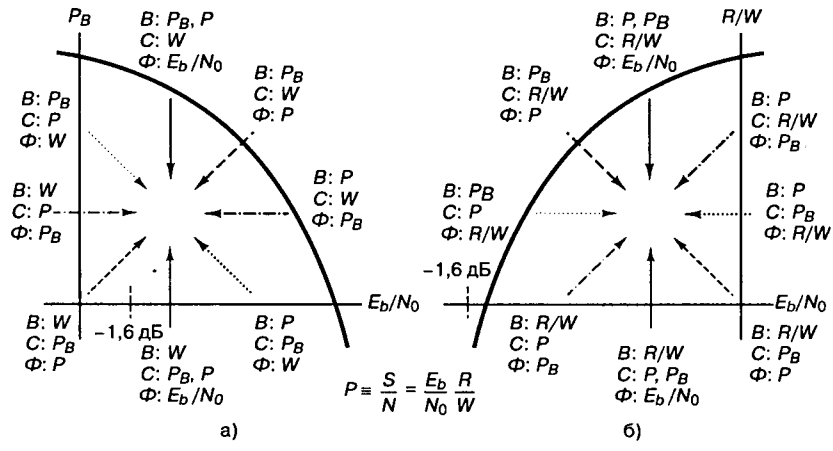


Рис. 9.7. Компромиссы при использовании модуляции и кодирования: а) график вероятности появления ошибки; б) график эффективности использования полосы частот

Довольно часто эти компромиссы изучаются при фиксированном значении P_B (ограничиваемом системными требованиями). Следовательно, наиболее интересующими нас стрелками на рисунке являются описывающие изменения при фиксированной вероятности появления ошибки (обозначены как $Ф: P_B$). На рис. 9.7 имеется четыре такие стрелки: две на графике вероятности ошибки и две на графике эффективности использования полосы частот. Стрелки, помеченные аналогичным образом, указывают соответствие между данными двумя графиками. Работу системы можно представлять с использованием любого из этих графиков. Эти графики — просто два

возможных взгляда на некоторые ключевые параметры системы; каждый из них подчеркивает несколько отличные аспекты разработки. В *системах ограниченной мощности* удобнее всего пользоваться графиком вероятности появления ошибки, поскольку при переходе от одной кривой к другой требования к полосе пропускания лишь подразумеваются, а явно выделяется вероятность появления битовой ошибки. График эффективности использования полосы частот, как правило, применяется в *системах ограниченной полосы пропускания*; здесь при переходе от одной кривой к другой на задний план отодвигается вероятность появления битовой ошибки, тогда как требования к полосе пропускания показываются явно.

Итак, для формирования эвристического взгляда на вопросы разработки компромиссов между вероятностью ошибки, полосой пропускания и мощностью были представлены два графика системных компромиссов, что применимо ко *многим схемам модуляции и кодирования*, но с одной оговоркой. Для *некоторых* кодов или комбинированных схем с модуляцией и кодированием кривые характеристик *не ведут себя настолько предсказуемо*, как в рассмотренном примере. Это связано с функциями коррекции ошибок и использования полосы пропускания конкретного кода. Например, на рис. 6.22 показана характеристика когерентной схемы PSK в сочетании с несколькими кодами. Обратим внимание на графики, описывающие два кода БХЧ, (127, 64) и (127, 36). Из их взаимного расположения видно, что код (127, 64) дает большую *эффективность* кодирования, чем код (127, 36). Это противоречит ожиданиям, поскольку код (127, 36) при тех же размерах блока имеет большую избыточность (и требует большей полосы пропускания), чем код (127, 64). В разделе 9.10, посвященном решетчатому кодированию, рассматриваются коды, которые могут обеспечить высокую эффективность кодирования без расширения полосы пропускания. Рабочие характеристики таких схем кодирования также будут вести себя не так, как характеристики, рассмотренные выше.

9.7. Определение, разработка и оценка систем цифровой связи

Этот раздел призван помочь в описании характерных этапов, которые следует рассматривать при удовлетворении требований, касающихся мощности, полосы пропускания и достоверности передачи в системе цифровой связи. Далее приводится несколько примеров систем, в которых подробно описываются критерии выбора схем кодирования и модуляции, исходя из типа системы — является ли она системой ограниченной мощности или системой ограниченной полосы пропускания. Подчеркиваются тонкие, но важные моменты преобразования битов данных в канальные биты, затем в символы и далее в элементарные сигналы.

Разработка любой системы цифровой связи начинается с описания канала (принимаемая мощность, доступная полоса пропускания, статистики шума и иных ухудшений качества сигнала, таких, например, как замирание) и определения системных требований (скорость передачи данных и вероятность появления ошибок). После описания канала нужно определиться с проектными решениями, которые позволят наилучшим образом использовать канал и удовлетворить требования производительности. Описание производительности системы включает в себя традиционный набор преобразований и расчетов. После того как такой подход станет понятным, его можно использовать как образец для оценки большинства систем связи. В последующих раз-

делах будут рассмотрены три примера систем: система ограниченной мощности без кодирования, система ограниченной полосы без кодирования и система ограниченной мощности и полосы с кодированием. В данном разделе представлены системы связи реального времени, в которых термин *кодированный* (или *некодированный*) означает наличие (или отсутствие) кода коррекции ошибок, включающего использование избыточных битов и увеличение ширины полосы пропускания.

Два основных ресурса связи — это *переданная мощность* и *ширина полосы пропускания*. В различных системах связи один из этих ресурсов дороже другого, и следовательно, большую часть систем можно классифицировать как системы ограниченной мощности или ограниченной полосы пропускания. В *системах ограниченной мощности* для экономии энергии за счет полосы пропускания можно применять схемы кодирования, эффективно использующие мощность, тогда как в *системах ограниченной полосы* можно использовать методы эффективной (с точки зрения используемого спектра) модуляции для экономии полосы частот за счет увеличения расхода энергии. В обоих случаях для экономии энергии или повышения достоверности передачи при расширении полосы пропускания можно применять кодирование с коррекцией ошибок (часто называемое *канальным кодированием*). Для повышения надежности передачи в каналах с ограниченной полосой пропускания *без* увеличения ширины полосы пропускания часто используется решетчатое кодирование (trellis-coded modulation — TCM) [6]. Эти методы рассматриваются в разделе 9.10.

9.7.1. *M*-арная передача сигналов

При использовании схемы, в которой за такт обрабатывается k бит, передача сигналов называется M-арной (см. раздел 3.8). Каждый символ M-арного алфавита можно однозначно связать с последовательностью из k бит, где

$$M = 2^k \quad \text{или} \quad k = \log_2 M \tag{9.13}$$

и M — размер алфавита. Если передача является цифровой, термин *символ* означает элемент M-арного алфавита, передаваемый за время символьного интервала T_s. Для передачи символ следует представить в виде сигнала напряжения или тока. Поскольку сигнал представляет символ, термины *символ* и *сигнал* иногда используются как синонимы. Поскольку один из M символов (или сигналов) передается за интервал T_s, скорость передачи данных R можно записать в следующем виде.

$$R = \frac{k}{T_s} = \frac{\log_2 M}{T_s} \text{ бит/с} \tag{9.14}$$

Из соотношения (9.14) *эффективную* длительность T_b каждого бита можно представить через длительность символа T_s или скорость передачи данных R_s.

$$T_b = \frac{1}{R} = \frac{T_s}{k} = \frac{1}{kR_s} \tag{9.15}$$

Далее на основе выражений (9.13) и (9.15) через скорость передачи битов R можно записать скорость передачи символов R_s.

$$R_s = \frac{R}{\log_2 M} \tag{9.16}$$

Из соотношений (9.14) и (9.15) видно, что в любой цифровой схеме передачи $k = (\log_2 M)$ бит за T_s секунд, при ширине полосы пропускания в W Гц, эффективность использования полосы частот записывается следующим образом.

$$\frac{R}{W} = \frac{\log_2 M}{WT_s} = \frac{1}{WT_b} \text{ бит/с/Гц} \tag{9.17}$$

В данном случае T_b — это эффективное время передачи каждого бита.

9.7.2. Системы ограниченной полосы пропускания

Из уравнения (9.17) видно, что в любой системе цифровой связи эффективность использования полосы частот возрастает при увеличении произведения WT_b. Следовательно, в системах ограниченной полосы пропускания часто применяются сигналы с малыми значениями произведения WT_b. Например, в системе GSM (Global System for Mobile — глобальная система мобильной связи) используется гауссова манипуляция с минимальным сдвигом (Gaussian minimum shift keying — GMSK), в которой произведение WT_b равно 0,3 Гц/бит/с [7], где W — ширина полосы частот по уровню 3 дБ.

При использовании системы ограниченной полосы пропускания без кодирования задачей является получение максимально возможного объема переданной информации в заданной полосе пропускания за счет E_b/N_0 (сохраняя при этом определенное значение P_B). На графике эффективности использования полосы частот (рис. 9.6) показаны рабочие точки когерентной M-арной схемы PSK (MPSK) при $P_B = 10^{-5}$. Предполагается, что немодулированный сигнал подвергается фильтрации по Найквисту (идеальной прямоугольной) [2], так что для модуляции MPSK минимальная двойная полоса пропускания, центрированная на промежуточной частоте (intermediate frequency — IF), связана со скоростью передачи символов.

$$W = \frac{1}{T_s} = R_s \tag{9.18}$$

Здесь T_s — время передачи символа, а R_s — скорость передачи символов. Фильтрация по Найквисту дает *минимальную* полосу пропускания, при которой существует нулевая межсимвольная интерференция; такая идеальная фильтрация определяет *минимальную ширину полосы по Найквисту*. Следует отметить, что при неортогональной передаче сигналов (например, MPSK или MQAM) полоса пропускания зависит не от плотности точек сигналов в группе, а только от скорости передачи сигналов. При передаче вектора сигнала система не различает, пришел ли этот сигнал из разреженного или уплотненного алфавита. Это и является свойством неортогональных сигналов, которое позволяет уплотнить пространство сигналов и, таким образом, повысить эффективность использования полосы частот за счет мощности передатчика. Из уравнений (9.17) и (9.18) запишем, насколько сигнал в модуляции MPSK эффективно использует полосу при фильтрации по Найквисту.

$$\frac{R}{W} = \log_2 M \text{ бит/с/Гц} \tag{9.19}$$

Точки MPSK, показанные на рис. 9.6, подтверждают соотношение (9.19). Отметим, что модуляция MPSK является схемой эффективного использования полосы. С увеличением M также растет R/W. Из рис. 9.6 можно убедиться, что модуляция MPSK

действительно может дать повышение эффективности использования полосы частот за счет увеличения E_b/N_0. Было найдено множество схем модуляции, позволяющих весьма эффективно использовать полосу частот [8], но их рассмотрение выходит за рамки данной книги.

На графике эффективности использования полосы частот (рис. 9.6) показаны две области — область ограниченной полосы пропускания и область ограниченной мощности. Отметим, что желаемые компромиссы, связанные с каждой из этих областей, не являются беспристрастными. В области ограниченной полосы желательным является большое значение R/W; в то же время с ростом E_b/N_0 выравнивается кривая предельной пропускной способности и для повышения R/W требуется дополнительное увеличение E_b/N_0. Аналогичная связь имеется в области ограниченной мощности. Здесь желательно малое отношение E_b/N_0, но кривая предельной пропускной способности становится более крутой и для незначительного снижения требуемого E_b/N_0 нужно уменьшить R/W.

9.7.3. Системы ограниченной мощности

Для систем ограниченной мощности, где имеется достаточная полоса пропускания, но существует дефицит мощности (например, линия космической связи), возможны следующие компромиссы (см. рис. 9.1, *а*): 1) уменьшение P_B за счет полосы пропускания при фиксированном E_b/N_0; 2) снижение E_b/N_0 за счет полосы пропускания при фиксированном P_B. "Естественным" вариантом при выборе модуляции для систем ограниченной мощности представляется M-арная FSK (MFSK). На рис. 9.6 показаны рабочие точки для некогерентной ортогональной модуляции MFSK при $P_B = 10^{-5}$. Для MFSK минимальная полоса частот по Найквисту определяется следующим выражением (см. раздел 4.5.4.1):

$$W = \frac{M}{T_s} = MR_s, \tag{9.20}$$

где T_s — длительность передачи символа, а R_s — скорость передачи символов. При использовании MFSK необходимая полоса пропускания расширяется в M раз по сравнению с двоичной FSK, поскольку теперь существует M различных ортогональных сигналов, каждый из которых требует полосы шириной $1/T_s$. Таким образом, из уравнений (9.17) и (9.20) эффективность использования полосы частот при некогерентной модуляции MFSK с фильтрацией по Найквисту можно выразить следующим образом.

$$\frac{R}{W} = \frac{\log_2 M}{M} \text{ бит/с/Гц} \tag{9.21}$$

Следует отметить важное различие между *эффективностью использования полосы* (R/W) схемой MPSK в уравнении (9.19) и схемой MFSK, представленной в уравнении (9.21). При MPSK R/W растет с увеличением размерности пространства сигналов M. При использовании MFSK работает два механизма. Числитель дроби R/W дает такой же эффект с увеличением M, как и в случае MPSK. Знаменатель же приводит к уменьшению значения R/W при росте M. Поскольку при увеличении M знаменатель растет быстрее числителя, это приводит к снижению R/W. Рабочие точки MFSK, показанные на рис. 9.6, подтверждают соотношение (9.21) — ортогональная передача сигналов (например, MFSK) является схемой с расширением полосы пропускания. Из

рис. 9.6 видно, что модуляция MFSK вполне подходит для снижения требуемого значения E_b/N_0 за счет увеличения полосы пропускания.

Здесь важно подчеркнуть, что в уравнениях (9.18) и (9.19) для MPSK, а также в уравнениях (9.20) и (9.21) для MFSK и всех рабочих точек, показанных на рис. 9.6, предполагается фильтрация по Найквисту (идеальная прямоугольная). На практике такие фильтры нереализуемы. Для *реальных* каналов и сигналов требуемая полоса пропускания должна быть *больше*, чтобы учитывать *реализуемость* фильтров.

Во всех последующих примерах будут рассматриваться радиоканалы с аддитивным белым гауссовым шумом (additive white Gaussian noise — AWGN), не имеющие иных факторов ухудшения качества сигнала. Для простоты выбор типа модуляции будет ограничен *схемами с постоянной огибающей* — MPSK или некогерентная ортогональная MFSK. Таким образом, если в системах без кодирования ограничена пропускная способность канала, выбирается схема MPSK, а если у канала ограничена мощность, применяется MFSK. Отметим, что *при рассмотрении кодирования с коррекцией ошибок* выбор типа модуляции не так прост, поскольку существуют методы кодирования [9], которые позволяют более эффективно выбрать компромисс между полосой пропускания и мощностью, чем схемы M-арной модуляции.

Следует сказать, что в общем случае M-арную передачу сигналов можно рассматривать как процедуру *кодирования формы сигнала*. Иными словами, если вместо двоичной выбрана M-арная модуляция, *по сути*, сигналы двоичной формы заменяются сигналами *лучшей* формы — лучшей или с точки зрения эффективности использования полосы (MPSK), или с точки зрения требуемой мощности (MFSK). Хотя передачу ортогональных сигналов MFSK можно рассматривать как систему с кодированием (ее можно представить как код Рида-Мюллера [10]), мы будем применять термин *система с кодированием* только к традиционным кодам коррекции ошибок, использующим избыточность, таким как блочные или сверточные коды.

9.7.4. Требования к передаче сигналов MPSK и MFSK

Основное соотношение между скоростью передачи символов (или сигналов) R_s и скоростью передачи битов R выражено в уравнении (9.16) и имеет следующий вид.

$$R_s = \frac{R}{\log_2 M}$$

На основе этого соотношения и уравнений (9.18)–(9.21) для скорости передачи данных $R = 9600$ бит/с была составлена табл. 9.1 [11]. В этой таблице сведены данные о скорости передачи символов, минимальной полосе пропускания по Найквисту, эффективности использования полосы частот для MPSK и некогерентной ортогональной MFSK при $M = 2$, 4, 8, 16 и 32. В табл. 9.1 также для каждого M показаны значения E_b/N_0, необходимые для получения вероятности ошибки 10^{-5} для MPSK и MFSK. Эти значения E_b/N_0 в таблице были получены исходя из соотношений, которые будут представлены далее, и соответствуют компромиссам, показанным на рис. 9.6. С ростом M передача сигналов MPSK позволяет более эффективно использовать полосу частот за счет увеличения E_b/N_0, в то время как передача сигналов MFSK позволяет снизить E_b/N_0 за счет расширения полосы пропускания. В следующих трех разделах будут подробно рассмотрены примеры из табл. 9.1.

Таблица 9.1. Скорость передачи символов, минимальная полоса по Найквисту, эффективность использования полосы и требуемое E_b/N_0 для схем MPSK и некогерентной ортогональной MFSK при скорости передачи данных 9600 бит/с

M	k	R (бит/с)	R_s (символ/с)	MPSK Минимальная полоса (Гц)	MPSK R/W	MPSK E_b/N_0 (дБ) $P_B = 10^{-5}$	Некогерентная ортогональная MFSK Минимальная полоса (Гц)	MFSK R/W	MFSK E_b/N_0 (дБ) $P_B = 10^{-5}$
2	1	9600	9600	9600	1	9,6	19200	1/2	13,4
4	2	9600	4800	4800	2	9,6	19200	1/2	10,6
8	3	9600	3200	3200	3	13,0	25600	1/3	9,1
16	4	9600	2400	2400	4	17,5	38400	1/4	8,1
32	5	9600	1920	1920	5	22,4	61440	5/32	7,4

9.7.5. Система ограниченной полосы пропускания без кодирования

Рассмотрим радиоканал с шумом AWGN и ограниченной полосой пропускания $W = 4000$ Гц. Пусть ограничения линии связи (мощность передатчика, коэффициент усиления антенны, потери в канале и т. д.) приводят к тому, что отношение мощности принятого сигнала к спектральной плотности мощности шума (P_r/N_0) равно 53 дБГц. Допустим, требуемое значение скорости передачи информации R равно 9600 бит/с, а требуемая вероятность появления битовой ошибки P_B *не должна превышать* 10^{-5}. Задача — выбрать схему модуляции, которая сможет удовлетворить требуемым рабочим характеристикам. В общем случае может потребоваться схема кодирования с коррекцией ошибок, если ни одна из доступных схем модуляции не может удовлетворить всем требованиям. Тем не менее в данном примере (как показывается далее) кодирование с коррекцией ошибок не понадобится.

Для любой цифровой системы связи соотношение между принимаемой мощностью и спектральной плотностью мощности шума (P_r/N_0), а также принимаемой энергией одного бита и спектральной плотностью мощности шума (E_b/N_0) приведено в формуле (5.20,в) и имеет следующий вид.

$$\frac{P_r}{N_0} = \frac{E_b}{N_0} R \tag{9.22}$$

Выразив из этого соотношения E_b/N_0 в децибелах, получаем следующее.

$$\frac{E_b}{N_0}(\text{дБ}) = \frac{P_r}{N_0}(\text{дБГц}) - R(\text{дБбит/с}) = \tag{9.23}$$

$$= 53 \text{ дБГц} - (10 \times \lg 9600) \text{ дБбит/с} = 13{,}2 \text{ дБ (или } 20{,}89)$$

Поскольку необходимая скорость передачи данных 9600 бит/с значительно больше, чем можно достичь с доступной полосой пропускания, составляющей 4000 Гц, канал можно считать *каналом ограниченной полосы пропускания*. Следовательно, в качестве схемы модуляции выбираем MPSK. Напомним, что при выборе возможной схемы модуляции было решено ограничиться модуляциями с постоянной огибающей; без такого ограничения можно найти тип модуляции с еще большей эффективностью использования полосы частот. Вычислим далее *минимально допустимое* значение M, при котором символьная скорость передачи данных *не превышает* доступной полосы пропускания 4000 Гц. Из табл. 9.1 видно, что наименьшим значением M, удовлетворяющим этим требованиям, является

$M = 8$. Следующая задача — выяснить, удовлетворяется ли требование к вероятности появления битовой ошибки $P_B \leq 10^{-5}$ при использовании 8-уровневой PSK или потребуется дополнительно вводить схему кодирования с коррекцией ошибок. Из табл. 9.1 видно, что 8-уровневая PSK удовлетворяет всем требованиям, поскольку отношение E_b/N_0 для 8-уровневой PSK меньше принятого E_b/N_0, выраженного в (9.23). Тем не менее, представим, что табл. 9.1 нет. Покажем, как определить, нужно ли кодирование с коррекцией ошибок.

На рис. 9.8 показана блочная диаграмма простого модулятора/демодулятора (модема), в которой отображены функциональные элементы разработки. В модуляторе в ходе преобразования битов данных в символы выходная скорость передачи символов равна R_s, т.е. в $(\log_2 M)$ раз меньше входной скорости передачи битов R, как видно из уравнения (9.16). Аналогично на входе демодулятора отношение энергии символа к спектральной плотности мощности шума E_s/N_0 в $(\log_2 M)$ больше E_b/N_0, поскольку каждый символ состоит из $(\log_2 M)$ бит. Поскольку E_s/N_0 больше E_b/N_0 в столько же раз, во сколько R_s меньше R, формулу (9.22) можно переписать следующим образом.

$$\frac{P_r}{N_0} = \frac{E_b}{N_0} R = \frac{E_s}{N_0} R_s \tag{9.24}$$

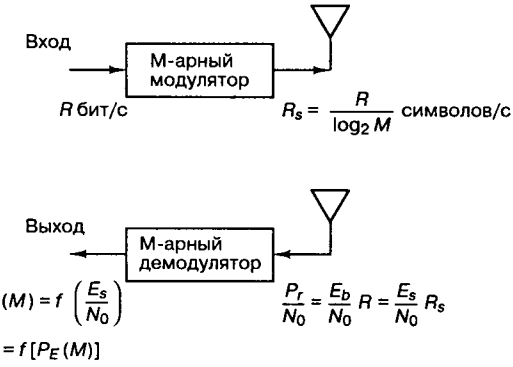


Рис. 9.8. Схема простого модулятора/демодулятора (модема) без канального кодирования

За каждый интервал T_s демодулятор принимает сигнал (в данном случае — один из $M = 8$ возможных сдвигов фаз). Вероятность $P_E(M)$ возникновения в демодуляторе символьной ошибки довольно точно описывается следующим приближенным выражением [12].

$$P_E \approx 2Q\left[\sqrt{\frac{2E_s}{N_0}} \sin\left(\frac{\pi}{M}\right)\right] \text{ для } M > 2 \tag{9.25}$$

Здесь $Q(x)$ — это *гауссов интеграл ошибок*, который был определен в выражении (3.43).

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{u^2}{2}\right) du$$

На рис. 9.8 и на всех последующих рисунках для обозначения некоторой функциональной зависимости вероятности от x будет использоваться не явное выражение, а обобщенная запись $f(x)$.

Как правило, для описания эффективности связи (по фактору мощности) или достоверности передачи в цифровых системах их выражают через E_b/N_0 в децибелах. Такое употребление E_b/N_0 является распространенной практикой. Тем не менее напомним, что на входе демодулятора/детектора нет битов, имеются только сигналы, которым присвоено битовое значение. Следовательно, принимаемое значение E_b/N_0 представляет собой пропорциональное распределение энергии принимаемых битов по сигналам. Более точное (но громоздкое) название — энергия *эффективного бита* на N_0. Для выражения $P_E(M)$ из уравнения (9.25) сначала нужно получить выражение для отношения энергии символа к спектральной плотности мощности шума, E_s/N_0. Поскольку (из выражения (9.23)) $E_b/N_0 = 13{,}2$ дБ (или 20,89) и каждый символ образуется ($\log_2 M$) битами, при $M = 8$ получаем следующее.

$$\frac{E_s}{N_0} = (\log_2 M)\frac{E_b}{N_0} = 3 \times 20{,}89 = 62{,}67 \tag{9.26}$$

Подставляя выражение (9.26) в (9.25), получаем вероятность появления символьной ошибки $P_E = 2{,}2 \times 10^{-5}$. Чтобы этот результат перевести в вероятность появления битовой ошибки, нужно воспользоваться соотношением между вероятностью появления битовой ошибки P_B и вероятностью появления символьной ошибки P_E для многофазной передачи сигналов [10]. Итак,

$$P_B \approx \frac{P_E}{\log_2 M} \quad (\text{для } P_E << 1) \tag{9.27}$$

Это является довольно хорошей аппроксимацией, если для отображения битов в символы применяется код Грея [12]. Последняя формула дает $P_B = 7{,}3 \times 10^{-6}$, что вполне удовлетворяет требованиям к вероятности появления битовых ошибок. Таким образом, в приведенном примере кодирование с коррекцией ошибок не потребовалось и 8-уровневая PSK удовлетворяет требованиям канала ограниченной полосы пропускания (что и было предсказано при изучении значений E_b/N_0 в табл. 9.1).

9.7.6. Система ограниченной мощности без кодирования

Рассмотрим теперь систему, где требуется такая же скорость передачи данных и такая же вероятность появления битовой ошибки, как и в случае, описанном в разделе 9.7.5. Однако в данном примере доступная полоса пропускания W пусть будет равна 45 кГц, а доступное P_r/N_0 — 48 дБГц. Как и ранее, задача — выбор схемы модуляции или модуляции/кодирования, которая смогла бы удовлетворить техническим требованиям. В данном случае кодирования с коррекцией ошибок снова не потребуется.

Очевидно, что в этом примере канал не имеет ограничений на полосу пропускания, так как имеющихся 45 кГц полосы более чем достаточно для обеспечения требуемой скорости передачи данных 9600 бит/с. Из уравнения (9.23) получаем принимаемое E_b/N_0.

$$\frac{E_b}{N_0}(\text{дБ}) = 48\ \text{дБГц} - (10 \times \lg 9600)\ \text{дББит/с} = 8{,}2\ \text{дБ}\ \ (\text{или } 6{,}61) \tag{9.28}$$

Поскольку полоса пропускания избыточна, а для получения нужной вероятности битовой ошибки доступно сравнительно небольшое E_b/N_0, канал можно назвать *каналом ограниченной мощности*. Следовательно, в качестве схемы модуляции выбирается MFSK. Для экономии мощности далее необходимо подобрать *максимальное* M, при котором минимальная полоса пропускания MFSK не будет превышать доступные 45 кГц. Следуя табл. 9.1, можно видеть, что это возможно при $M = 16$. Следующая задача — выяснить, можно ли удовле-

творить требованию $P_B \leq 10^{-5}$ с помощью лишь 16-уровневой FSK, без привлечения какого-либо кодирования с коррекцией ошибок. Подобно рассмотренному ранее случаю, из табл. 9.1 видно, что 16-уровневая FSK *может* удовлетворить требованиям, поскольку требуемое E_b/N_0, взятое для 16-уровневой FSK, меньше полученного из уравнения (9.28). Тем не менее мы получим данный результат, не обращаясь к табл. 9.1. Покажем, как определить, нужно ли кодирование с коррекцией ошибок.

Как и ранее, блочная диаграмма на рис. 9.8 отображает соотношение между скоростью передачи символов R_s и скоростью передачи битов R и между E_s/N_0 и E_b/N_0; эти соотношения аналогичны полученным в предыдущем примере системы ограниченной полосы. В данном случае демодулятор 16-уровневой схемы FSK принимает сигнал (одну из 16 возможных частот) за интервал T_s. При некогерентной MFSK вероятность возникновения в демодуляторе символьной ошибки аппроксимируется следующим выражением [13].

$$P_E(M) \leq \frac{M-1}{2}\exp\left(-\frac{E_s}{2N_0}\right) \tag{9.29}$$

Для вычисления $P_E(M)$ из формулы (9.29) требуется, как и в предыдущем примере, найти E_s/N_0. Подставляя выражение (9.28) в (9.26) при $M = 16$, получаем следующее.

$$\frac{E_s}{N_0} = (\log_2 M)\frac{E_b}{N_0} = 4 \times 6{,}61 = 26{,}44 \tag{9.30}$$

Далее формулу (9.30) подставляем в (9.29), что дает вероятность появления символьной ошибки $P_E = 1{,}4 \times 10^{-5}$. Для преобразования этой величины в вероятность появления битовой ошибки P_B нужно воспользоваться соотношением между P_B и P_E для передачи ортогональных сигналов [13], которое имеет следующий вид.

$$P_B = \frac{2^{k-1}}{2^k - 1}P_E \tag{9.31}$$

Из последней формулы получаем, что $P_B = 7{,}3 \times 10^{-6}$; это вполне удовлетворяет требуемой вероятности появления битовых ошибок. Таким образом, с помощью 16-уровневой FSK можно удовлетворить требованиям спецификации данного канала ограниченной мощности, не используя дополнительно никакого кодирования с коррекцией ошибок (что и было предсказано при изучении значений E_b/N_0 в табл. 9.1).

9.7.7. Система ограниченной мощности и полосы пропускания с кодированием

В этом примере начальные параметры будут такими же, как и в предыдущем примере системы ограниченной полосы пропускания (раздел 9.7.5), а именно $W = 4000$ Гц, $P/N_0 = 53$ дБГц и $R = 9600$ бит/с, за одним исключением. В данном случае предполагается, что вероятность появления битовой ошибки должна быть *не больше* 10^{-9}. Поскольку полоса пропускания составляет 4000 Гц, а из уравнения (9.23) находим $E_b/N_0 = 13{,}2$ дБ, то из табл. 9.1 ясно, что данная система ограничена и по полосе пропускания и по доступной мощности (для удовлетворения требованиям к полосе пропускания можно использовать 8-уровневую схему PSK; но имеющихся 13,2 дБ отношения E_b/N_0 совсем не достаточно для обеспечения требуемой вероятности появления битовой ошибки 10^{-9}). При таких малых значениях P_B, системы, изображенной на рис. 9.8, явно недостаточно, значит, надо по-

смотреть, какое повышение производительности сможет дать кодирование с коррекцией ошибок (в пределах доступной полосы пропускания). В общем случае можно использовать сверточный или блочный код. Для упрощения будем применять блочный код. Коды Боуза-Чоудхури-Хоквенгема (Bose, Chaudhuri, Hocquenghem — BCH, БХЧ) образуют большой класс мощных циклических (блочных) кодов коррекции ошибок [14]. В данном примере выберем из семейства кодов один конкретный. Рассмотрим табл. 9.2, где приведены некоторые коды БХЧ, определяемые параметрами n, k и t. Здесь k — количество информационных битов, которые код преобразует в более длинные блоки из n кодовых битов (их также называют *канальными битами* или *канальными символами*), а t — максимальное число неправильных канальных битов, не поддающихся исправлению, в блоке размером n бит. *Степень кодирования* кода определяется как отношение k/n; а величина, обратная данной, является мерой избыточности кода.

Таблица 9.2. Коды БХЧ (неполный перечень)

n	k	t
7	4	1
15	11	1
	7	2
	5	3
31	26	1
	21	2
	16	3
	11	4
		5
63	57	1
	51	2
	45	3
	39	4
	36	5
	30	6
127	120	1
	113	2
	106	3
	99	4
	92	5
	85	6
	78	7
	71	9
	64	10
	57	11
	50	13
	43	14
	36	15
	29	21
	22	23
	15	27
	8	31

Поскольку ограничения системы аналогичны использованным в разделе 9.7.5, удовлетворить требования к полосе пропускания можно с помощью 8-уровневой схемы PSK. Тем не менее для снижения вероятности появления ошибки до $P_B \le 10^{-9}$ придется воспользоваться кодом коррекции ошибок. При выборе оптимального кода из табл. 9.2 нужно иметь в виду следующее.

1. Выходная вероятность появления битовой ошибки в комбинированной системе модуляции/кодирования должна удовлетворять системным требованиям достоверности передачи.
2. Степень кодирования кода не должна требовать увеличения полосы пропускания до значения, большего доступного.
3. Код должен быть максимально простым. Вообще, чем короче код, тем проще его реализовать.

Минимальная полоса пропускания для 8-уровневой схемы PSK без кодирования составляет 3200 Гц (см. табл. 9.1), а доступная полоса пропускания канала — 4000 Гц. Следовательно, полосу пропускания некодированного сигнала можно увеличить *не более чем* в 1,25 раза (или расширить на 25%). Таким образом, самым первым шагом в данном (упрощенном) примере выбора кода будет отбрасывание тех кодов из табл. 9.2, которые потребуют расширения полосы пропускания более чем на 25%. В результате мы получим набор кодов, "совместимых" с полосой пропускания (табл. 9.3). В этой таблице добавлены два столбца, которые обозначены как "эффективность кодирования", G, причем эта величина определяется следующим образом.

$$G(\text{дБ}) = \left(\frac{E_b}{N_0}\right)_{\text{некодированное}} (\text{дБ}) - \left(\frac{E_b}{N_0}\right)_{\text{кодированное}} (\text{дБ}) \qquad (9.32)$$

Таблица 9.3. Коды БХЧ, "совместимые" с полосой пропускания

			Эффективность кодирования, G (дБ)	
n	k	t	$P_B = 10^{-5}$	$P_B = 10^{-9}$
31	26	1	1,8	2,0
63	57	1	1,8	2,2
	51	2	2,6	3,2
127	120	1	1,7	2,2
	113	2	2,6	3,4
	106	3	3,1	4,0

Из уравнения (9.32) эффективность кодирования можно описать как меру *снижения* величины требуемого E_b/N_0 (в децибелах), которую нужно обеспечить с помощью свойств кода, касающихся обнаружения и исправления ошибок. Эффективность кодирования зависит от типа модуляции и вероятности возникновения битовых ошибок. В табл. 9.3 эффективность кодирования G рассчитана для значений $P_B = 10^{-5}$ и $P_B = 10^{-9}$. При модуляции MPSK, G относительно независима от значения M. Следовательно, при конкретной вероятности возникновения битовой ошибки данный код будет иметь приблизительно равную эффективность с любой модуляцией MPSK. Эффективность кодирования в табл. 9.3 рассчитана согласно процедуре, описываемой в разделе 9.7.7.1.

На рис. 9.9 изображена блочная диаграмма, включающая кодер и модулятор/демодулятор (модем). Если сравнить рис. 9.9 и 9.8, то видно, что введение блоков кодера/декодера влечет за собой дополнительные преобразования. На рис. 9.9 в блоке кодер/модулятор показано, как преобразовывается скорость передачи: из R (бит/с) в R_c (канальных бит/с), а затем в R_s (символ/с).

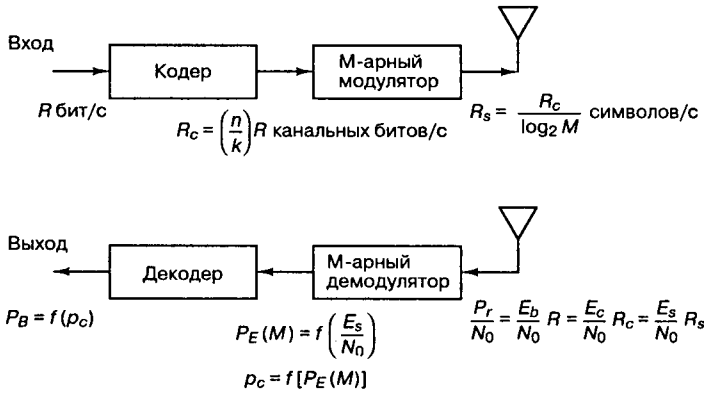


Рис. 9.9. Схема модема с канальным кодированием

Предполагается, что рассматриваемая система связи является системой реального времени, а значит, в ней недопустимы задержки при передаче сообщений. Следовательно, скорость передачи канальных битов R_c должна *превышать* битовую скорость передачи данных R в n/k раз. Более того, каждый передаваемый символ образован ($\log_2 M$) канальными битами, так что символьная скорость передачи R_s *меньше* R_c в ($\log_2 M$) раз. Для систем с модуляцией и кодированием преобразования скорости имеют следующий вид.

$$R_c = \left(\frac{n}{k}\right) R \tag{9.33}$$

$$R_s = \frac{R_c}{\log_2 M} \tag{9.34}$$

В блоке демодулятор/декодер, показанном на рис. 9.9, преобразования энергии битов данных, энергии канальных битов и энергии символов связаны теми же множителями, что и преобразования скоростей, показанные в выражениях (9.33) и (9.34). Поскольку при преобразовании кодирования k информационных битов заменяются n канальными битами, отношение энергии канального бита к спектральной плотности мощности шума, E_c/N_0, — это результат умножения E_b/N_0 на коэффициент k/n. Кроме того, поскольку каждый передаваемый символ состоит из ($\log_2 M$) канальных битов, E_s/N_0, необходимое в (9.25) для получения P_E, вычисляется путем умножения E_c/N_0 на коэффициент ($\log_2 M$). Для систем, содержащих одновременно и модуляцию, и кодирование, преобразования отношений энергии к спектральной плотности мощности шума будут следующими.

$$\frac{E_c}{N_0} = \left(\frac{n}{k}\right) \frac{E_b}{N_0} \tag{9.35}$$

$$\frac{E_s}{N_0} = (\log_2 M)\frac{E_c}{N_0} \tag{9.36}$$

Следовательно, исходя из уравнений (9.33)–(9.36), можно обобщить выражение для P_r/N_0 в уравнении (9.24).

$$\frac{P_r}{N_0} = \frac{E_b}{N_0}R = \frac{E_c}{N_0}R_c = \frac{E_s}{N_0}R_s \tag{9.37}$$

Как и ранее, канал связи описывается величиной E_b/N_0, выражаемой в децибелах. Тем не менее на входе демодулятора/детектора *нет ни информационных, ни канальных битов*. Есть *только* сигналы (символы передачи), которым присваивается битовое значение, а следовательно, их можно описывать через пропорциональное распределение энергии по битам. Из формулы (9.37) видно, что додетекторная точка приемника — это удобная опорная точка, в которой можно соотнести *эффективную* энергию и *эффективную* скорость различных параметров. Слово "эффективный" используется потому, что единственные сигналы в додетекторной точке — это импульсы, которые мы называем символами. Конечно, эти символы связаны с канальными битами, которые, в свою очередь, связаны с информационными битами. Чтобы подчеркнуть тот момент, что уравнение (9.37) весьма удобно при учете системных ресурсов, рассмотрим систему, в которой поток некоторого числа битов, например 273 бит, настолько часто появляется в виде отдельного блока, что этой группе присваивается собственное имя; все это идет отдельной "порцией". Инженеры делают это постоянно, например восемь бит называют байтом. Как только мы определили новый объект, его сразу можно связать с параметрами уравнения (9.37), поскольку P_r/N_0 — это теперь энергия блока на N_0, умноженная на скорость передачи блока. Нечто такое будет использовано в главе 12, где подобное расширение формулы (9.37) будет применяться к элементарным сигналам расширенного спектра.

Поскольку значения P_r/N_0 и R равны 53 дБГц и 9600 бит/с, (по аналогии с предыдущим случаем) из уравнения (9.23) находим, что принятое $E_b/N_0 = 13{,}2$ дБ. Отметим, что принимаемое E_b/N_0 фиксированно и не зависит от параметров кода n и k, а также от параметра модуляции M. Как было установлено при изучении табл. 9.3, для идеального кода, удовлетворяющего всем требованиям, можно итеративно повторить расчеты, представленные на рис. 9.9. Полезно запрограммировать на ПК (или калькуляторе) следующие четыре шага как функцию от n, k и t. Шаг первый начинается с подстановки уравнения (9.35) в (9.36).

Шаг 1
$$\frac{E_s}{N_0} = (\log_2 M)\frac{E_c}{N_0} = (\log_2 M)\left(\frac{k}{n}\right)\frac{E_b}{N_0} \tag{9.38}$$

Шаг 2
$$P_E(M) \approx 2Q\left[\sqrt{\frac{2E_s}{N_0}}\sin\left(\frac{\pi}{M}\right)\right] \tag{9.39}$$

Выражение (9.39) — это аппроксимация (для M-арной PSK) вероятности символьной ошибки P_E, которая уже приводилась в формуле (9.25). В каждый промежуток передачи символа демодулятор принимает решение относительно значения символа и подает на декодер последовательность канальных битов, представляющую этот символ. Если на демодуляторе канальные биты квантуются на два уровня, обозначаемых 1 и 0, го-

ворят, что демодулятор принимает *жесткое решение* (hard decision). Если выход демодулятора квантуется более чем на два уровня — демодулятор принимает *мягкое решение* (soft decision). В этом разделе предполагается принятие жестких решений.

Теперь, когда в системе присутствует блок декодера, вероятность появления ошибки в канальном бите вне демодулятора и на декодере будем обозначать как p_c, а вероятность появления ошибки в бите *вне декодера*, как и ранее, будем обозначать через P_B (декодированная вероятность битовой ошибки). С помощью p_c уравнение (9.27) можно переписать следующим образом.

Шаг 3
$$p_c \approx \frac{P_E}{\log_2 M} \quad (\text{для } P_E << 1) \tag{9.40}$$

Третий шаг связывает вероятность появления ошибки в канальном бите с вероятностью появления ошибки в символе вне демодулятора (предполагается использование кода Грея, как это было в уравнении (9.27)).

В системах связи реального времени, использующих традиционные схемы кодирования, при фиксированном значении P_r/N_0 величина E_s/N_0 с кодированием *всегда будет меньше* величины E_s/N_0 без кодирования. Поскольку при кодировании демодулятор принимает сигнал с меньшим E_s/N_0, он делает больше ошибок! Тем не менее при использовании кодирования достоверность передачи зависит от характеристик не только демодулятора, но и декодера. Следовательно, для повышения достоверности передачи при кодировании декодер должен осуществлять коррекцию ошибок так, чтобы *перекрывать* слабую производительность демодулятора. Итоговая декодированная вероятность битовой ошибки P_B на выходе зависит от конкретного кода, декодера и вероятности появления ошибки в канальном бите p_c. Эту зависимость можно аппроксимировать следующим выражением [15].

Шаг 4
$$P_B \approx \frac{1}{n} \sum_{j=t+1}^{n} j \binom{n}{j} p_c^j (1-p_c)^{n-j} \tag{9.41}$$

На четвертом шаге t — это наибольшее число канальных битов, которые код способен исправить в блоке размером n бит. Исходя из уравнений (9.38)–(9.41), определяющих четыре упомянутых выше шага, декодированную вероятность появления битовой ошибки P_B можно рассчитать как функцию n, k и t для всех кодов, представленных в табл. 9.3. Нужная позиция таблицы, удовлетворяющая установленным требованиям к вероятности возникновения ошибки с *наибольшей возможной* степенью кодирования и *наименьшим* n, — это код с коррекцией двойных ошибок (63, 51). Ниже приводятся соответствующие расчеты.

Шаг 1
$$\frac{E_s}{N_0} = 3\left(\frac{51}{63}\right)20{,}89 = 50{,}73,$$

где $M = 8$, а принятое $E_b/N_0 = 13{,}2$ дБ (или 20,89).

Шаг 2
$$P_E \approx 2Q\left[\sqrt{101{,}5} \times \sin\left(\frac{\pi}{8}\right)\right] = 2Q(3{,}86) = 1{,}2 \times 10^{-4}$$

Шаг 3
$$p_c \approx \frac{1{,}2 \times 10^{-4}}{3} = 4 \times 10^{-5}$$

$$P_B \approx \frac{3}{63}\binom{63}{3}(4\times10^{-5})^3(1-4\times10^{-5})^{60} +$$

Шаг 4

$$+\frac{4}{63}\binom{63}{4}(4\times10^{-5})^4(1-4\times10^{-5})^{59} + \ldots =$$

$$= 1{,}2\times10^{-10}$$

На четвертом шаге способность кода к исправлению битовых ошибок $t = 2$. Для получения P_B на четвертом шаге, учитываются только первые два члена суммы в уравнении (9.41), так как остальные слагаемые дают пренебрежимо малый вклад при малых значениях p_c или при разумно большом E_b/N_0. Важно отметить, что при выполнении этих расчетов на компьютере стоит (на всякий случай) *всегда* учитывать все слагаемые в формуле (9.41), так как приближенное решение может сильно отличаться от правильного при малых значениях E_b/N_0. Теперь, когда мы выбрали код (63, 51), рассчитаем скорость передачи данных в канальных битах R_c и скорость передачи символов R_s с помощью уравнений (9.33) и (9.34), при $M = 8$.

$$R_c = \left(\frac{n}{k}\right)R = \left(\frac{63}{51}\right)9600 \approx 11{,}859 \quad \text{канальных битов/с}$$

$$R_s = \frac{R_c}{\log_2 M} = \frac{11{,}859}{3} = 3953 \quad \text{символов/с}$$

9.7.7.1. Расчет эффективности кодирования

Более прямой способ поиска простейшего кода, удовлетворяющего требованиям, указанным в разделе 9.7.7, состоит в следующем. Вначале для схемы 8-PSK *без кодирования* рассчитывается, насколько большее (относительно доступных 13,2 дБ) значение E_b/N_0 требуется для получения $P_B = 10^{-9}$. Это дополнительное E_b/N_0 является требуемой эффективностью кодирования. Используя формулы (9.27) и (9.39), находим E_s/N_0 *без использования кодирования*, которое даст вероятность появления ошибки $P_B = 10^{-9}$.

$$P_B \approx \frac{P_E}{\log_2 M} \approx \frac{2Q\left[\sqrt{\frac{2E_s}{N_0}}\sin\left(\frac{\pi}{M}\right)\right]}{\log_2 M} = 10^{-9} \tag{9.42}$$

При таком низком значении вероятности битовой ошибки, правомерно использовать приведенную в (3.44) аппроксимацию $Q(x)$. Методом проб и ошибок (с помощью программируемого калькулятора) находим, что E_s/N_0 *без кодирования* равно 120,67 (20,8 дБ), и поскольку каждый символ состоит из $(\log_2 8) = 3$ бит, требуемое (E_b/N_0) (без кодирования) = 120,67/3 = 40,22 = 16 дБ. Из параметров примера и уравнения (9.23) мы знаем, что (E_b/N_0) (с кодированием) = 13,2 дБ. Следовательно, используя формулу (9.32), видим, что эффективность кодирования, удовлетворяющая условию $P_B = 10^{-9}$, равна следующему.

$$G(\text{дБ}) = \left(\frac{E_b}{N_0}\right)_{\text{без кодирования}}(\text{дБ}) - \left(\frac{E_b}{N_0}\right)_{\text{с кодированием}}(\text{дБ}) = 16\ \text{дБ} - 13{,}2\ \text{дБ} = 2{,}8\ \text{дБ}$$

Чтобы приведенный выше расчет был точным, все значения E_b/N_0 должны точно соответствовать одинаковым значениям вероятности битовой ошибки. В нашей ситуации это не совсем так: два значения E_b/N_0 соответствуют $P_B = 10^{-9}$ и $P_B = 1{,}2 \times 10^{-10}$. Тем не менее при таких низких значениях вероятности (даже при таком отличии) расчеты дают хорошее приближенное значение требуемой эффективности кодирования. Изучая табл. 9.3 на предмет выбора простейшего кода, дающего эффективность кодирования *не меньше* 2,8 дБ, видим, что это код (63, 51); тот же, что и был выбран ранее. Отметим, что эффективность кодирования нужно всегда определять для конкретной вероятности появления ошибки и типа модуляции, как в табл. 9.3.

9.7.7.2. Выбор кода

Рассмотрим систему связи реального времени, которая, согласно спецификации, относится к системам ограниченной мощности, но в то же время обладает достаточной полосой пропускания и должна иметь очень низкую вероятность возникновения ошибки. В данной ситуации необходимо кодирование с коррекцией ошибок. Пусть для кодирования нужно выбрать один из кодов БХЧ, которые представлены в табл. 9.2. Поскольку система имеет достаточную полосу пропускания, а требования относительно вероятности ошибок довольно строги, может возникнуть соблазн выбора самого мощного кода, из указанных в табл. 9.2, а именно — кода (127, 8), способного исправлять комбинации до 31 искаженных бит в блоке размером 127 кодовых бит. Будет ли кто-либо использовать такой код в системе связи реального времени? Конечно же, нет. Объясним, почему такой выбор *неразумен.*

Если в системе связи реального времени применяется код коррекции ошибок, то на достоверность передачи оказывают влияние два фактора. Один вызывает улучшение достоверности передачи, а другой — снижение. Первый фактор — это кодирование; чем больше избыточность кода, тем выше способность кода к коррекции ошибок. Второй фактор — это уменьшение энергии, приходящейся на канальный символ или кодовый бит (по сравнению с информационным битом). Такое уменьшение энергии вызвано повышением избыточности (что влечет за собой увеличение скорости передачи в системе связи реального времени). Меньшая энергия символа — это большее число ошибок. В конце концов, второй фактор подавляет первый, и при очень низких степенях кодирования резко возрастает вероятность появления ошибки. (Эти рассуждения иллюстрируются ниже, в примере 9.4.) Следует отметить, что сказанное справедливо только для систем связи реального времени, где задержки передачи сообщения недопустимы. В иных системах можно "играть на компромиссах" между задержкой при передаче сообщения и избыточностью кода (не снижая энергию, приходящуюся на символ).

Пример 9.4. Выбор кода, удовлетворяющего требованиям спецификации

Даны следующие параметры системы: $P_r/N_0 = 67$ дБГц, скорость передачи данных $R = 10^6$ бит/с, доступная полоса пропускания $W = 20$ МГц, декодированная вероятность битовой ошибки $P_B \leq 10^{-7}$, модуляция BPSK. Выберите из табл. 9.2 код, удовлетворяющий этим требованиям. Рассмотрение начать с кода (127, 8). Привлекательность этого кода объясняется наивысшей (из представленных кодов) способностью к коррекции ошибок.

Решение

Код (127, 8) расширяет полосу пропускания в $127/8 = 15{,}875$ раз. Следовательно, при использовании этого кода скорость передачи 1 Мбит/с (определяющая номинальную полосу

пропускания в 1 МГц) возрастает до 15,875 МГц. Таким образом, передаваемый сигнал находится в пределах полосы 20 МГц, что позволяет увеличение полосы еще на 25% для целей фильтрации. После выбора кода оценим вероятность ошибки, использовав шаги, описанные в разделе 9.7.7.

$$\frac{E_b}{N_0} = \frac{P_r}{N_0}\left(\frac{1}{R}\right) = 67 \text{ дБ} - 60 \text{ дБ} = 7 \text{ дБ (или 5)}$$

$$\frac{E_s}{N_0} = \frac{E_c}{N_0} = \left(\frac{k}{n}\right)\frac{E_b}{N_0} = \left(\frac{8}{127}\right)5 = 0{,}314$$

Поскольку применяется двоичная модуляция, $p_c = P_E$, так что имеем следующее.

$$p_c = P_E \approx Q\left(\sqrt{\frac{2E_s}{N_0}}\right) = Q(\sqrt{0{,}628}) = Q(0{,}7936) = 0{,}2156$$

Код (127, 8) способен исправлять последовательности до $t = 31$ ошибочных бит, поэтому, используя формулу (9.41), получаем следующую вероятность появления ошибки в декодированном бите.

$$P_B \approx \frac{1}{n}\sum_{j=t+1}^{n} j\binom{n}{j}p_c^j(1-p_c)^{n-j} = \frac{1}{127}\sum_{j=32}^{127} j\binom{127}{j}(0{,}2156)^j(1-0{,}2156)^{127-j}$$

При очень малом p_c достаточно взять лишь первые несколько членов суммы. Но если p_c большое, как в данном случае, то помощь компьютера будет очень кстати. После выполнения расчетов с $p_c = 0{,}2156$ вероятность появления ошибки в декодированном бите P_B получаем равной 0,05, что очень сильно отличается от требуемых 10^{-7}. Возьмем теперь код, степень кодирования которого близка к очень популярному значению 1/2, т.е. код (127, 64). Возможности этого кода не столь значительны, как у первого кода. Он может исправить 10 искаженных битов в блоке из 127 кодовых битов. Впрочем, исследуем этот код. Выполняя те же шаги, что и выше, получаем следующее.

$$\frac{E_s}{N_0} = \frac{E_c}{N_0} = \left(\frac{k}{n}\right)\frac{E_b}{N_0} = \left(\frac{64}{127}\right)5 = 2{,}519$$

Отметив, насколько большее E_s/N_0 получено, по сравнению с кодом (127, 8), продолжим вычисление.

$$p_c = Q(\sqrt{2 \times 2{,}519}) = Q(2{,}245) = 0{,}0124$$

$$P_B \approx \frac{1}{127}\sum_{j=11}^{127} j\binom{127}{j}(0{,}0124)^j(1-0{,}0124)^{127-j}$$

В итоге, $P_B = 5{,}6 \times 10^{-8}$, что удовлетворяет системным требованиям. Из этого примера можно видеть, что выбор кода нужно делать, рассматривая тип модуляции и имеющееся E_b/N_0. При выборе можно руководствоваться тем, что очень высокие и очень низкие степени кодирования, в основном, оказываются малоэффективными в системах связи реального времени, что ясно видно из поведения кривых на рис. 8.6 (глава 8).

9.8. Модуляция с эффективным использованием полосы частот

Основной задачей спектрально эффективных модуляций является максимизация эффективности использования полосы частот. Увеличение спроса на цифровые каналы передачи привело к исследованиям спектрально эффективных методов модуляции [8, 16], направленных на максимально эффективное использование полосы частот и, следовательно, призванных ослабить проблему спектральной перегрузки каналов связи.

В некоторых системах, помимо требования эффективности использования спектра, имеются и другие. Например, в спутниковых системах с сильно нелинейными транспондерами требуется модуляция с постоянной огибающей. Это связано с тем, что при прохождении сигнала с большими флуктуациями амплитуды нелинейные транспондеры создают паразитные боковые полосы (причина — механизм, называемый преобразованием амплитудной модуляции в фазовую). Эти боковые полосы отбирают у информационного сигнала часть мощности транспондера, а также могут интерферировать с сигналами соседних каналов (помеха соседнего канала) или других систем связи (внутриканальная помеха). Двумя примерами модуляций с постоянной огибающей, подходящими для систем с нелинейными транспондерами, являются *квадратурная фазовая манипуляция со сдвигом* (Offset QPSK — OQPSK) и *манипуляция с минимальным сдвигом* (minimum shift keying — MSK).

9.8.1. Передача сигналов с модуляцией QPSK и OQPSK

На рис. 9.10 показано разбиение типичного потока импульсов при модуляции QPSK. На рис. 9.10, *а* представлен исходный поток данных $d_k(t) = d_0, d_1, d_2, \ldots$, состоящий из биполярных импульсов, т.е. d_k принимают значения +1 или −1, представляющие двоичную единицу и двоичный нуль. Этот поток импульсов разделяется на синфазный поток, $d_I(t)$, и квадратурный, $d_Q(t)$, как показано на рис. 9.10, *б*.

$$\begin{aligned} d_I(t) &= d_0, d_2, d_4, \ldots \text{ (четные биты)} \\ d_Q(t) &= d_1, d_3, d_5, \ldots \text{ (нечетные биты)} \end{aligned} \tag{9.43}$$

Отметим, что скорости потоков $d_I(t)$ и $d_Q(t)$ равны половине скорости передачи потока $d_k(t)$. Удобную ортогональную реализацию сигнала QPSK, $s(t)$, можно получить, используя амплитудную модуляцию синфазного и квадратурного потоков на синусной и косинусной функциях от несущей.

$$s(t) = \frac{1}{\sqrt{2}} d_I(t) \cos\left(2\pi f_0 t + \frac{\pi}{4}\right) + \frac{1}{\sqrt{2}} d_Q(t) \sin\left(2\pi f_0 t + \frac{\pi}{4}\right) \tag{9.44}$$

С помощью тригонометрических тождеств (Г.5) и (Г.6) уравнение (9.44) можно представить в следующем виде.

$$s(t) = \cos\left[2\pi f_0 t + \theta(t)\right] \tag{9.45}$$

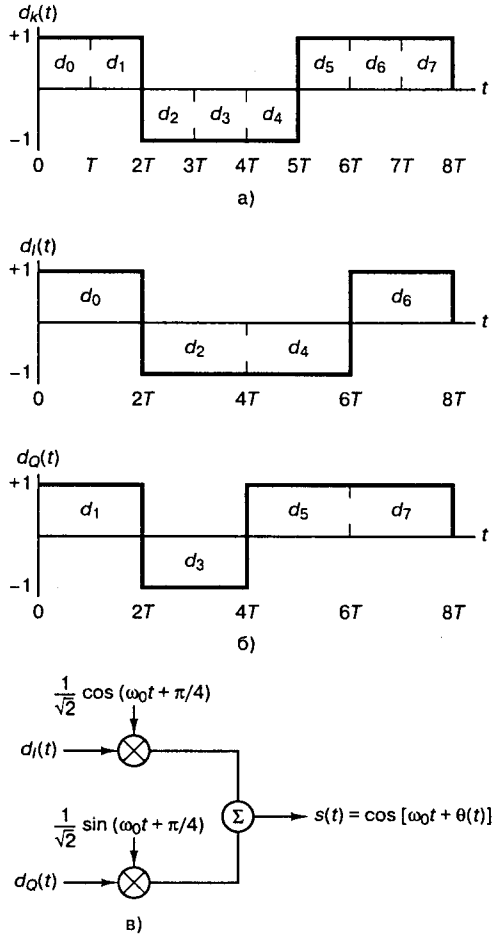


Рис. 9.10. Модуляция QPSK

Модулятор QPSK, показанный на рис. 9.10, *в*, использует сумму синусоидального и косинусоидального слагаемых, тогда как аналогичное устройство, описанное в разделе 4.6, применяет разность таких слагаемых. Материл данного раздела представлен так, как это сделано в работе [17]. Поскольку когерентный приемник должен разрешать любую неопределенность фазы, использование в передатчике иного формата фазы можно рассматривать как часть подобной неопределенности. Поток импульсов $d_I(t)$ используется для амплитудной модуляции (с амплитудой +1 или –1) косинусоиды. Это равноценно сдвигу фазы косинусоиды на 0 или π; следовательно, в результате получаем сигнал BPSK. Аналогично поток импульсов $d_Q(t)$ модулирует синусоиду, что дает сигнал BPSK, ортогональный предыдущему. При суммировании этих двух ортогональных компонентов несущей получается сигнал QPSK. Величина $\theta(t)$ будет соответствовать одному из четырех возможных сочетаний $d_I(t)$ и $d_Q(t)$ в уравнении (9.44): $\theta(t) = 0°$, $\pm 90°$ или $180°$; результирующие векторы сигналов показаны в сигнальном пространстве на рис. 9.11. Так как $\cos(2\pi f_0 t + \pi/4)$ и $\sin(2\pi f_0 t + \pi/4)$ ортогональны, два сигнала BPSK можно обнаруживать раздельно.

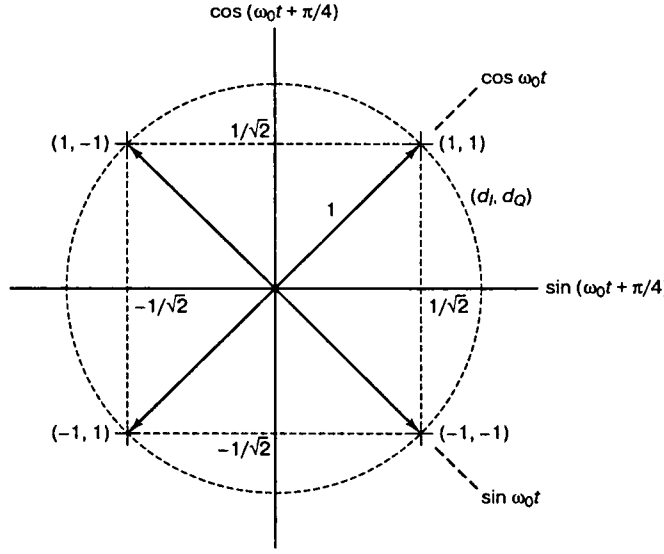


Рис. 9.11. Сигнальное пространство для схем QPSK и OQPSK

Передачу сигналов OQPSK также можно представить формулами (9.44) и (9.45); различие между двумя схемами модуляции, QPSK и OQPSK, состоит только в *ориентации* двух модулированных сигналов. Как показано на рис. 9.10, длительность каждого исходного импульса равна T (рис. 9.10, *а*); следовательно, в потоках на рис. 9.10, *б* длительность каждого импульса равна $2T$. В обычной QPSK потоки четных и нечетных импульсов передаются со скоростью $1/(2T)$ бит/с, причем они синхронизированы так, что их переходы совпадают, как показано на рис. 9.10, *б*. В OQPSK, которую иногда называют *QPSK с разнесением* (staggered QPSK — SQPSK), используется также разделение потока данных и ортогональная передача; разница в том, что потоки $d_I(t)$ и $d_Q(t)$ синхронизированы со сдвигом на T. Этот сдвиг показан на рис. 9.12.

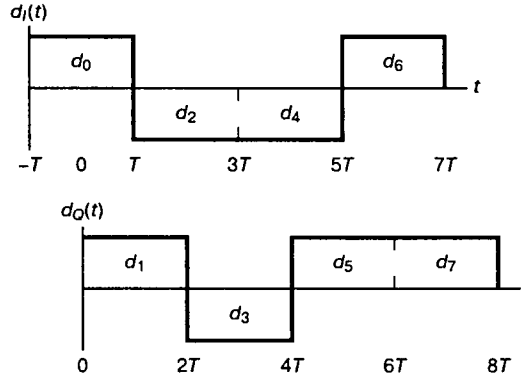


Рис. 9.12. Потоки данных при модуляции OQPSK

При стандартной QPSK из-за синхронизации $d_I(t)$ и $d_Q(t)$ за промежуток $2T$ фаза несущей может изменяться только раз. В зависимости от значений $d_I(t)$ и $d_Q(t)$ в лю-

бом промежутке $2T$, фаза несущей на этом промежутке может принимать одно из четырех значений, показанных на рис. 9.11. В течение следующего интервала $2T$ фаза несущей остается такой же, если ни один из потоков не меняет знака. Если только один из потоков импульсов изменит знак, происходит сдвиг фазы на ±90°. Изменение знака у обоих потоков приводит к сдвигу фазы на 180°. На рис. 9.13, *а* изображен типичный сигнал QPSK для последовательности $d_I(t)$ и $d_Q(t)$, показанной на рис. 9.10.

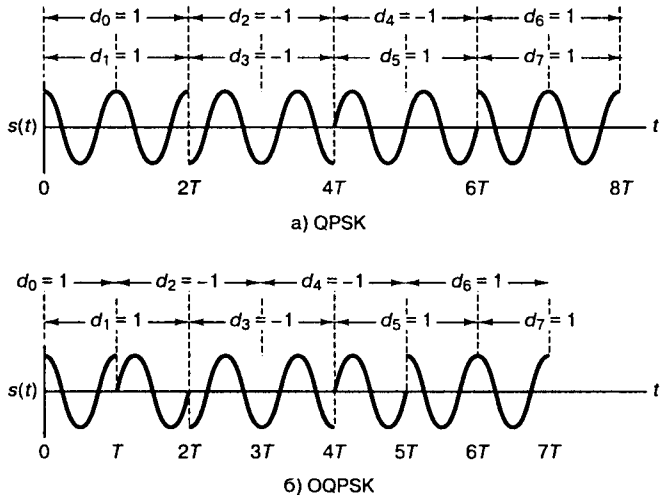


Рис. 9.13. Сигналы: а) QPSK; б) OQPSK. (Перепечатано с разрешения автора из работы Pasupathy S. "Minimum Shift Keying: A Spectrally Efficient Modulation," IEEE Commun. Mag., *July, 1979, Fig. 4, p. 17. © 1979, IEEE.)*

Если сигнал, модулированный QPSK, подвергается фильтрации для уменьшения побочных максимумов спектра, результирующий сигнал больше не будет иметь постоянной огибающей и, фактически, случайный фазовый сдвиг на 180° вызовет моментальное обращение огибающей в нуль (рис. 9.13, *а*). Если эти сигналы применяются в спутниковых каналах, где используются нелинейные усилители, постоянная огибающая будет восстанавливаться. Однако в то же время восстанавливаться будут и все *нежелательные* частотные боковые максимумы, которые могут интерферировать с сигналами соседних каналов и других систем связи.

При модуляции QPSK потоки импульсов $d_I(t)$ и $d_Q(t)$ разнесены и, следовательно, не могут одновременно изменить состояние. Несущая не может изменять фазу на 180°, поскольку за один раз переход может сделать только один из компонентов. За каждые T секунд фаза может измениться только на 0° или ±90°. На рис. 9.13, *б* показан типичный сигнал OQPSK для последовательности, представленной на рис. 9.12. Если сигнал OQPSK становится сигналом с ограниченной полосой, возникающая межсимвольная интерференция приводит к легкому спаду огибающей в области переходов фазы на ±90°, но поскольку переходов на 180° при OQPSK нет, огибающая не обращается в нуль, как это происходит при QPSK. Если сигнал OQPSK с ограниченной полосой проходит через нелинейный транспондер, спад огибающей устраняется; в то же время высокочастотные компоненты, связанные с исчезновением огибающей, не усиливаются. Таким образом, отсутствует внеполосная интерференция [17].

9.8.2. Манипуляция с минимальным сдвигом

Главное преимущество OQPSK перед QPSK (устранение внеполосной интерференции) наводит на мысль, что можно дополнительно усилить формат OQPSK, устранив разрывные переходы фазы. Это стало мотивацией разработки схем модуляции без разрыва фазы (continuous phase modulation — CPM). Одной из таких схем является *манипуляция с минимальным сдвигом* (minimum shift keying — MSK) [17, 20]. MSK можно рассматривать как частный случай *частотной манипуляции без разрыва фазы* (continuous-phase frequency shift keying — CPFSK) или как частный случай OQPSK с синусоидальным взвешиванием символов. В первом случае сигнал MSK можно представить следующим образом [18].

$$s(t) = \cos\left[2\pi\left(f_0 + \frac{d_k}{4T}\right)t + x_k\right] \quad kT < t < (k+1)T \tag{9.46}$$

Здесь f_0 — несущая частота, $d_k = \pm 1$ представляет биполярные данные, которые передаются со скоростью $R = 1/T$, а x_k — это фазовая постоянная для k-го интервала передачи двоичных данных. Отметим, что при $d_k = 1$ передаваемая частота — это $f_0 + 1/4T$, а при $d_k = -1$ — это $f_0 - 1/4T$. Следовательно, разнесение тонов в MSK составляет половину от используемого при ортогональной FSK с некогерентной демодуляцией, откуда и название — манипуляция с *минимальным* сдвигом. В течение каждого T-секундного интервала передачи данных значение x_k постоянно, т.е. $x_k = 0$ или π, что диктуется требованием непрерывности фазы сигнала в моменты $t = kT$. Это требование накладывает ограничение на фазу, которое можно представить следующим рекурсивным соотношением для x_k.

$$x_k = \left[x_{k-1} + \frac{\pi k}{2}(d_{k-1} - d_k)\right] \text{ по модулю } 2\pi \tag{9.47}$$

С помощью тождеств (Г.5) и (Г.6) уравнение (9.46) можно переписать в квадратурном представлении.

$$s(t) = a_k \cos\frac{\pi t}{2T}\cos 2\pi f_0 t - b_k \sin\frac{\pi t}{2T}\sin 2\pi f_0 t \quad kT < t < (k+1)T, \tag{9.48}$$

где

$$\begin{aligned} a_k &= \cos x_k = \pm 1 \\ b_k &= d_k \cos x_k = \pm 1. \end{aligned} \tag{9.49}$$

Синфазный компонент обозначается как $a_k \cos(\pi t/2T)\cos 2\pi f_0 t$, где $\cos 2\pi f_0 t$ — несущая, $\cos(\pi t/2T)$ — *синусоидальное взвешивание символов*, a_k — информационно-зависимый член. Подобным образом квадратурный компонент — это $b_k \sin(\pi t/2T)\sin 2\pi f_0 t$, где $\sin 2\pi f_0 t$ — квадратурное слагаемое несущей, $\sin(\pi t/2T)$ — такое же синусоидальное взвешивание символов, а b_k — информационно-зависимый член. Может показаться, что величины a_k и b_k могут изменять свое значение каждые T секунд. Однако из-за требования непрерывности фазы величина a_k может измениться лишь при переходе функции $\cos(\pi t/2T)$ через нуль, а b_k — только при переходе через нуль $\sin(\pi t/2T)$. Следовательно, взвешивание символов в синфазном или квадратурном канале — это синусоидальный импульс с

периодом $2T$ и переменным знаком. Как и в случае OQPSK, синфазный и квадратурный компоненты сдвинуты относительно друг друга на T секунд.

Отметим, что x_k в уравнении (9.46) — это функция разности между прежним и текущим информационными битами (дифференциальное кодирование). Таким образом, величины a_k и b_k в уравнении (9.48) можно рассматривать как *дифференциально кодированные* компоненты исходных данных d_k. Однако чтобы биты данных d_k были независимы между собой, знаки последовательных импульсов квадратурного и синфазного каналов различных интервалов, длительностью $2T$ секунд, должны быть случайными импульсами. Таким образом, если уравнение (9.48) рассматривать как частный случай модуляции OQPSK, его можно переписать в иной (недифференциальной) форме [18].

$$s(t) = d_I(t)\cos\frac{\pi t}{2T}\cos 2\pi f_0 t - d_Q(t)\sin\frac{\pi t}{2T}\sin 2\pi f_0 t \tag{9.50}$$

Здесь $d_I(t)$ и $d_Q(t)$ имеют такой же смысл синфазного и квадратурного потоков данных, как и в уравнении (9.43). Схема MSK, записанная в форме (9.50), иногда называется MSK с *предварительным кодированием* (precoded MSK). Графическое представление уравнения (9.50) дано на рис. 9.14. На рис. 9.14, *а* и *в* показано синусоидальное взвешивание импульсов синфазного и квадратурного каналов. Эти последовательности представляют собой те же информационные последовательности, что и на рис. 9.12, но здесь умножение на синусоиду дает более плавные переходы фазы, чем в исходном представлении данных. На рис. 9.14, *б* и *г* показана модуляция ортогональных компонентов $\cos(2\pi f_0 T)$ и $\sin(2\pi f_0 T)$ синусоидальными потоками данных. На рис. 9.14, *д* представлено суммирование ортогональных компонентов, изображенных на рис. 9.14, *б* и *г*. Итак, из уравнения (9.50) и рис. 9.14 можно заключить следующее: 1) сигнал $s(t)$ имеет постоянную огибающую; 2) фаза радиочастотной несущей непрерывна при битовых переходах; 3) сигнал $s(t)$ можно рассматривать как сигнал, модулированный FSK, с частотами передачи $f_0 + 1/4T$ и $f_0 - 1/4T$. Таким образом, минимальное разнесение тонов, требуемое при модуляции MSK, можно записать следующим образом.

$$\left(f_0 + \frac{1}{4T}\right) - \left(f_0 - \frac{1}{4T}\right) = \frac{1}{2T}, \tag{9.51}$$

что равно половине скорости передачи битов. Отметим, что разнесение тонов, требуемое для MSK, — это половина ($1/T$) разнесения, необходимого при некогерентном обнаружении сигналов, модулированных FSK (см. раздел 4.5.4). Это объясняется тем, что фаза несущей известна и непрерывна, что позволяет осуществить когерентную демодуляцию сигнала.

Спектральная плотность мощности $G(f)$ для QPSK и OQPSK имеет следующий вид [18].

$$G(f) = 2PT\left(\frac{\sin 2\pi fT}{2\pi fT}\right)^2, \tag{9.52}$$

где P — средняя мощность модулированного сигнала. При MSK $G(f)$ будет иметь следующий вид [18].

$$G(f) = \frac{16PT}{\pi^2}\left(\frac{\cos 2\pi fT}{1 - 16f^2T^2}\right)^2 \tag{9.53}$$

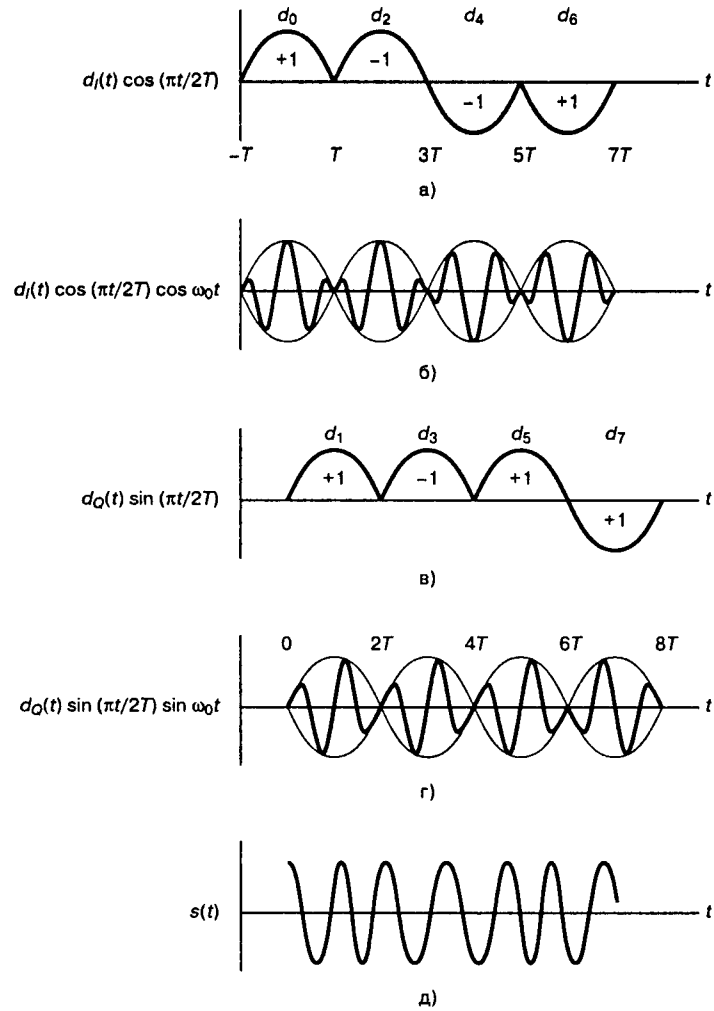


Рис. 9.14. Манипуляция с минимальным сдвигом (minimum shift keying — MSK): а) модифицированный синфазный поток битов; б) произведение синфазного потока битов и несущей; в) модифицированный квадратурный поток битов; г) произведение квадратурного потока битов и несущей; д) сигнал MSK. (Перепечатано с разрешения автора из работы Pasupathy S. "Minimum Shift Keying: A Spectrally Efficient Modulation," IEEE Commun. Mag., *July, 1979, Fig. 5, p. 18. © 1979, IEEE.)*

Нормированная спектральная плотность мощности ($P = 1$ Вт) для QPSK, OQPSK и MSK изображена на рис. 9.15. Для сравнения здесь же приводится спектральный график BPSK. Не должно удивлять, что BPSK требует большей полосы пропускания, чем другие типы модуляции, при том же уровне спектральной плотности. В разделе 9.5.1 и на рис. 9.6 было показано, что теоретическая эффективность использования полосы частот схемы BPSK вдвое меньше, чем схемы QPSK. Из рис. 9.15 видно, что боковые максимумы графика MSK ниже, чем графика QPSK или OQPSK. Это является следствием умножения потока данных на синусоиду и дает большое количество *плавных фазовых переходов*. Чем плавнее переход, тем быстрее спектральные хвосты стремятся к нулю. Модуляция MSK *спектраль-*

но эффективнее QPSK или OQPSK; тем не менее, как видно из рис. 9.15, спектр MSK имеет более широкий основной максимум, чем спектр QPSK или OQPSK. Следовательно, MSK нельзя назвать удачным выбором при наличии узкополосных линий связи. В то же время MSK стоит использовать в системах с несколькими несущими, поскольку ее относительно малые побочные максимумы спектра позволяют избежать значительных помех соседних каналов (adjacent channel interference — ACI). То, что спектр QPSK имеет более узкий основной максимум, чем MSK, объясняется тем, что при данной скорости передачи битов скорость передачи символов QPSK вдвое меньше скорости передачи символов MSK.

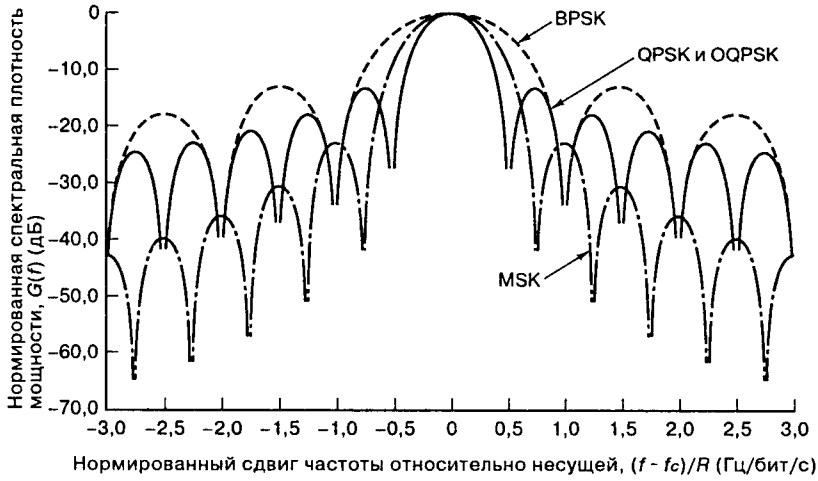


Рис. 9.15. Нормированная спектральная плотность мощности для BPSK, QPSK, OQPSK и MSK. (Перепечатано с разрешения автора из работы Amoroso F. "The Bandwidth of Digital Data Signals," IEEE Commun. Mag., vol. 18, n. 6, November, 1980, Fig. 2A, p. 16. © 1980, IEEE.)

9.8.2.1. Вероятность ошибки при модуляциях OQPSK и MSK

Ранее говорилось, что BPSK и QPSK имеют одинаковую вероятность появления битовой ошибки, поскольку QPSK сконфигурирована как два сигнала BPSK на ортогональных компонентах несущей. Так как разнесение потоков данных не меняет ортогональности несущих, схема OQPSK имеет ту же теоретическую вероятность появления битовой ошибки, что и BPSK и QPSK.

Для модуляции двух квадратурных компонентов несущей манипуляция с минимальным сдвигом использует сигналы антиподной формы, $\pm\cos(\pi t/2T)$ и $\pm\sin(\pi t/2T)$, с периодом $2T$. Следовательно, если для независимого восстановления данных из каждого ортогонального компонента используются согласованные фильтры, то модуляция MSK, определенная в формуле (9.50), имеет ту же вероятность появления ошибки, что и BPSK, QPSK и OQPSK [17]. Однако если сигнал, модулированный MSK, когерентно обнаруживается в интервале T секунд как сигнал, модулированный FSK, то эта вероятность будет ниже, чем у BPSK, на 3 дБ [17]. У MSK с дифференциально кодированными данными, определенной в выражении (9.46), вероятность появления ошибки будет такой же, как и при когерентном обнаружении дифференциально кодированных данных в модуляции PSK. Сигналы MSK также можно обнаруживать некогерентно

[19]. Это позволяет осуществлять дешевую демодуляцию (если это позволяет величина принятого E_b/N_0).

9.8.3. Квадратурная амплитудная модуляция

Когерентная M-арная фазовая манипуляция (M-ary phase shift keying — MPSK) — это хорошо известный метод, позволяющий сузить полосу пропускания. Здесь используется не бинарный алфавит с передачей одного информационного бита за период передачи канального символа, а алфавит из M символов, что позволяет передавать $k = \log_2 M$ битов за каждый символьный интервал. Поскольку использование M-арных символов в k раз повышает скорость передачи информации при той же полосе пропускания, то при фиксированной скорости применение M-арной PSK сужает необходимую полосу пропускания в k раз (см. раздел 4.8.3).

Из уравнения (9.44) можно видеть, что модуляция QPSK состоит из двух независимых потоков. Один поток модулирует амплитуду косинусоидальной функции несущей на уровни +1 и –1, а другой — аналогичным образом синусоидальную функцию. Результирующий сигнал называется двухполосным сигналом с подавлением несущей (double-sideband suppressed-carrier — DSB-SC), поскольку полоса радиочастот вдвое больше полосы немодулированного сигнала (см. раздел 1.7.1) и не содержит выделенной несущей. *Квадратурную амплитудную модуляцию* (quadrature amplitude modulation — QAM) можно считать логическим продолжением QPSK, поскольку сигнал QAM также состоит из двух независимых амплитудно-модулированных несущих. Каждый блок из k бит (k полагается четным) можно разделить на два блока из $k/2$ бит, подаваемых на цифро-аналоговые преобразователи (ЦАП), которые обеспечивают требующее модулирующее напряжение для несущих. В приемнике оба сигнала обнаруживаются независимо с помощью согласованных фильтров. Передачу сигналов, модулированных QAM, можно также рассматривать как комбинацию амплитудной (amplitude shift keying — ASK) и фазовой (phase shift keying — PSK) манипуляций, откуда альтернативное название *амплитудно-фазовая манипуляция* (amplitude phase keying — APK). И наконец, ее можно считать двухмерной амплитудной манипуляцией, откуда еще одно название — *квадратурная амплитудная манипуляция* (quadrature amplitude shift keying — QASK).

На рис. 9.16, *а* показано двухмерное пространство сигналов и набор векторов сигналов, модулированных 16-ричной QAM и изображенных точками, которые расположены в виде прямоугольной совокупности. На рис. 9.16, *б* показан канонический модулятор QAM. На рис. 9.16, *в* изображен пример модели канала, в которой предполагается наличие лишь гауссова шума. Сигналы передаются в виде пары (x, y). На модели показано, что координаты сигнальной точки (x, y) передаются по раздельным каналам и независимо возмущаются переменными гауссова шума (n_x, n_y), каждый компонент которого имеет нулевое среднее и дисперсию N. Можно также сказать, что двухмерная точка сигнала возмущается двухмерной переменной гауссова шума. Если средняя энергия сигнала (среднеквадратическое значение координат сигнала) равна S, тогда отношение сигнал/шум равно S/N. Простейший метод цифровой передачи сигналов через подобные системы — это применение одномерной амплитудно-импульсной модуляции (pulse amplitude modulation — PAM) независимо к каждой координате сигнала. При модуляции PAM для передачи k битов/размерность по гауссову каналу каждая точка сигнала принимает значение одной из 2^k равновероятных экви-

дистантных амплитуд. Точки сигналов принято группировать в окрестности пространства на амплитудах ±1, ±3, ..., $\pm(2^k - 1)$.

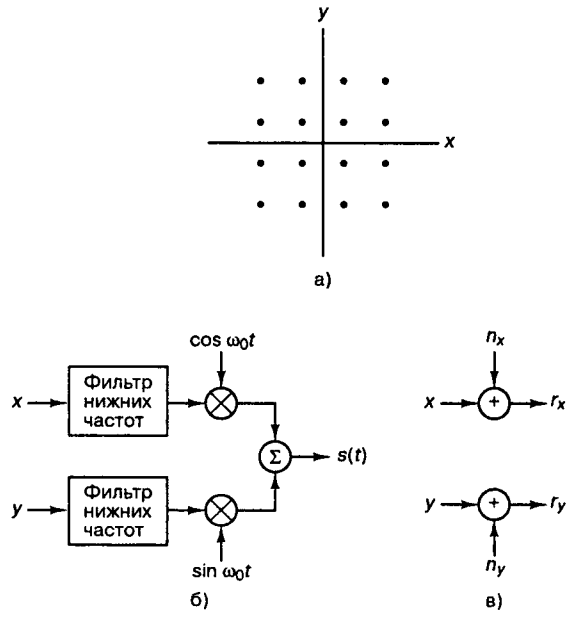


Рис. 9.16. Схема модуляции QAM: а) 16-ричное пространство сигналов; б) канонический модулятор QAM; в) модель канала QAM

9.8.3.1. Вероятность битовой ошибки при модуляции QAM

Для прямоугольной совокупности, гауссова канала и приема с помощью согласованных фильтров, вероятность появления битовой ошибки выражается следующим образом [12].

$$P_B \approx \frac{2(1-L^{-1})}{\log_2 L} Q\left[\sqrt{\left(\frac{3\log_2 L}{L^2-1}\right)\frac{2E_b}{N_0}}\right] \tag{9.54}$$

Здесь $Q(x)$ определено в формуле (3.43), а L представляет количество уровней амплитуды в одном измерении. Предполагается, что при отображении последовательности $\log_2 L$ бит в L-арный символ используется код Грея (см. раздел 4.9.4).

9.8.3.2. Компромисс между полосой пропускания и мощностью

На рис. 9.6 представлена плоскость эффективности использования полосы частот, на которой показан компромисс между полосой пропускания и мощностью при M-арной модуляции QAM, если вероятность битовой ошибки равна 10^{-5}, а значения на оси абсцисс измеряются в среднем отношении E_b/N_0. Предполагается, что немодулированные импульсы фильтруются по Найквисту, так что двусторонняя полоса пропускания на промежуточной частоте (Intermediate Frequency — IF) равна $W_{IF} = 1/T$, где T — длительность передачи символа. Следовательно, эффективность использования полосы частот равна $R/W = \log_2 M$, где M — размер набора символов. Для реальных ка-

налов и сигналов достоверность передачи ниже указанной, поскольку для реализации реальных фильтров требуется большая полоса пропускания. Из рис. 9.6 видно, что QAM — это метод снижения требований к полосе пропускания при передаче цифровых данных. Как и при M-арной PSK, за счет снижения эффективности использования полосы частот можно получить выигрыш в мощности или E_b/N_0; однако при QAM можно реализовать *более выгодный компромисс*, чем при M-арной PSK.

Пример 9.5. Выбор схемы модуляции

Пусть поток данных со скоростью R = 144 Мбит/с передается по радиочастотному каналу с использованием двухполосной схемы модуляции. Предполагается фильтрация по Найквисту и наличие двусторонней полосы 36 МГц. Какую модуляцию стоит выбрать при данных требованиях? Если доступное E_b/N_0 равно 20, какой будет вероятность битовой ошибки?

Решение

Запишем требуемую спектральную эффективность.

$$\frac{R}{W} = \frac{144 \text{ Мбит/с}}{36 \text{ МГц}} = 4 \text{ бит/с/Гц}$$

Из рис. 9.6 видно, что 16-ричная QAM с теоретической спектральной эффективностью 4 бит/с/Гц требует более низкого значения E_b/N_0, чем 16-арная PSK, при том же значении P_B. Исходя из этого выбираем модем QAM.

Считая E_b/N_0 равным 20 и используя формулу (9.54), вычисляем ожидаемую вероятность битовой ошибки.

$$P_B \approx \frac{3}{4} Q\left(\sqrt{\frac{4}{5}\frac{E_b}{N_0}}\right) = 2{,}5 \times 10^{-5}$$

Пример 9.6. Спектральная эффективность

а) Объясните схему расчета спектральной эффективности схемы QAM в примере 9.5, считая что сигнал, модулированный QAM, передается на ортогональных компонентах несущей.

б) Поскольку двусторонняя полоса пропускания в примере 9.5 равна 36 МГц, рассмотрим использование половины этого значения для передачи потока данных со скоростью 144 Мбит/с при многоуровневой схеме PAM. Какая спектральная эффективность нужна для осуществления этого и какое количество уровней необходимо в схеме PAM? Предполагается фильтрация по Найквисту.

Решение

а) *Полосовой канал с использованием схемы QAM*: поток данных со скоростью 144 Мбит/с разделяется на синфазный поток со скоростью 72 Мбит/с и квадратурный поток с такой же скоростью (72 Мбит/с); один поток модулирует амплитуду косинусоидальной функции несущей в полосе пропускания 36 МГц, а другой поток аналогичным образом модулирует синусоидальную функцию. Поскольку каждый поток со скоростью 72 Мбит/с модулирует ортогональный компонент несущей, 36 МГц достаточно для обоих потоков или для передачи со скоростью 144 Мбит/с. Следовательно, спектральная эффективность равна (144 Мбит/с)/36 МГц = 4 бит/с/Гц.

б) *Требуемая спектральная эффективность при узкополосной передаче равна следующему.*

$$\frac{R}{W} = \frac{144 \text{ Мбит/с}}{18 \text{ МГц}} = 8 \text{ бит/с/Гц}$$

Если предполагается фильтрация по Найквисту, полоса пропускания 18 МГц поддерживает максимальную скорость передачи символов $R_s = 2W = 3 \times 10^6$ символ/с (см. уравнение (3.80)). Следовательно, каждый импульс, модулированный PAM, должен иметь l-битовое значение.

$$R = lR_s$$

Откуда

$$\frac{R}{W} = \frac{144 \text{ Мбит/с}}{36 \times 10^6 \text{ выборок/с}} = 4 \text{ бит/импульс},$$

где $l = \log_2 L$, а $L = 16$ уровней.

9.9. Модуляция и кодирование в каналах ограниченной полосы

Методы канального кодирования, описанные в главах 6–8, обычно *не* применяются в телефонных каналах (хотя первые испытания последовательного декодирования сверточных кодов проводились именно по телефонной линии). Недавно, однако, возник существенный интерес к методам, которые могут обеспечить эффективное кодирование в узкополосных каналах. Это связано с желанием получить надежную передачу по телефонным линиям при *высоких скоростях передачи данных*. Потенциальная эффективность составляет порядка 3 бит/символ (при данном отношении сигнал/шум) или, что то же самое, при данной вероятности ошибки можно достичь экономии мощности до 9 дБ [21].

Наибольший интерес представляют следующие три отдельные области исследования кодирования.

1. Оптимальные границы совокупностей сигналов (выбор наиболее плотно упакованного подмножества сигналов из любого регулярного массива или решетки возможных точек).
2. Структуры решеток с высокой плотностью (улучшение выбора подмножества сигналов за счет начала рассмотрения с наиболее плотной из возможных решеток пространства).
3. Решетчатое кодирование (комбинация методов модуляции и кодирования для получения эффективного кодирования в узкополосных каналах).

Первые две области не являются "истинными" схемами кодирования с защитой от ошибок. Под словами "истинная схема кодирования с защитой от ошибок" подразумевается метод, использующий некоторую структурную избыточность для снижения вероятности ошибки. Избыточность включает лишь третья позиция списка, решетчатое кодирование. Перечисленные области исследования кодирования и ожидаемые от них улучшения производительности обсуждаются ниже.

9.9.1. Коммерческие модемы

В использовании эффективных методов модуляции традиционно заинтересована телекоммуникационная индустрия, поскольку основные ресурсы телефонных компаний — это жестко ограниченные речевые (телефонные) каналы. Типичный телефонный ка-

нал характеризуется высоким отношением сигнал/шум (signal-to-noise ratio — SNR) порядка 30 дБ и полосой пропускания порядка 3 кГц. В табл. 9.4 представлена эволюция некоммутируемых телефонных модемов, а в табл. 9.5 — эволюция стандартов коммутируемых телефонных модемов.

9.9.2. Границы совокупности сигналов

Исследователи [22–26] изучили большое количество возможных совокупностей сигналов QAM, пытаясь найти структуру, которая снизит вероятность появления ошибок при данном среднем отношении сигнал/шум. На рис. 9.17 показано несколько примеров совокупностей символов для $M = 4$, 8 и 16, которые рассматривались в [22]. Циклические наборы обозначаются как (a, b, ...), где a — количество сигналов во внутреннем круге, b — сигналы следующего круга и т.д. В общем, правило совокупности, известное как правило построения Кампопьяно-Глейзера (Campopiano-Glazer) [24], которое дает оптимальные характеристики множества сигналов, можно сформулировать так: *из бесконечного массива точек, плотно упакованных в* регулярный массив или решетку, *в качестве совокупности сигналов выбрать плотно упакованное подмножество* 2^k *точек*. В данном случае "оптимальный" означает среднюю минимальную среднюю или пиковую мощность при данной вероятности ошибки. В двухмерном пространстве сигналов оптимальная граница, окружающая массив точек, стремится к окружности. На рис. 9.18 показаны примеры 64-арной ($k = 6$) и 128-арной ($k = 7$) совокупностей сигналов из прямоугольного массива. Крестообразные границы — это компромиссное представление оптимальной окружности. Совокупность $k = 6$ использована в модеме Paradyne 14,4 Кбит/с. По сравнению с квадратной, кольцевая граница дает улучшение характеристик всего на 0,2 дБ [21].

Таблица 9.4. Эволюция некоммутируемых телефонных модемов

Год	Название	Максимальная скорость передачи битов (бит/с)	Скорость передачи сигналов (символов/с)	Метод модуляции	Эффективность передачи сигналов (бит/символ)
1962	Bell 201	2400	1200	4-PSK	2
1967	Milgo 4400/48	4800	1600	8-PSK	3
1971	Codex 9600C	9600	2400	16-QAM	4
1980	Paradyne MP14400	14 400	2400	64-QAM	6
1981	Codex SP14.4	14 400	2400	64-QAM	6
1984	Codex 2660	16 800	2400	Решетчатая 256-QAM	7
1985	Codex 2680	19 200	2743	8-D решетчатая 160-QAM	7

Таблица 9.5. Эволюция стандартов коммутируемых телефонных модемов

Год	Название	Максимальная скорость передачи битов (бит/с)	Скорость передачи сигналов (символов/с)	Метод модуляции	Эффективность передачи сигналов (бит/символ)
1984	V.32	9600	2400	2-D решетчатая 32-QAM	4
1991	V.32bis	14 400	2400	2-D решетчатая 128-QAM	6
1994	V.34	28 800	2400, 2743, 2800, 3000, 3200, 3429	4-D решетчатая 960-QAM	≈ 9
1996	V.34	33 600	2400, 2743, 2800, 3000, 3200, 3429	4-D решетчатая 1664-QAM	≈ 10
1998	4.90	по направлению основного потока: 56 000	8000	PCM*(*M*-PAM)	7
		против направления основного потока: 33 600	как в V.34	как в V.34	≈ 10
2000	V.92	по направлению основного потока: 56 000	8000	PCM*(*M*-PAM)	7
		против направления основного потока: 48 000	8000	Решетчатая PCM*	6

* В Рекомендации ITU-T G.711 "PCM" — это термин, используемый для *M*-арной передачи сигналов по схеме PAM.

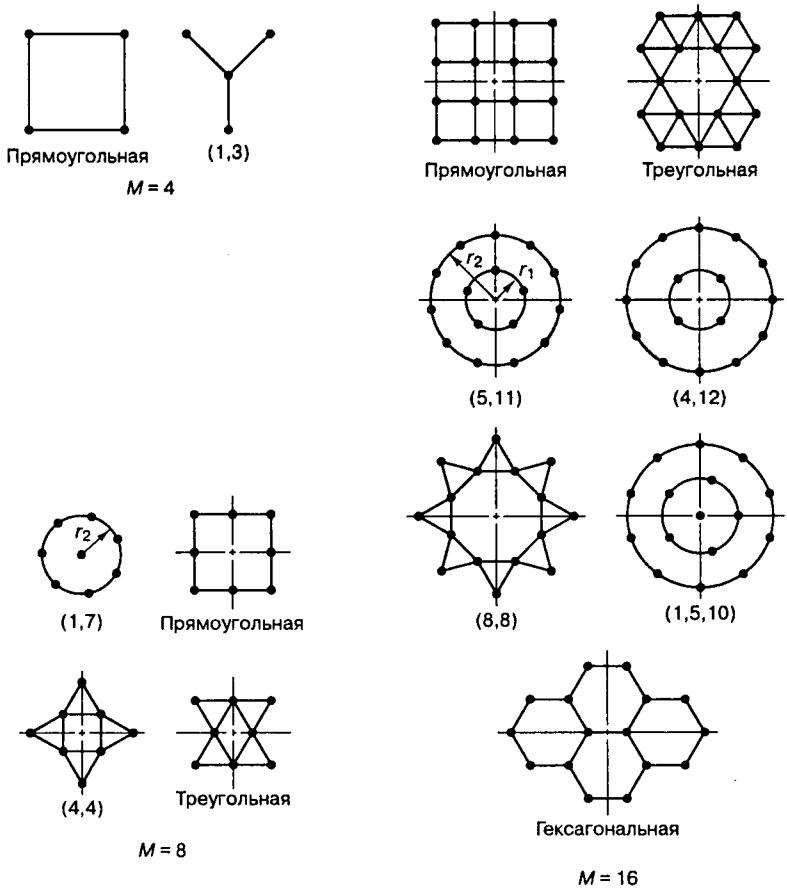


Рис. 9.17. Совокупности M-арных символов. (Перепечатано с разрешения авторов из работы Thomas C. M., Weidner M. Y. and Durrani S. N. "Digital Amplitude-Phase Shift Keying with M-ary Alphabets," IEEE Trans. Commun., *vol. COM22, n. 2, February, 1974, Figs. 2 and 3, p. 170. © 1974, IEEE.)*

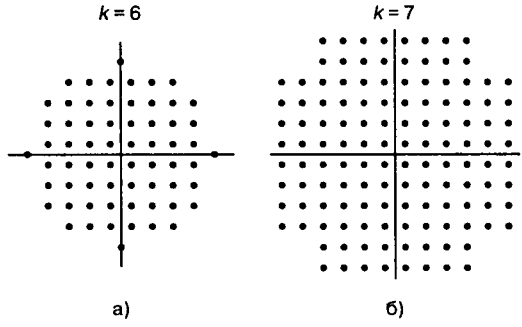


Рис. 9.18. Примеры M-арных совокупностей, использующих прямоугольную решетку

9.9.3. Совокупности сигналов высших размерностей

Для любой скорости передачи информации и шумового процесса в канале, который независимо и одинаково распределен в двух измерениях, передача сигнала в двухмерном пространстве может дать такую же вероятность ошибки при меньшей средней (или пиковой) мощности, как и передача сигналов в одномерном импульсно-амплитудном (pulse-amplitude — PAM) пространстве. Это выполняется посредством выбора точек сигналов на двухмерной решетке в пределах кольцевой, а не квадратной границы. Аналогичным образом, переходя к измерению высшей размерности N и выбирая точки на n-мерной решетке в пределах не N-мерного куба, а N-мерной сферы, можно еще больше сэкономить энергию [27–30]. Задачей подобного формирования совокупности является снижение требуемой средней энергии точек сигнала, расположенных внутри N-мерной сферы, по сравнению с энергией точек, расположенных внутри N-мерного куба. Такое снижение требуемой энергии при данной вероятности ошибки называется *эффективностью выбора формы* (shaping gain) [16]. В табл. 9.6 показано, как можно сэкономить энергию в N измерениях. Если устремить N к бесконечности, эффективность будет стремиться к 1,53 дБ; как правило, эффективность порядка 1 дБ получить нетрудно [16, 21].

Таблица 9.6. Экономия энергии при замене N-мерного куба N-мерной сферой (эффективность выбора формы)

Размерность (N)	Эффективность (дБ)
2	0,20
4	0,45
8	0,73
16	0,98
24	1,01
32	1,17
48	1,26
64	1,31

Источник: G. D. Forney, Jr., et. al. "Efficient Modulation for Bandlimited Channels," *IEEE J. Sel. Areas Commun.*, vol. SAC2, n. 5, September, 1984, pp. 632–647.

Канал, по сути, является двухмерным, поскольку символы, представленные на двухмерной плоскости в виде точек, передаются квадратурным образом. Многомерная передача сигналов обычно означает передачу с использованием двух или большего числа таких плоскостей. Для передачи n бит/символ при N-мерной (N четное и большее 2) передаче входящие биты группируются в блоки размером $nN/2$. Затем требуется выполнить отображение, при котором значения информационных битов присваиваются $2^{nN/2}$ N-мерным векторам, имеющим минимальную энергию среди всех векторов пространства. Соответствующее обратное отображение производится приемником.

Рассмотрим пример отображения сигналов из двухмерного пространства в четырехмерное. Сначала имеется двухмерная M-арная совокупность сигналов, например, M-QAM с $M = 16$. Здесь переданный символ, рассматриваемый как точка на плоскости, представляется $n = 4$ бит (две квадратурные амплитуды, по два бита

на амплитуду). Каждая передача символа состоит из передачи вектора, принадлежащего пространству из 16 возможных векторов. При четырехмерной передаче сигналов переданный символ (рассматриваемый как две точки, по одной на каждой из двух плоскостей) представляется 8 бит. Тогда, каждая (двухточечная) передача состоит из передачи вектора из пространства 16 × 16 = 256 векторов. Вообще, исходные биты данных группируются в блоки размером $nN/2$ бит. В данном примере четырехмерной передачи сигналов информационные биты группируются в блоки из 8 бит (2 плоскости × n = 4 бит/плоскость). Такую 8-битовую передачу можно рассматривать как отображение из пространства 2^n двухмерных векторов в пространство $2^{nN/2}$ четырехмерных векторов. Для четырехмерной системы, изображенной на рис. 9.19, данный источник производит один из 256 четырехмерных векторов m_i (i = 1, 2, ..., 256) путем группирования двух 16-ричных символов (двух плоскостей) за раз и передает сигналы $a_j s(t)$, $b_j s(t)$, $c_j s(t)$, $d_j s(t)$, где j = 1, ..., 4 представляет одно 4-ричное значение амплитуды. Эти узкополосные или полосовые сигналы передаются по раздельным неинтерферирующим каналам. В каждом канале сигналы независимо искажаются AWGN, и в приемнике они демодулируются с помощью согласованных фильтров. Передавать N-мерный сигнал можно посредством следующих способов.

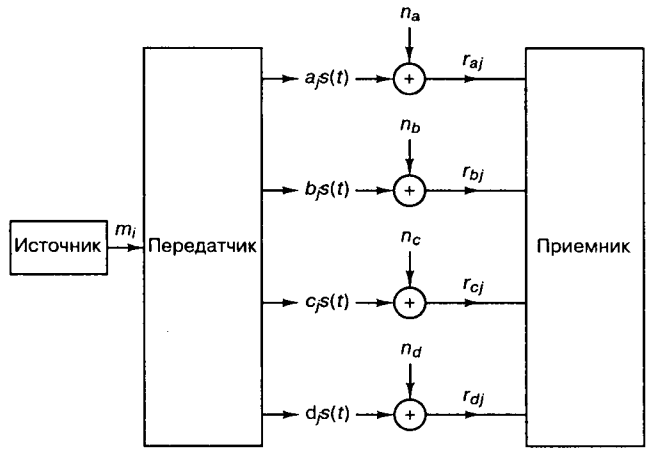


Рис. 9.19. Конфигурация четырехмерной системы

1. С помощью четырех отдельных проводников, представляющих четыре узкополосных канала.
2. С использованием двух полосовых каналов, в каждом из которых раздельно модулированы синфазный и квадратурный компоненты.
3. Путем уплотнения с частотным и временным разделением для создания нескольких узкополосных или полосовых каналов в общей линии передачи.
4. При помощи ортогональной поляризации электромагнитных волн.

Таким образом, если пример на рис. 9.19 представляет радиосистему, можно следовать методу 2 и квадратурным образом модулировать сигналы $a_j s(t)$ и $b_j s(t)$ на одной несущей, а сигналы $c_j s(t)$ и $d_j s(t)$ — на другой. Таким образом, в течение

каждого интервала, длительностью $2T$ секунд, можно передать четыре 4-ричных числа, представляющих 8 бит или вектор из 256-ричного пространства. Дополнительной эффективности выбора формы можно достичь аналогичным образом при использовании 16-ричных символов на плоскости с шестимерной передачей сигналов, если передача 16-ричного символа со всех трех плоскостей происходит каждые $3T$ секунд. Таким образом, каждый шестимерный сигнал содержит три 16-ричные величины, представляющие 12 бит или точку в пространстве 4096 сигналов. Важно подчеркнуть, что это — не просто эффективная группировка 16-ричных символов. Эффективность проявляется вследствие того, что обнаружение, выполняемое в большем пространстве сигналов, может дать нужную достоверность передачи при более низком значении E_b/N_0. При передаче 16-ричных символов с помощью шестимерной передачи сигналов каждые $3T$ секунд обнаруживается последовательность из 12 бит (не 4 бит за T секунд!). Обнаружение в пространстве большей размерности требует более сложной реализации. В основном, уменьшение сложности отображения происходит за счет снижения эффективности использования энергии.

9.9.4. Решетчатые структуры высокой плотности

В разделе 9.9.3 описывался выбор плотно упакованного подмножества точек из регулярного массива или решетки. Здесь будет рассмотрено дополнительное улучшение, поэтому мы начнем с *наиболее плотной решетки* пространства. В двухмерном пространстве сигналов наиболее плотной решеткой является гексагональная (проверьте, пытаясь наиболее плотно уложить монеты на столе!). Результатом замены прямоугольной решетки, подобной показанным на рис. 9.18, на гексагональную является экономия средней энергии до 0,6 дБ. На рис. 9.20 показано несколько примеров гексагональной упаковки. Представленная на рис. 9.20, *а* совокупность была открыта Фоскини (Foschini) и др. [26] и является самым лучшим методом из известных 16-ричных размещений. Расположение точек, показанное на рис. 9.20, *б*, было использовано в модеме Codex SP14.4.

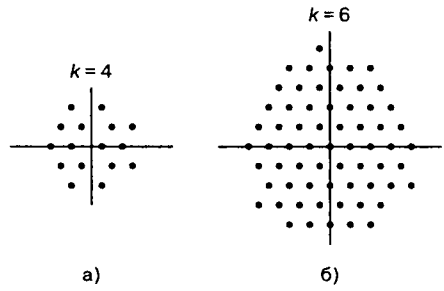


Рис. 9.20. Пример М-арных совокупностей с гексагональным расположением элементов

Гексагональная решетка является оптимальной для двух измерений. В пространствах более высоких размерностей имеются другие решетчатые структуры, которые дают наиболее плотную упаковку. В табл. 9.7 приводится улучшение (в децибелах), возникающее при переходе от применения прямоугольных решеток к лучшим из известных на настоящее время способам плотной упаковки.

Таблица 9.7. Экономия энергии при плотной решетке по сравнению с прямоугольной

Размерность (N)	Эффективность плотной решетки (дБ)
2	0,62
4	1,51
8	3,01
16	4,52
24	6,02
32	6,02
48	7,78
64	8,09

Источник: G. D. Forney, Jr., et. al. "Efficient Modulation for Bandlimited Channels," *IEEE J. Sel. Areas Commun.*, vol. SAC2, n. 5, September, 1984, pp. 632–647.

9.9.5. Комбинированная эффективность: отображение на *N*-мерную сферу и плотная решетка

Преимущества границы Кампопьяно-Глейзера в N измерениях можно объединить с преимуществами наиболее плотной решетки в N-мерном пространстве. Суммарный выигрыш — это комбинация выигрыша N-мерной сферы по сравнению с N-мерным кубом (табл. 9.6) и преимущества плотной упаковки решетки (табл. 9.7). Экономия энергии, получаемая в результате такого объединения, представлена в табл. 9.8.

Таблица 9.8. Суммарная экономия энергии при использовании максимально плотной решетки и замене *N*-мерного куба *N*-мерной сферой

Размерность (N)	Экономия энергии (дБ)
2	0,82
4	1,96
8	3,74
16	5,50
24	7,12
32	7,19
48	9,04
64	9,40

Источник: G. D. Forney, Jr., et. al. "Efficient Modulation for Bandlimited Channels," *IEEE J. Sel. Areas Commun.*, vol. SAC2, n. 5, September, 1984, pp. 632–647.

9.10. Решетчатое кодирование

При использовании в системах связи реального времени кодов коррекции ошибок, описанных в главах 6–8, достоверность передачи улучшается за счет расширения полосы частот. Как для блочных, так и для сверточных кодов преобразование каждого k-кортежа входных данных в более длинный n-кортеж кодового слова требует дополнительного расширения полосы пропускания. Вследствие этого в прошлом кодирование не было

особенно популярно в узкополосных каналах (таких, как телефонные), в которых расширять полосу частот сигнала было нецелесообразно. Однако приблизительно с 1984 года возникает активный интерес к схемам, где модуляция объединяется с кодированием; такие схемы называются *решетчатым кодированием* (trellis-coded modulation — TCM). Эти схемы позволяют повысить достоверность передачи, не расширяя при этом полосу частот сигнала. Схемы TCM используют избыточную небинарную модуляцию плюс *конечный автомат* (кодер). Что такое "конечный автомат" (finite-state machine) и какой смысл имеют его состояния? Конечный автомат (или автомат с конечным числом состояний) — это общее название устройств, обладающих памятью о прошлых сигналах; прилагательное *конечный* подчеркивает то, что существует ограниченное число однозначных состояний, которые может принимать система. Какой смысл заложен в понятие *состояние* конечного автомата? В наиболее общем смысле, состояние состоит из минимального объема информации, который, совместно с текущими данными на входе, может предсказывать данные на выходе системы. Состояние несет информацию о прошлых событиях и ограниченном наборе возможных данных на выходе в будущем. Будущие состояния ограничиваются прошлыми состояниями.

Кодер TCM с конечным числом состояний для каждого символьного интервала из набора сигналов выбирает один, формируя, таким образом, передаваемую последовательность кодированных сигналов. Полученный зашумленный сигнал обнаруживается и декодируется детектором/декодером, работающим согласно принципу максимального правдоподобия на основе мягкой схемы принятия решений. В стандартных системах, включающих модуляцию и кодирование, обычно принято отдельно описывать и реализовать детектор и декодер. Однако в системах TCM эти функции должны рассматриваться совместно. Можно добиться эффективного кодирования, не жертвуя скоростью передачи данных или не увеличивая ни ширину полосы частот, ни мощность [6, 31]. Вначале может показаться, что это утверждение нарушает некоторые основные принципы компромисса между мощностью или шириной полосы частот и вероятностью ошибки. Отметим, что компромисс здесь все же присутствует, поскольку TCM позволяет достичь эффективности кодирования за счет усложнения декодера.

При решетчатом кодировании набор сигналов многоуровневой/фазовой модуляции комбинируется со *схемой решетчатого кодирования*. Термин "схема решетчатого кодирования" применим к любой кодовой системе, которая обладает памятью (конечный автомат), такой например, как сверточный код. Сигналы многоуровневой/фазовой модуляции имеют совокупности, содержащие множественные амплитуды, множественные фазы или комбинации этих амплитуд и фаз. Иными словами, набор сигналов TCM наилучшим образом представляется любым набором сигналов (более чем двоичным), векторное представление которого может быть отображено на плоскости, подобной показанной на рис. 9.16, *а* для сигналов QAM. Схема решетчатого кодирования — это схема, которую можно охарактеризовать (решетчатой) диаграммой состояния, подобной решетчатым диаграммам, описывающим сверточные коды. Отметим, что хотя сверточные коды, представленные в главе 7, линейны, в общем случае решетчатые коды линейными быть не обязаны. Эффективность кодирования можно получить с помощью блочных или решетчатых кодов, однако здесь будут рассматриваться только решетчатые коды, поскольку наличие *алгоритма декодирования Витерби* делает решетчатое кодирование простым и эффективным. Унгербоек (Ungerboeck) показал, что при наличии шума AWGN схема TCM довольно просто может дать суммарную эффективность кодирования порядка 3 дБ по сравнению с некодированной системой, а при увеличении сложности можно получить эффективность порядка 6 дБ.

9.10.1. Истоки решетчатого кодирования

При TCM канальное кодирование и модуляция осуществляются вместе; невозможно просто определить, где начинается одно и заканчивается другое. Что же могло натолкнуть на разработку TCM? Возможно, все началось с мысли о том, что "не все подмножества сигналов (в совокупности) имеют равные пространственные свойства". Другими словами, для неортогонального множества сигналов, такого как MPSK, антиподные сигналы будут иметь наилучшие пространственные характеристики с точки зрения различения сигналов, в то время как ближайшие соседние сигналы будут иметь относительно плохие пространственные характеристики. Возможно, изначально идея кодовой модуляции возникла именно при попытке использовать эти различия.

Понять общие задачи TCM может помочь простая аналогия. Пусть в передатчике есть всезнающий волшебник. Как только биты сообщения попадают в систему, волшебник обнаруживает, что некоторые биты наиболее уязвимы к искажению, вызываемому каналом; следовательно, им присваиваются модулирующие сигналы, имеющие наилучшие пространственные характеристики. Подобным образом другие биты признаются весьма устойчивыми, поэтому они передаются с использованием сигналов с худшими пространственными характеристиками. Модуляция и кодирование осуществляются одновременно. Волшебник присваивает сигналы битам (модуляция), однако присвоение выполняется согласно критерию лучших или худших пространственных характеристик (канальное кодирование).

9.10.1.1. Увеличение избыточности сигнала

Схему TCM можно реализовать с помощью сверточного кодера, где k текущих битов и $K-1$ предыдущих битов используются для получения $n=k+p$ кодовых битов, где K — длина кодового ограничения кодера (см. главу 7), а p — число битов четности. Отметим, что кодирование увеличивает размер множества сигнала с 2^k до 2^{k+p}. Унгербоек [31] исследовал повышение пропускной способности, достигаемое благодаря увеличению набора сигналов, и пришел к заключению, что максимальную эффективность кодирования при обычной многоуровневой модуляции без кодирования можно реализовать, удваивая передаваемый некодированный набор ($p=1$). Этого можно достичь путем кодирования со степенью $k/(k+1)$ с последующим отображением групп из $(k+1)$ бит в набор из 2^{k+1} сигналов. На рис. 9.21, *а* показан набор сигналов, модулированных 4-PAM, до и после кодирования кодом со степенью кодирования 2/3 (после кодирования получаются 8-ричные сигналы PAM). Аналогично на рис. 9.21, *б* показан набор сигналов с 4-ричной модуляцией PSK (QPSK) до и после перекодирования кодом со степенью кодирования 2/3 в 8-ричные сигналы PSK. Подобным образом на рис. 9.21, *в* показаны некодированные 16-ричные сигналы QAM до и после перекодирования кодом со степенью кодирования 4/5 в 32-ричные сигналы QAM. В каждом из случаев, показанных на рис. 9.21, система сконфигурирована таким образом, чтобы до и после кодирования средняя мощность сигнала была одинаковой. Кроме того, для обеспечения необходимой избыточности при кодировании размер набора сигналов увеличивается с $M=2^k$ до $M'=2^{k+1}$. Таким образом, $M'=2M$; однако увеличение размера алфавита не приводит к увеличению требуемой ширины полосы частот. Напомним из раздела 9.7.2, что ширина полосы пропускания при неортогональной передаче сигнала не зависит от плотности точек сигналов в совокупности; она зависит только от скорости передачи сигнала. Расширенный набор сигнала *приводит* к уменьшению расстояния между соседними точками символов (для наборов

сигналов с постоянной средней мощностью), как видно из рис. 9.21. В *некодированной* системе такое уменьшение расстояния снижает достоверность передачи. Однако вследствие избыточности, вносимой кодом, это уменьшение расстояния уже не сильно влияет на вероятность ошибки. Напротив, достоверность определяет *просвет* — минимальное расстояние между членами набора *разрешенных* кодовых последовательностей. Просвет описывает "наиболее простой способ" совершения ошибки декодером (см. раздел 9.10.3.1). Независимо от используемого кода, пространство сигналов — это не самое удобное место для изучения улучшения достоверности, которое можно получить за счет кодирования. Это объясняется тем, что код определяется правилами и ограничениями, которые не видны в пространстве сигналов. Когда два сигнала находятся в непосредственной близости друг от друга в сигнальном пространстве кодовой системы, их близость может и не иметь существенного значения (с точки зрения вероятности ошибки), поскольку правила кода могут не допускать перехода между двумя такими якобы уязвимыми точками сигналов. Что же нужно для определения допустимых кодовых последовательностей и пространственных характеристик? Решетчатые диаграммы! При их использовании задача TCM — присвоение сигналов переходам в решетке, чтобы увеличить просвет между теми сигналами, которые вероятнее всего могут быть спутаны.

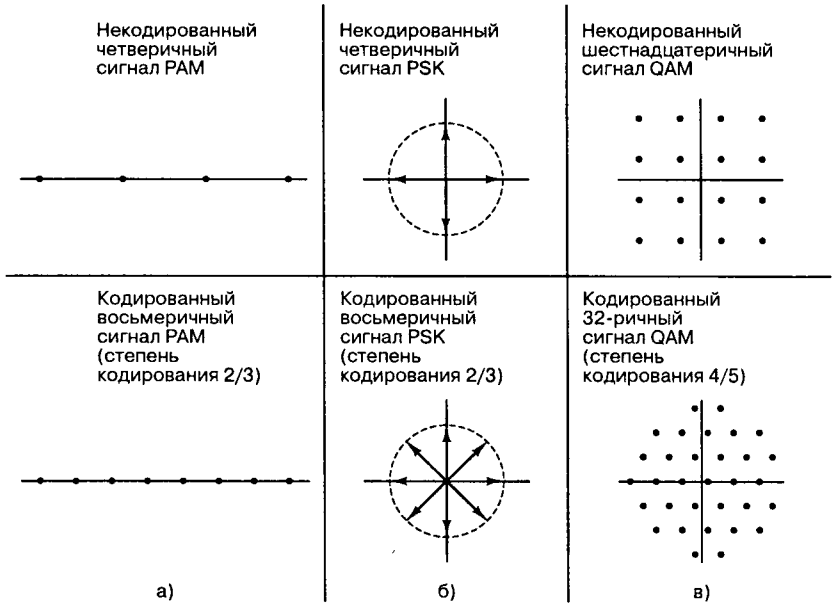


Рис. 9.21. Увеличение размера множества сигнала для решетчатого кодирования

9.10.2. Кодирование TCM

9.10.2.1. Разбиение Унгербоека

Пусть приемник использует мягкую схему принятия решений, так что подходящей будет евклидова метрика расстояния. Для максимизации просвета (измеряемого по Евклиду) Унгербоек [31] предложил отображение кода в сигнал, следующее из последова-

тельного разбиения совокупности модулирующих сигналов на подмножества с возрастающими минимальными расстояниями $d_0 < d_1 < d_2 \ldots$ между элементами подмножеств. Эта идея продемонстрирована на рис. 9.22 для сигнального множества 8-PSK. На рис. 9.22 исходная совокупность сигнала обозначена через A_0, а отдельные сигналы последовательно пронумерованы от 0 до 7. Если средняя мощность сигнала (квадрат амплитуды) выбрана равной единице, то расстояние d_0 между любыми двумя соседними сигналами, очевидно, равно $2 \sin(\pi/8) = 0{,}765$. На первом уровне разбиения получаются подмножества B_0 и B_1, где расстояние между соседними сигналами равно $d_1 = \sqrt{2}$. На следующем уровне образуются подмножества с C_0 по C_3, где расстояние между соседними сигналами равно уже $d_2 = 2$. Структуру простых кодов (до восьми состояний) можно определить эвристически. В первую очередь выбирается подходящая решетчатая структура, что можно сделать, не задумываясь о конкретном кодере. TCM относится к классу методов кодирования формой сигнала, поскольку для описания этой концепции требуется только подходящая решетка и набор модулирующих сигналов; даже не нужно вводить понятие битов. Сигналы из расширенного множества $M' = 2^{k+1}$ сигналов присваиваются переходам в решетке таким образом, чтобы максимизировать просвет (напомним, используется евклидово расстояние). При рассмотрении сверточных кодов в главе 7, переходы в решетке кодера (отражающие поведение цепи кодирования) помечались кодовыми битами. Для схемы TCM переходы в решетке помечаются модулирующими сигналами. Некодированный набор сигналов 4-PSK будет служить эталоном для кодированного набора 8-PSK. Этот эталонный набор, как показано на рис. 9.23, имеет тривиальную решетчатую диаграмму с одним состоянием и четырьмя параллельными переходами. Эта решетка тривиальна, поскольку решетка с одним состоянием означает, что в системе отсутствует память. Нет никаких ограничений или препятствий, чтобы в течение любого промежутка времени могли быть переданы сигналы 4-PSK; поэтому для такого некодированного случая оптимальный детектор просто независимо принимает ближайшие решения для каждого полученного зашумленного сигнала 4-PSK.

9.10.2.2. Отображение сигналов на переходы решетки

Унгербоек разработал эвристический набор правил [31] присвоения сигналам ветвей переходов решетки для получения эффективности кодирования, который позволяет сделать адекватный выбор состояний решетки. Правила построения решетки и разбиения множества сигнала (для модуляции 8-PSK) можно кратко изложить следующим образом.

1. Если за один интервал модуляции кодируется k бит, решетка должна разрешать для каждого состояния 2^k возможных перехода в последующее состояние.
2. Между парой состояний может существовать более одного перехода.
3. Все сигналы должны появляться с равной частотой и обладать высокой регулярностью и симметрией.
4. Переходы с одинаковым исходным состоянием присваиваются сигналам либо из подмножества B_0, либо B_1 — их смешение недопустимо.
5. Переходы с одинаковым конечным состоянием присваиваются сигналам либо из подмножества B_0, либо B_1 — их смешение недопустимо.
6. Параллельные переходы присваиваются сигналам либо из подмножества C_0, либо C_1, либо C_2, либо C_3 — их смешение недопустимо.

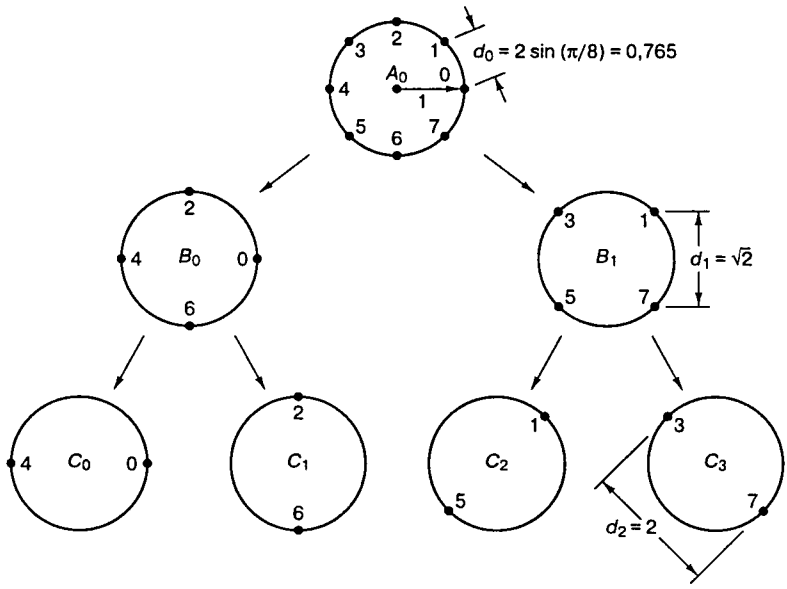


Рис. 9.22. Разбиение Унгербоека набора сигналов 8-PSK

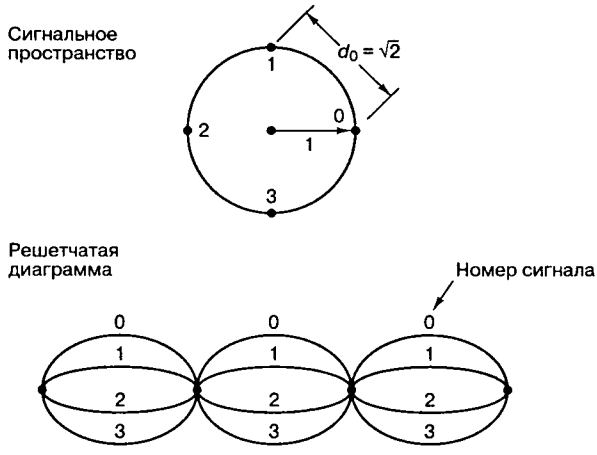


Рис. 9.23. Некодированное множество сигналов 4-PSK и его решетчатая диаграмма с одним состоянием

Правила гарантируют, что код, построенный таким образом, будет иметь регулярную структуру и просвет, всегда превышающий минимальное расстояние между точками сигнала исходной некодированной модуляции. На рис. 9.24 показано возможное отображение кода в сигнал с использованием решетки с четырьмя состояниями с параллельными путями. Присвоение сигналу кода производится посредством изучения разбитого пространства сигналов (рис. 9.22), решетчатой диаграммы, показанной на рис. 9.24, и правил, перечисленных выше. На переходах решетки написаны номера сигналов, присвоенных этим переходам согласно правилам разбиения. Отметим, что для модуляции 8-PSK присвоение сигнала осуществлялось согласно правилу 1: имеется $k + 1 = 3$ кодовых бита, следовательно

$k = 2$ информационных бита, а на входе и выходе каждого состояния имеется $2^2 = 4$ перехода. Присвоение сигналов осуществлялось согласно правилу 6, поскольку каждой паре параллельных переходов был присвоен сигнал одного из наборов C_0, C_1, C_2 или C_3. Кроме того, присвоение согласуется с правилами 4 и 5, поскольку четырем ветвям, выходящим в состояние (или покидающим состояние), были присвоены сигналы из набора B_0 или B_1. На рис. 9.24 состояния решетки различаются согласно типам сигналов, которые могут появиться на переходах, покидающих это состояние. Таким образом, состояния можно обозначить с помощью подмножеств сигнала как состояние C_0C_1 или C_2C_3 либо (другой возможный способ обозначения с помощью номеров сигнала) как состояние 0426, 1537 и т.д. На рис. 9.24 показаны обе системы обозначений. Из этого присвоения модулирующих сигналов переходам в решетке согласно правилам разбиения следует спецификация решетчатого *кодера*. Отметим, что окончательное присвоение битов кода сигналу (отображение кодового слова в переход) можно теперь выполнить произвольно. Хотя может показаться несколько странным, что теперь можно безнаказанно присваивать биты переходам в решетке и сигналам, стоит напомнить, что схемы кодера еще не существует. Следовательно, еще нет битов и переходы в решетке могут иметь только тот смысл, который для них выберем мы. Каковы же последствия такого *произвольного* присвоения? Выбор различных отображений кодовых слов в переходы отразится на структуре кодера. Следовательно, если повезет, будет реализована схема кодера, выходные биты которого будут соответствовать способу, которым осуществлялось их присваивание переходам между состояниями. В противном случае такое конструктивное решение реализовать будет сложно. При некотором выборе способа присвоения кодовых слов конструкция кодера будет проще, в то время как другой выбор может обусловить громоздкость его конструкции.

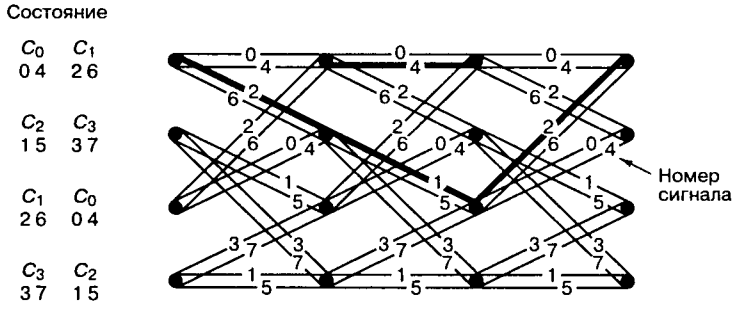


Рис. 9.24. Решетка с четырьмя состояниями с параллельными путями

Решетка, аналогичная показанной на рис. 9.24, вскоре будет исследована в контексте обнаружения и декодирования, чтобы проверить, обеспечивается ли эффективность кодирования при учете в процессе кодирования правил Унгербоека.

9.10.3. Декодирование TCM

9.10.3.1. Ошибочное событие и просвет

Задача сверточного декодера заключается в определении пути, пройденного сообщением в кодирующей решетке. Если все входящие последовательности сообщений равновероятны, декодером с минимальной вероятностью появления ошибки будет декодер, сравнивающий условные вероятности $P(\mathbf{Z}|\mathbf{U}^{(m)})$ (где $\mathbf{Z}$ — полученная последовательность сигналов,

а $\mathbf{U}^{(m)}$ — одна из возможных переданных последовательностей сигналов) и выбирающий максимальную. Этот критерий принятия решений, известный как *критерий максимального правдоподобия*, описан в разделе 7.3.1. Нахождение последовательности $\mathbf{U}^{(m')}$, которая максимизирует $P(\mathbf{Z}|\mathbf{U}^{(m)})$, эквивалентно нахождению последовательности $\mathbf{U}^{(m')}$, которая наиболее похожа на $\mathbf{Z}$. Поскольку декодер, работающий по принципу максимального правдоподобия, выберет такой путь по решетке, которому будет соответствовать последовательность $\mathbf{U}^{(m')}$, находящаяся на минимальном расстоянии от полученной последовательности $\mathbf{Z}$, задача определения максимального правдоподобия будет идентична задаче нахождения самого короткого расстояния по решетчатой диаграмме.

Поскольку сверточный код — это групповой (или линейный) код, набор расстояний, которые нужно проверить, не зависит от того, какая последовательность выбрана в качестве проверочной. Вследствие этого, не теряя общности, в качестве проверочной можно выбрать последовательность, целиком состоящую из нулей, показанную на рис. 9.25 пунктирной линией. В предположении, что была передана нулевая последовательность, ошибочное событие определяется как отклонение от нулевого пути с последующим возвратом на этот путь. Ошибочные события начинаются и заканчиваются состоянием a и не возвращаются в это состояние нигде в промежуточной области. На рис. 9.25 показано ошибочное сообщение в решетчатом коде, т.е. на рисунке изображена переданная нулевая последовательность, помеченная как $\mathbf{U} = \ldots, U_1, U_2, U_3, \ldots$, и альтернативная последовательность, помеченная как $\mathbf{V} = \ldots, V_1, V_2, V_3, \ldots$. Видно, что альтернативная последовательность сначала отклоняется, а затем снова сливается с переданной последовательностью. Если предположить, что осуществляется мягкое декодирование, сообщение принимается ошибочно тогда, когда полученные символы ближе (евклидово расстояние) к некоторой возможной последовательности $\mathbf{V}$, чем к реальной переданной последовательности $\mathbf{U}$. Из этого следует, что коды для сигналов многоуровневой/фазовой модуляции должны строиться таким образом, чтобы достигать максимального евклидова просвета; чем больше просвет, тем меньше вероятность ошибки. Следовательно, присвоение сигналов переходам решетки в кодере таким образом, чтобы максимизировать евклидов просвет (см. раздел 9.10.2), — это ключ к оптимизации решетчатых кодов.

9.10.3.2. Эффективность кодирования

Рассмотрим мягкую схему принятия решений, декодирование по принципу максимального правдоподобия, единичную среднюю мощность сигнала и гауссово распределение шума с дисперсией σ^2 на размерность. В этом случае нижний предел вероятности ошибочного события можно выразить через просвет d_f [32].

$$P_e \geq Q\left(\frac{d_f}{2\sigma}\right), \tag{9.55}$$

где $Q(\cdot)$ — гауссов интеграл ошибок, определенный в формуле (3.43). Использование термина "ошибочное событие" (error event) вместо "битовая ошибка" (bit-error) объясняется тем, что ошибка может распространяться на более чем один бит. При большом значении отношения сигнал/шум (signal-to-noise ratio — SNR) предел в уравнении (9.55) асимптотически точен. *Асимптотическая эффективность кодирования* G в децибелах относительно некоторой некодированной эталонной системы с аналогичными средней мощностью сигнала и дисперсией шума выражается как отношение расстояний или квадратов расстояний и записывается в следующем виде.

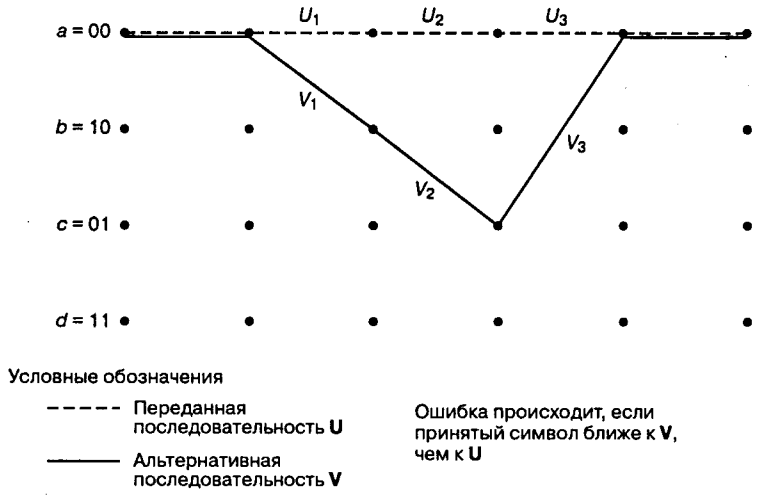


Рис. 9.25. Пример ошибочного события

$$G(\text{дБ}) = 20 \times \lg\left(\frac{d_f}{d_{\text{эт}}}\right) \text{ или } G(\text{дБ}) = 10 \times \lg\left(\frac{d_f^2}{d_{\text{эт}}^2}\right), \tag{9.56}$$

где d_f и $d_{\text{эт}}$ — евклидов просвет кодированной системы и некодированной эталонной системы. Отметим, что для больших значений SNR и данной вероятности появления ошибки формула (9.56) дает те же результаты, что и выражение для эффективности кодирования (6.19), повторно приведенное ниже.

$$G(\text{дБ}) = \left(\frac{E_b}{N_0}\right)_u (\text{дБ}) - \left(\frac{E_b}{N_0}\right)_c (\text{дБ}) \tag{9.57}$$

Здесь $(E_b/N_0)_u$ и $(E_b/N_0)_c$ являются требуемыми E_b/N_0 (в децибелах) для некодированной и кодированной систем. Следует помнить, что эффективность кодирования, выраженная в виде (9.56), дает ту же информацию (при больших значениях SNR), что и более привычное выражение для повышения достоверности передачи (9.57). По сути, формула (9.56) резюмирует основную задачу кода TCM. Эта задача — добиться просвета, превышающего минимальное расстояние между некодированными модулирующими сигналами (при той же скорости передачи информации, ширине полосы частот и мощности).

9.10.3.3. Эффективность кодирования для схемы 8-PSK при использовании решетки с четырьмя состояниями

Вычислим теперь эффективность кодирования для решетки с четырьмя состояниями в схеме 8-PSK, разработанной согласно правилам кодирования из раздела 9.10.2.2. Решетка на рис. 9.24 теперь будет исследоваться в контексте процедуры декодирования. Сначала в качестве настроечной выбирается нулевая последовательность. Иными словами, предполагается, что передатчик отправил последовательность, содержащую только копии сигнала номер 0. Чтобы продемонстрировать преимущества такой системы TCM (используя алгоритм декодирования Витерби), нужно пока-

зать, что самый простой способ совершения ошибки в кодированной системе сложнее самого простого способа совершения ошибки в некодированной системе. Необходимо изучить всевозможные отклонения от верного пути с последующим слиянием с верным путем (нулевой последовательностью) и найти тот, который имеет минимальное евклидово расстояние до правильного пути. Рассмотрим сначала возможный путь ошибочного события (рис. 9.24), который затемнен и помечен номерами сигнала 2, 1, 2. Квадрат расстояния до нулевого пути вычисляется как сумма квадратов отдельных расстояний между сигналами 2 и 0; 1 и 0; и 2 и 0. Отдельные расстояния берутся из диаграммы разбиения на рис. 9.22, в результате чего получаем следующее.

$$d^2 = d_1^2 + d_0^2 + d_1^2 = 2 + 0{,}585 + 2 = 4{,}585$$

или

$$d = \sqrt{4{,}585} = 2{,}2 \tag{9.58}$$

В уравнении (9.58) евклидово расстояние d получается точно так же, как и результирующий вектор в евклидовом пространстве, т.е. как квадратный корень из суммы квадратов отдельных компонентов (расстояний). На рис. 9.24 есть путь с отклонением и повторным слиянием, который имеет евклидово расстояние, меньшее $d = 2{,}2$. Это затененное ошибочное событие (помеченное как сигнал 4) происходит, если (при использовании декодирования Витерби) вместо правильного пути, связанного с сигналом 0, выживает параллельный. Может возникнуть вопрос: если декодер выбирает параллельный путь (т.е. последующее состояние одинаково в обоих случаях), будет ли это в действительности серьезной ошибкой. Если параллельный путь — это неправильно выбранный путь (это все-таки путь с отклонением и повторным слиянием, даже если он занимает только один промежуток времени), то позже, когда будут введены схемы кодеров и биты, выживший сигнал 4 даст в результате неверное значение бита. Расстояние от пути сигнала 4 до пути сигнала 0 равно, как видно из рис. 9.22, $d = 2$. Это расстояние меньше, чем расстояние для любого другого ошибочного события (можете проверить!); поэтому евклидов просвет для этой кодированной системы равен $d_f = 2$. Минимальное евклидово расстояние для набора некодированных эталонных сигналов на рис. 9.23 равно $d_{эт} = \sqrt{2}$. Теперь для вычисления асимптотической эффективности кодирования следует воспользоваться уравнением (9.56), что даст следующее.

$$G(\text{дБ}) = 10 \lg\left(\frac{d_f^2}{d_{эт}^2}\right) = 10 \lg\left(\frac{4}{2}\right) = 3 \text{ (дБ)} \tag{9.59}$$

9.10.4. Другие решетчатые коды

9.10.4.1. Параллельные пути

Если число состояний меньше размера набора кодированных сигналов M', решетчатая диаграмма требует параллельных путей. Следовательно, решетка с четырьмя состояниями для модуляции 8-PSK требует наличия параллельных путей. Чтобы лучше понять причины этого, обратимся еще раз к первому правилу Унгербоека: *если за один интервал модуляции кодируется k бит, решетка должна разрешать для каждого состояния 2^k возможных перехода в последующее состояние.* Для рассматриваемого случая 8-PSK каждый сигнал представляет $k + 1 = 3$ кодовых бит или $k = 2$ бит данных. Поэтому

из первого правила следует наличие $2^k = 2^2 = 4$ переходов в каждое последующее состояние. На первый взгляд решетка с четырьмя состояниями без параллельных путей может удовлетворить такому условию, если реализовать полностью замкнутую решетку (каждое состояние связано со всеми последующими состояниями). Однако попробуйте нарисовать полностью замкнутую решетку с четырьмя состояниями без параллельных путей, удовлетворяя при этом правилам 4 и 5 для системы 8-PSK. Это невозможно! Нарушение правил приведет к результатам, близким к оптимальным. В следующем разделе показана решетка с восемью состояниями для схемы 8-PSK (количество состояний уже не меньше M'), где могут быть соблюдены все правила разбиения без требования наличия параллельных путей.

9.10.4.2. Решетка с восемью состояниями

После экспериментирования с использованием различных структур решетки и присвоением канальных сигналов, в качестве оптимального для восьми состояний был выбран код 8-PSK, показанный на рис. 9.26 [31]. Путь ошибочного события с минимальным расстоянием до нулевого пути помечен номерами сигналов 6, 7, 6. Поскольку здесь отсутствуют параллельные пути, ограничивающие евклидов просвет, квадрат этого просвета равен $d_f^2 = d_1^2 + d_0^2 + d_1^2 = 4{,}585$, где расстояния d_0 и d_1 получены из рис. 9.22. Асимптотическая эффективность кодирования системы TCM с восемью состояниями относительно эталонной системы 4-PSK равна следующему.

$$G(\text{дБ}) = 10 \times \lg \frac{(d_1^2 + d_0^2 + d_1^2)_{\text{кодированная 8-PSK}}}{(d_{\text{эт}}^2)_{\text{некодированная 4-PSK}}} = 10 \times \lg\left(\frac{4{,}585}{2}\right) = 3{,}6\ (\text{дБ}) \qquad (9.60)$$

Подобным образом можно показать, что решетчатая структура с шестнадцатью состояниями для кодированной совокупности 8-PSK дает эффективность кодирования 4,1 дБ, по сравнению с некодированной схемой 4-PSK [31]. Если состояний меньше восьми, дополнительная эффективность кодирования может быть получена путем введения асимметрии в совокупность модулирующих сигналов [33].

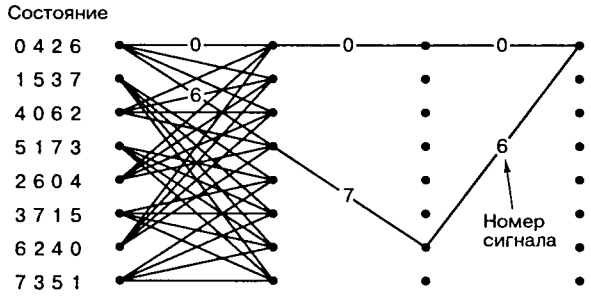


Рис. 9.26. Решетчатая диаграмма с восьмью состояниями для кода 8-PSK

9.10.4.3. Решетчатое кодирование для схемы QAM

Метод разбиения набора сигналов можно применять и к другим типам модуляции. Рассмотрим использование кодированной схемы 16-QAM с тремя информационными

битами на интервал модуляции, где в качестве эталонной системы выбрана некодированная 8-PSK. Для нормированного пространства 16-QAM выберем среднее значение квадрата амплитуды набора сигналов, равное единице, что дает $d_0 = 2/\sqrt{10}$. На рис. 9.27 показано разбиение сигналов 16-QAM на подмножества с возрастающими расстояниями между элементами ($d_0 < d_1 < d_2 < d_3$). Кодовая система 16-QAM с восемью состояниями, полученная путем разбиения набора согласно описанной ранее процедуре, показана на рис. 9.28 [31]. Путь ошибочной комбинации с минимальным расстоянием обозначен как D_6, D_5, D_2. Хотя при использовании схемы TCM имеется эффективность кодирования, при декодировании расширенного пространства сигнала существует потенциальная неопределенность фазы, которая может серьезно ухудшить достоверность передачи. Вей (Wei) [34] применил концепцию дифференциального кодирования к методам TCM; полученные при этом коды не зависят от поворотов элементарных сигналов на углы 90°, 180° и 270°.

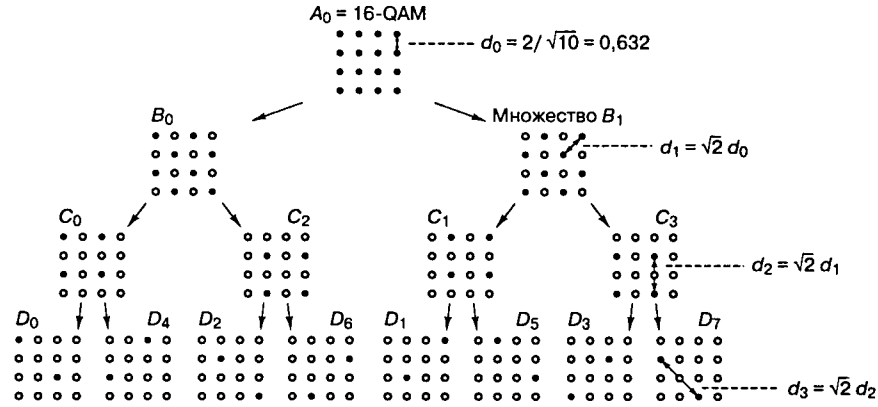


Рис. 9.27. Разбиение Унгербоека сигналов 16-QAM

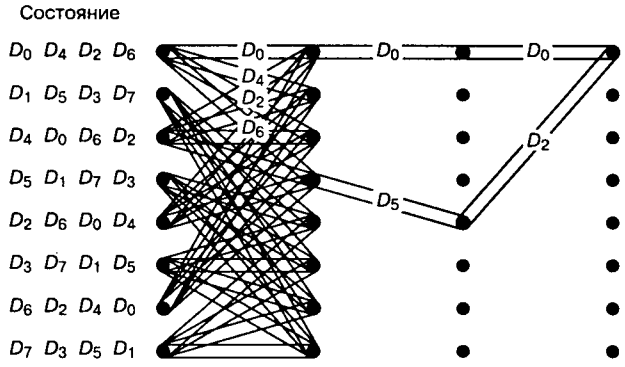


Рис. 9.28. Решетчатая диаграмма с восемью состояниями для передачи сигнала 16-QAM

Вкратце можно сказать, что решетчатое кодирование в узкополосных каналах включает больший алфавит сигналов (т.е. M-арные схемы PAM, PSK или QAM) для компенсации избыточности, которая вводится при кодировании; таким образом, ши-

рина полосы частот канала не возрастает. Даже если увеличение размера набора сигналов уменьшает минимальное расстояние между сигналами, евклидов просвет между *разрешенными* кодовыми последовательностями *превышает* величину, необходимую для компенсации этого уменьшения. В результате полная эффективность кодирования равна от 3 до 6 дБ без какого-либо расширения полосы частот [6, 31]. В следующем разделе эти идеи будут дополнительно проиллюстрированы на примере.

9.10.5. Пример решетчатого кодирования

В предыдущем разделе обсуждалось отображение сигналов в переходы решетки безотносительно к конечному отображению канальных символов (кодовых битов или кодовых слов) в переходы решетки. В этом разделе пример решетчатого кодирования начнется с рассмотрения точного определения структуры кодера. Структура кодера автоматически определяет решетчатую диаграмму и присвоение кодовых слов переходам решетки. Следовательно, в этом примере, если сигналы присвоены переходам решетки (а значит, подразумевающимся кодовым словам), уже нет возможности произвольно присваивать кодовые слова сигналам, как это делалось ранее при отсутствии схемы кодера.

Рассмотрим кодер, использующий сверточный код со степенью кодирования 2/3 для передачи двух бит информации за один интервал модуляции. Пример подобного кодера показан на рис. 9.29. Степень кодирования 2/3 достигается путем передачи без изменения одного бита из каждой пары битов исходной последовательности и кодирования второго бита двумя кодовыми битами (выполняется кодером со степенью кодирования 1/2 и длиной кодового ограничения $K = 3$). Как показано на рисунке, биты из входящей последовательности попадают в сдвиговый регистр только через один (m_2, m_4, ...). Может возникнуть вопрос: насколько может быть хорошей такая система, если преимущества, определяемые избыточностью, получают только 50% бит. Напомним пример с волшебником, который определял, что некоторые биты довольно уязвимы и поэтому они присваивались модулирующим сигналам с наилучшими пространственными характеристиками, в то время как другие считались устойчивыми и присваивались сигналам с худшими пространственными характеристиками. Модуляция и кодирование происходят одновременно; якобы "некодированные" не будут забыты, они выиграют от присвоения наилучших сигналов. Следует подчеркнуть, что кодирование и декодирование в схеме TCM происходит преимущественно на сигнальном уровне (в нашем первом описании TCM о каком-либо кодере не упоминалось), тогда как в традиционном коде с исправлением ошибок кодирование и декодирование происходит только на битовом уровне.

Решетчатая диаграмма на рис. 9.30 описывает схему кодера с рис. 9.29. Как и в главе 7, названия состояний соответствуют содержимому крайних правых $K - 1 = 2$ разрядов регистра сдвига. Параллельные переходы на решетке (рис. 9.30) обусловлены некодированными битами; некодированный бит представляется крайним левым битом каждого перехода решетки. В каждом состоянии начинается четыре перехода. Для каждого состояния имеется два верхних перехода — от пары входных информационных битов (m_1m_2 равны 00 и 10); два нижних перехода проистекают от пары 01 и 11. На рис. 9.30 показана решетчатая структура, подобная показанной на рис. 9.24, за исключением того, что каждый переход на рис. 9.30 обозначен назначенным ему кодовым словом. Стоит повторить, что схема кодера определяет, какие кодовые слова появляются на переходах решетки; разработчик системы только присваивает сигналы переходам. Следовательно, когда уже имеется схема (поведение которой описывается

решеткой), любой сигнал, присвоенный переходу в решетке, автоматически становится носителем кодового слова, которое соответствует этому переходу.

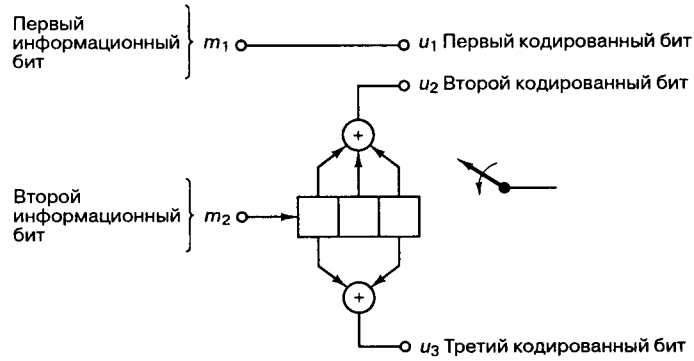


Рис. 9.29. Сверточный кодер со степенью кодирования 2/3

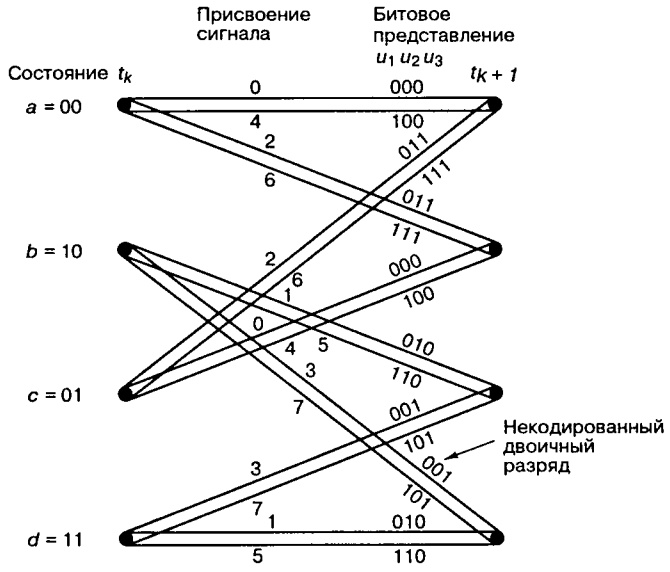


Рис. 9.30. Решетчатая диаграмма для кода со степенью кодирования 2/3

Пусть кодовая модуляция — это 8-ричная амплитудно-импульсная модуляция (8-ary pulse amplitude modulation — 8-PAM), как показано на рис. 9.31. На рис. 9.31, *а* показан кодированный набор сигналов, где для каждого сигнала евклидово расстояние до центра пространства сигналов показано в некоторых произвольных единицах, причем сигналы расположены на равных расстояниях один от другого и симметрично относительно нуля. На рис. 9.31, *б* показан эталонный набор 4-ричной схемы PAM, в котором точки сигнала и расстояния помечены аналогичным образом. Важным этапом в разработке кодера является присвоение 8-ричных сигналов PAM переходам ре-

шетки согласно правилам разбиения Унгербоека (рис. 9.32). Изучение этих правил может привести к такому же присвоению номеров сигналов переходам решетки, как показано на рис. 9.24. Подобное присвоение сигналов, а также кодовые слова, присвоенные схемой кодера, показаны на рис. 9.30. Наиболее несопоставимая пара сигналов (с расстоянием $d_2 = 8$) была присвоена наиболее уязвимым (в плане появления ошибок) параллельным переходам. Кроме того, как следует из правил Унгербоека, сигналы со следующим наибольшим расстоянием ($d_1 = 4$) были присвоены переходам, выходящим или входящим в одно и то же состояние. Для удобства на рис. 9.31, *а* показано также присвоение кодовых слов сигналам (результат отображения сигналов в переходы решетки).

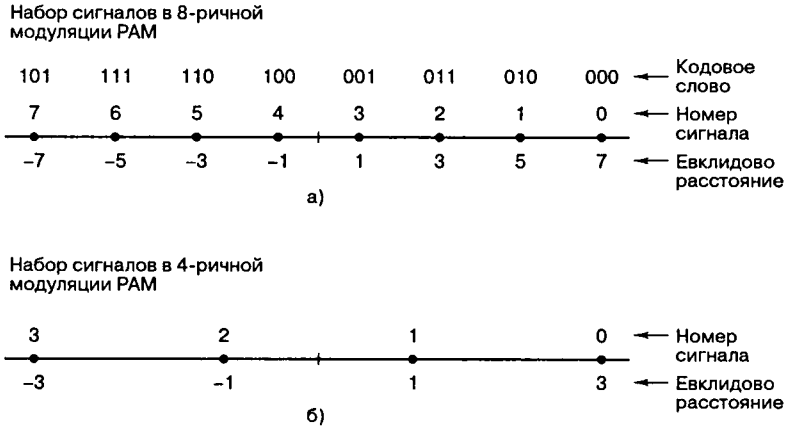


Рис. 9.31. Множества сигналов: а) кодированная 8-ричная PAM; б) некодированная 4-ричная PAM

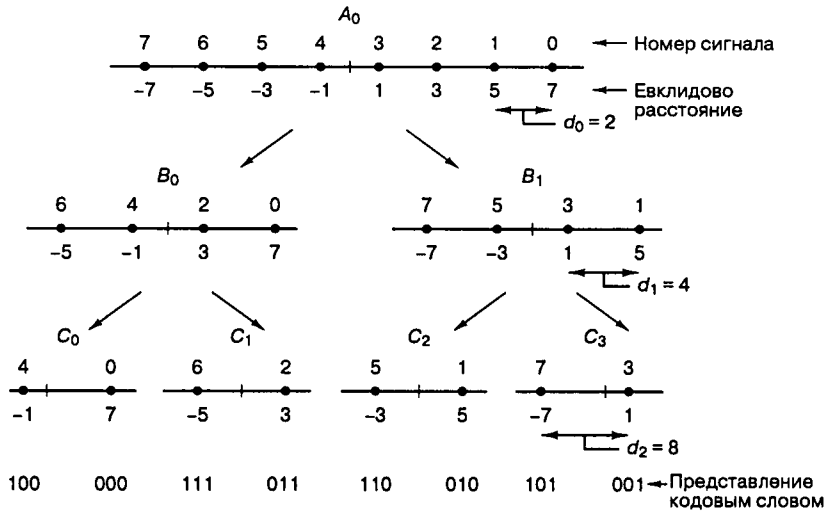


Рис. 9.32. Разбиение Унгербоека сигналов 8-PAM

На рис. 9.24 путь ошибочного события, помеченный номерами сигналов 2, 1, 2, — это путь с минимальным расстоянием для нашего примера модуляции 8-PAM. Расстояние до нулевого пути вычисляется с использованием формулы (9.58). В этом примере, если взять отдельные расстояния с рис. 9.32, d_f вычисляется следующим образом.

$$d^2 = d_1^2 + d_0^2 + d_1^2 = 16 + 4 + 16 = 36$$

или

$$d_f = 6 \tag{9.61}$$

Можно легко убедиться, что для такого типа модуляции параллельный путь (с $d = 8$) не будет ошибочным путем с минимальным расстоянием (как это было для 8-PSK). Далее для нахождения эталонного расстояния для 4-PAM из рис. 9.31, *б* находим, что $d_{эт} = 2$. Теперь для этого примера можем вычислить асимптотическую эффективность кодирования, сравнивая квадрат евклидова просвета кодированной системы с евклидовым просветом эталонной системы. Однако тут необходимо убедиться в том, что средняя мощность сигналов в каждом наборе одинакова. В предыдущем примере схемы 8-PSK выбор единичной окружности для кодированной и некодированной систем означал, что средняя мощность сигнала была одинакова в обоих наборах. Однако в этом примере ситуация несколько иная. Следовательно, для вычисления асимптотической эффективности кодирования требуется нормировать следствие неравенства средней мощности набора сигналов, т.е. видоизменить выражение (9.56) [35]. Соответственно записываем

$$G(\text{дБ}) = 10 \times \lg\left(\frac{d_f^2 / S_{ср}}{d_{эт}^2 / S'_{ср}}\right), \tag{9.62}$$

где $S_{ср}$ и $S'_{ср}$ — средняя мощность сигналов в кодированном и эталонном наборах. Расстояние соответствует амплитуде сигнала или напряжению; таким образом, квадрат расстояния соответствует квадрату напряжения, или мощности. Следовательно, средняя мощность сигнала из совокупности вычисляется как

$$S_{ср} = \frac{d_1^2 + d_2^2 + \ldots + d_M^2}{M}, \tag{9.63}$$

где d_i — евклидово расстояние от центра пространства до i-го сигнала, а M — количество кодовых символов в этом множестве. Для набора сигналов 8-PAM, показанного на рис. 9.31, *а*, уравнение (9.63) дает значение $S_{ср} = 21$. Для эталонного набора сигналов 4-PAM, показанного на рис. 9.31, *б*, уравнение (9.63) дает значение $S'_{ср} = 5$.

При использовании уравнения (9.62) асимптотическая эффективность кодирования для системы 8-PAM будет иметь следующий вид.

$$G(\text{дБ}) = 10 \times \lg\left(\frac{36/21}{4/5}\right) = 3{,}3\ (\text{дБ}) \tag{9.64}$$

Увеличивая количество состояний решетки (большая длина кодового ограничения) за счет возрастающей сложности декодирования, можно добиться большей эффективности кодирования. При кодировании сигналов 8-PAM со степенью кодирования 2/3 решетка с 256 состояниями даст эффективность кодирования, на 5,83 дБ большую от-

носительно набора сигналов 4-PAM [9]. В этом случае вследствие использования решетчатого кодирования будет иметь место только незначительное увеличение сложности передатчика. Задача декодирования в приемнике становится более сложной, однако использование больших интегральных схем (large scale integrated — LSI, БИС) и сверхскоростных интегральных схем (high-speed integrated circuit — VHSIC, ССИС) делает такой метод кодирования чрезвычайно привлекательным для достижения значительной эффективности кодирования без расширения полосы пропускания.

9.10.6. Многомерное решетчатое кодирование

В разделе 9.9.3 подчеркивалось, что при данной скорости передачи данных передача сигналов в двухмерном пространстве может давать ту же достоверность, что и передача в одномерном пространстве PAM, но при меньшей средней мощности. Это достигается путем выбора точек сигналов на двухмерной решетке из области с кольцевой, а не прямоугольной границей. Выполняя подобное при более высоких размерностях, можем видеть, что потенциальная экономия энергии приближается к 1,53 дБ при N, стремящемся к бесконечности. В реальных системах при такой многомерной передаче сигналов можно достичь экономии энергии (*эффективность выбора формы*) порядка 1 дБ относительно одномерной передачи [21, 36, 37]. В стандарте высокоскоростных модемов V.34 определена 16-мерная модуляция QAM; используемый метод отображения битов в точки пространства высшей размерности называется *отображением оболочки* (shell mapping); соответствующая эффективность выбора формы равна 0,8 дБ [16]. Используя четырех-, восьми- и шестнадцатимерную совокупности сигналов, можно получить некоторые преимущества по сравнению с обычными двухмерными схемами — меньшие двухмерные блоки совокупности, повышение устойчивости к неопределенности фазы, более выгодные компромиссы между эффективностью кодирования и сложностью реализации. Множество подобных систем представлено и охарактеризовано в работе [36]. (Читателям, заинтересованным в дальнейшем изучении кодовой модуляции, в частности решетчатого кодирования, рекомендуется обратиться к работам [38–46].)

9.11. Резюме

В этой главе объединены некоторые вопросы модуляции и кодирования, рассмотренные в предыдущих главах. Здесь пересмотрены основные задачи разработки системы: получение максимальной скорости передачи информации при одновременном снижении вероятности возникновения ошибки и значения E_b/N_0, сужении полосы пропускания и уменьшении сложности. Компромиссы были изучены эвристически в двух плоскостях: вероятность появления ошибки и эффективность использования полосы частот. Первая явно иллюстрирует компромисс между E_b/N_0 и P_B, плюс неявно отображает расход полосы пропускания. На второй показан компромисс между R/W и E_b/N_0 при неявном изображении поведения P_B. Кроме того, в этой главе описаны типичные шаги, которые предпринимаются при удовлетворении требований к полосе пропускания, мощности и вероятности появления ошибок в системе цифровой связи. Здесь также рассматриваются некоторые ограничения, которые делают невозможным неограниченное повышение производительности. Согласно критерию Найквиста, полосу пропускания нельзя сужать бесконечно. Существует теоретический предел; для передачи R_s символов/с без межсимвольной интерференции нужно задействовать, как минимум, $R_s/2$ Гц полосы пропускания. Теорема Шеннона-Хартли связана с компро-

миссом между мощностью и полосой пропускания, а также определяет другое важное ограничение — предел Шеннона. Предел Шеннона, равный –1,6 дБ, — это минимальное теоретически возможное значение E_b/N_0, которое (совместно с канальным кодированием) необходимо для получения сколь угодно низкой вероятности возникновения ошибки в канале AWGN. Более общим ограничением является значение пропускной способности канала, превышение которой автоматически запрещает безошибочную передачу сигналов. В этой главе также изучены некоторые схемы модуляции с эффективным использованием полосы пропускания, такие как манипуляция с минимальным сдвигом (minimum shift keying — MSK), квадратурная амплитудная модуляция (quadrature amplitude modulation — QAM) и решетчатое кодирование. Последний метод позволяет достичь эффективного кодирования без потерь в полосе пропускания.

Литература

1. *IEEE Personal Communications.* Special Issue on Software Radio, vol. 6, n. 4, August, 1999.
2. Nyquist H. *Certain Topic on Telegraph Transmission Theory.* Trans. AIEE, vol. 47, April, 1928, pp. 617–644.
3. Shannon C. E. *A Mathematical Theory of Communication.* BSTJ, vol. 27, 1948, pp. 379–423, 623–657.
4. Shannon C. E. *Communication in the Presence of Noise.* Proc. IRE, vol. 37, n. 1, January, 1949, pp. 10–21.
5. Bedrosian E. *Spectrum Conservation by Efficient Channel Utilization.* Rand Corp., Report WN-9275-ARPA, Contract DAHC-15–73–C–0181, Santa Monica, California, October, 1975.
6. Ungerboeck G. *Trellis-Coded Modulation with Redundant Signal Sets.* Part I and Part II. IEEE Comunications Magazine, vol. 25, February, 1987, pp. 5–21.
7. Hodges M. R. L. *The GSM Radio Interface.* British Telecom Tech. J., vol. 8, n. 1, January, 1990, pp. 31–43.
8. Anderson J. B. and Sundberg C-E. W. *Advances in Constant Envelope Coded Modulation.* IEEE Commun., Mag., vol. 29, n. 12, December, 1991, pp. 36–45.
9. Clark G. C. Jr. and Cain J. B. *Error Correction Coding for Digital Communications.* Plenum Press, New York, 1981.
10. Lindsey W. C. and Simon M. K. *Telecommunication Systems Engineering.* Prentice-Hall, Englwood Cliffs, NJ, 1973.
11. Sklar B. *Defining, Designing, and Evaluating Digital Communication Systems.* IEEE Commun. Mag., vol. 31, n. 11, November, 1993, pp. 92–101.
12. Korn I. *Digital Communications.* Van Nostrand Reinhold Co., New York, 1985.
13. Viterbi A. J. *Principles of Coherent Communications.* McGraw-Hill Book Co., New York, 1966.
14. Lin S. and Costello D. J., Jr. *Error Control Coding: Fundamental and Applications.* Prentice-Hall, Englwood Cliffs, NJ, 1983.
15. Odenwalder J. P. *Error Control Coding Handbook.* Linkabit Corporation, San Diego, California, July, 15, 1976.
16. Forney G. D., Jr., et. al. *The V.34 High-Speed Modem Standard.* IEEE Communications Magazine. December, 1996.
17. Pasupathy S. *Minimum Shift Keying: A Spectrally Efficient Modulation.* IEEE Commun. Mag., July, 1979, pp.14–22.
18. Gronemeyer S. A. and McBride A. L. *MSK and Offset QPSK Modulation.* IEEE Trans. Commun., vol. COM-24, August, 1976, pp. 809–820.
19. Simon M. K. *A Generalization of Minimum Shift Keying (MSK) Type Signaling Based Upon Input Data Symbol Pulse Shaping.* IEEE Trans. Commun., vol. COM-24, August, 1976, pp. 845–857.
20. Leib H. and Pasupathy S. *Inherent Error Control Properties of Minimum Shift Keying.* IEEE Communications Mag., vol. 31, n. 1, January, 1993, pp. 52–61.

21. Forney G. D. Jr. et. al. *Efficient Modulation for Bandlimited Channels.* IEEE J. Selectd Areas in Commun., vol. SAC-2, n. 5, September, 1984, pp. 632–647.
22. Thomas C. M., Weidner M. Y. and Durrani S. H. *Digital Amplitude-Phase Keying with M-ary Alphabets.* IEEE Trans. Commun., vol. COM-22, n. 2, February, 1974, pp. 168–180.
23. Lucky R. W. and Hancock J. C. *On the Optimum Performance of N-ary Systems Having Two Degrees of Freedom.* IRE Trans. on Commun. Sys., vol. CS-10, June, 1962, pp. 185–192.
24. Campopiano C. N. and Glazer B. G. *A Coherent Digital Amplitude and Phase Modulation Scheme.* IRE Trans. on Commun. Sys., vol. CS-10, June, 1962, pp. 90–95.
25. Cahn C. R. *Combined Digital Phase and Amplitude Modulation Communication Systems.* IRE Trans. on Commun. Tech., September, 1960.
26. Foschini G. J. and Gitlin R. D. *Optimization of Two Dimensional Signal Constellations in the Presence of Gaussian Noise.* IEEE Trans. Commun., vol. COM-22, n. 1, January, 1974, pp. 23–38.
27. Welti G. R. and Jhong S. L. *Digital Transmission with Coherent Four-Dimensional Modulation.* IEEE Trans. Inform. Theory, vol. IT-20, n. 4, July, 1974, pp. 497–502.
28. Gersho A. and Lawrence V. B. *Multidimensional Signal Constellations for Voice-band Data Transmission.* IEEE J. Selected Areas in Commun., vol. SAC-2, n. 5, September, 1984, pp. 687–702.
29. Zetterberg L. H. and Brandstrom H. *Codes for Combined Phase and Amplitude Modulated Signals in a Four-Dimensional Space.* IEEE Trans. Commun., vol. COM-25, n. 9, September, 1977, pp. 943–950.
30. Wilon S. G., Sleeper H. A. and Stinath N. K. *Four-Dimensional Modulation and Coding: An Alternative to Frequency Reuse.* IEEE 1984 Int'l. Commun. Conf., pp. 919–923.
31. Ungerboeck G. *Channel Coding with Multilevel/Phase Signals.* IEEE Trans. Inform. Theory, vol. IT-28, January, 1982, pp. 55–67.
32. Forney G. D. *The Veterbi Algorithm.* Proceedings of the IEEE, vol. 61, n. 3, March, 1978, pp. 268–278.
33. Divsalar D., Simon M. K. and Yuen J. H. *Trellis Coding with Asymmetric Modulations.* IEEE Trans. Commun., vol. COM-35, n. 2, February, 1987.
34. Wei J.-F. *Rotationally Invariant Convolutional Channel Coding with Expanded Signal Space — Parts I and II.* IEEE J. Sel. Areas Commun., vol. SAC-2, no. 5, September, 1984, pp. 659–686.
35. Thapar H. K. *Real-Time Application of Trellis Coding to Highspeed Voiceband Data Transmission.* IEEE J. Sel. Areas Commun., vol. SAC-2, n. 5, September, 1984, pp. 648–658.
36. Wei J.-F. *Trellis-Coded Modulation with Multidimensional Constellations.* IEEE Trans. Inforantion Theory, vol. IT-33, n. 4, July, 1987, pp. 483–501.
37. Tretter S. A. *An Eight-Dimensional 64-State Trellis code for Transmitting 4 Bits Per 2-D Symbol.* IEEE J. on Sel. Areas of Commun., vol. 7, n. 9, December, 1989, pp. 1392–1395.
38. Kato S., Morikura M. and Kubota S. *Implementation of Coded Modems.* IEEE Communications Magazine, vol. 29, n. 12, December, 1991, pp. 88–97.
39. *Special Issue on Coded Modulation.* IEEE Communication Magazine, vol. 29, n. 12, December, 1991.
40. Biglieri E., et. al. *Introduction to Trellis-Coded Modulation with Applicatin.* MacMillan, New York, NY, 1991.
41. Edbauer F. *Performance of Interleaved Trellis-Code Differential 8-PSK Modulation over Fading Channels.* IEEE J. on Selected Areas in Commun., vol. 7, n. 9, December, 1989, pp. 1340–1346.
42. Rimoldi B. *Design of Coded CPFSK Modulation Systems for Bandwidth and Energy Efficiency.* IEEE Transactions on Communications, vol. 37, n. 9, September, 1989, pp. 897–905.
43. Viterbi A. J., et. al. *A Pragmatic Approach to Trellis-Coded Modulation.* IEEE Communications Magazine, vol. 27, n. 7, July, 1989, pp. 11-19.
44. Divsalar D. and Simon M. K. *The Design of Trellis Coded MPSK for Fading Channels: Performance Criteria.* IEEE Trans. on Comm., vol. 36, n. 9, September, 1988, pp. 1004–1012.
45. Divsalar D. and Simon M. K. *The Design of Trellis Coded MPSK for Fading Channels: Set Partitioning for Optimum Code Design.* IEEE Trans. on Comm., vol. 36, n. 9, September, 1988, pp. 1013–1021.
46. Divsalar D. and Simon M. K. *Multiple Trellis Coded Modulation (MTCM).* IEEE Trans. on Commun., vol. 36, n. 4, April, 1988, pp. 410–419.

Задачи

9.1. Рассмотрим телефонный канал связи с полосой пропускания 3 кГц. Пусть данный канал можно смоделировать как канал AWGN.

а) Чему равна пропускная способность такой схемы, если SNR равно 30 дБ?

б) Какое минимальное значение SNR требуется для получения скорости передачи данных 4800 бит/с?

в) Повторить расчеты п. б для скорости передачи информации 19 200 бит/с.

9.2. Рассмотрим передачу по телефонному каналу потока данных со скоростью 100 Кбит/с (при полосе пропускания 3 кГц). Можно ли получить безошибочную передачу при SNR, равном 10 дБ? Ответ обоснуйте. Если это невозможно, предложите модификацию системы, которая бы это позволила.

9.3. Рассмотрим источник, который производит шесть сообщений с вероятностями 1/2, 1/4, 1/8, 1/16, 1/32 и 1/32. Определите среднее информационное содержание сообщения (в битах).

9.4. Данный исходный алфавит состоит из 300 слов, из которых 15 появляются с вероятностью 0,06 каждое, а остальные 285 слов — с вероятностью 0,00035 каждое. Если в секунду передается 1000 слов, то какова средняя скорость передачи информации?

9.5. а) Найдите среднюю пропускную способность (в битах за секунду), которая требуется для передачи черно-белого телевизионного сигнала высокого разрешения со скоростью 32 кадра в секунду, если каждый кадр состоит из 2×10^6 элементов изображения и 16 градаций уровня яркости. Все элементы изображения считаются независимыми, и все уровни яркости появляются с одинаковой вероятностью.

б) Для цветного телевидения в описанной выше системе дополнительно вводится 64 оттенка цвета. Какая дополнительная пропускная способность потребуется в цветной системе по сравнению с черно-белой?

в) Определите требуемую пропускную способность, если 100 возможных комбинаций цвета и яркости появляются с вероятностью 0,003 каждая, 300 комбинаций — с вероятностью 0,001 и 624 комбинации — с вероятностью 0,00064.

9.6. Докажите, что энтропия максимальна, когда все выходы источника имеют равную вероятность.

9.7. Рассчитайте неопределенность или неоднозначность сообщения в битах на знак для текстовой передачи с использованием 7-битового кода ASCII. Считайте, что все знаки равновероятны и что вследствие шума в канале вероятность ошибки равна 0,01.

9.8. Предполагается, что линия связи с некогерентной FSK имеет максимальную скорость передачи данных 2,4 Кбит/с без ISI в канале с номинальной полосой пропускания 2,4 кГц. Предложите способы повышения скорости передачи данных при следующих системных ограничениях.

а) Ограничена мощность.

б) Ограничена полоса пропускания.

в) Одновременно ограничены и полоса пропускания, и мощность.

9.9. В табл. 39.1 описаны четыре разные линии связи "спутник/наземный терминал". Для каждой линии связи потери в пространстве составляют 196 дБ, резерв — 0 дБ, случайные потери отсутствуют. Для каждой линии связи укажите рабочую точку на плоскости эффективности использования полосы частот, зависимости R/W от E_b/N_0 и охарактеризуйте линию согласно одному из следующих описаний: ограниченная полоса пропускания, строго ограниченная полоса пропускания, ограниченная мощность и строго ограниченная мощность. Ответ обоснуйте.

Таблица 39.1. Пропускная способность линии связи для четырех спутниковых линий связи

Спутник	Принимающий терминал	Максимальная скорость передачи данных
INTELSAT IV EIRP = 22,5 дБВт Полоса пропускания = 36 МГц	Большая стационарная станция, диаметр антенны = 30 м G/T = 40,7 дБ/К	165 Мбит/с
DSCS II EIRP = 28 дБВт Полоса пропускания = 50 МГц	Корабль, диаметр антенны = 4 фута G/T = 10 дБ/К	100 Кбит/с
DSCS II EIRP = 28 дБВт Полоса пропускания = 50 МГц	Большая стационарная станция, диаметр антенны = 60 футов G/T = 39 дБ/К	72 Мбит/с
GAPSAT/MARISAT EIRP = 28 дБВт Полоса пропускания = 500 кГц	Самолет, коэффициент усиления антенны = 0 дБ G/T = –30 дБ/К	500 бит/с

9.10. Нужно выбрать модуляцию и код коррекции ошибок для системы связи реального времени, работающей с каналом AWGN при доступной полосе пропускания 2400 Гц. E_b/N_0 = 14 дБ. Требуемая скорость передачи информации и вероятность битовой ошибки равны 9600 бит/с и 10^{-5}. Выбирать можно из двух типов модуляции — некогерентные ортогональные 8-FSK или 16-QAM при обнаружении с использованием согласованных фильтров. При выборе кода также возможны две альтернативы — код БХЧ (127, 92) или сверточный код со степенью кодирования 1/2, дающие эффективность кодирования 5 дБ при вероятности битовой ошибки 10^{-5}. Предполагая идеальную фильтрацию, докажите, что сделанный выбор удовлетворяет желаемым требованиям относительно полосы пропускания и вероятности ошибки.

9.11. В условиях задачи 9.10 полоса пропускания расширена до 40 кГц, а доступное E_b/N_0 снижено до 7,3 дБ. Выберите подходящие схемы модуляции и кодирования и докажите, что сделанный выбор удовлетворяет желаемым требованиям относительно полосы пропускания и вероятности ошибки.

9.12. В условиях задачи 9.10 в канале AWGN теперь возможно исчезновение сигнала, которое длится до 1000 мс. Доступная полоса пропускания равна 3400 Гц, а E_b/N_0 равно 10 дБ. Помимо выбора схем модуляции и кодирования теперь требуется разработать устройство чередования (см. раздел 8.2) для борьбы с проблемой исчезновения сигнала. Возможны две альтернативы — блочное устройство чередования 16 × 32 и сверточное 150 × 300. Докажите, что сделанный выбор удовлетворяет желаемым требованиям относительно полосы пропускания и вероятности ошибки, и продумайте способ борьбы с более длительными исчезновениями сигнала.

9.13. а) Рассмотрим систему связи реального времени, работающую с каналом AWGN, в которой применяется модуляция 8-PSK и код Грея. Выберите код коррекции ошибок, который сможет дать вероятность ошибки в декодированном бите не больше 10^{-7}, если принимаемое P_r/N_0 равно 70 дБГц, а скорость передачи информации равна 1 Мбит/с. Выбирать можно из следующих кодов: расширенный код Голея (24, 12), код БХЧ (127, 64) или код БХЧ (127, 36). Передаточные функции этих кодов показаны на рис. 6.21. Для облегчения процесса выбора считайте, что $P_B = 10^{-7}$, а передаточная функция пересекается с осью абсцисс в таких точках: для кода (24, 12) — в точке 3×10^{-3}, для кода (127, 64) — в точке $1{,}3 \times 10^{-2}$, для кода (127, 36) — в точке 3×10^{-2}.

б) С помощью внешнего вида передаточной функции кода можно интуитивно представить, какой код является лучшим при установленных технических требованиях. Совпадает ли ваш конечный выбор с первоначальной гипотезой? Не удивил ли вас ответ на п. а? Объясните полученные результаты в контексте двух механизмов, которые проявляются при использовании кодирования с коррекцией ошибок в системе связи реального времени.

в) Какую эффективность кодирования в децибелах обеспечивает код, выбранный вами в п. а?

9.14. Рассмотрим спутниковую систему связи реального времени, работающую с каналом AWGN (возмущаемую периодическим исчезновением сигнала). Вся линия связи описывается следующими требованиями для мобильного передатчика и спутникового приемника на низкой околоземной орбите.

Скорость передачи данных $R = 9600$ бит/с

Доступная полоса пропускания $W = 3000$ Гц

Энергетический резерв линии связи $M = 0$ дБ (см. раздел 5.6)

Несущая частота $f_c = 1{,}5$ ГГц

EIRP = 6 дБ

Расстояние между передатчиком и приемником $d = 1000$ км

Добротность спутникового приемника $G/T = 30$ дБ[i]

Температура принимающей антенны $T_A^\circ = 290K$

Потери в линии связи между принимающей антенной и приемником $L = 3$ дБ

Коэффициент шума приемника $F = 10$ дБ

Потери вследствие замирания $L_f = 20$ дБ

Прочие потери $L_o = 6$ дБ

Нужно так выбрать одну из двух схем модуляции (MPSK с применением кода Грея или некогерентную ортогональную MFSK), чтобы не было превышения имеющейся полосы пропускания и сохранялась мощность. Для кодирования с коррекцией ошибок выбирается один из кодов БХЧ (127, k), представленных в табл. 9.2, обеспечивающий наибольшую избыточность и при этом удовлетворяющий ограничениям на полосу пропускания. Рассчитайте вероятность появления ошибки в декодированном бите. Какая эффективность кодирования (если таковая имеется) соответствует предложенному выбору. *Подсказка*: параметры стоит вычислять в следующем порядке: E_b/N_0, E_s/N_0, $P_E(M)$, p_c, P_B. При использовании уравнения (9.41) для расчета декодированной вероятности появления битовой ошибки низкое значение E_b/N_0 вынуждает учитывать большое количество слагаемых в сумме. Следовательно, очень кстати будет помощь компьютера.

9.15. Требуется, чтобы система связи реального времени поддерживала скорость передачи данных 9600 бит/с при вероятности появления битовой ошибки, не превышающей 10^{-5} с полосой пропускания 2700 Гц. P_r/N_0 до детектирования составляет 54,8 дБГц. Выберите одну из двух схем модуляции (MPSK с применением кода Грея или некогерентную ортогональную MFSK) так, чтобы не было превышения имеющейся полосы пропускания и сохранялась мощность. Если необходимо применить кодирование с коррекцией ошибок, выбирайте самый простой (самый короткий) код из представленных в табл. 9.3, обеспечивающий требуемую достоверность передачи без превышения необходимой полосы пропускания. Докажите, что ваш выбор удовлетворяет системным требованиям.

9.16. а) При фиксированной вероятности появления ошибок покажите, что связь между размером алфавита M и требуемой средней мощностью для MPSK и QAM можно представить следующим образом.

$$\frac{\text{средняя мощность MPSK}}{\text{средняя мощность QAM}} \approx \frac{3M^2}{2(M-1)\pi^2}$$

б) Обсудите преимущества одного типа передачи сигналов перед другим.

9.17. Рассмотрим телефонный модем, работающий со скоростью 28,8 Кбит/с и использующий решетчатое кодирование QAM.

а) Рассчитайте эффективность использования полосы частот, считая, что полоса пропускания канала равна 3429 Гц.

б) Предполагая, что $E_b/N_0 = 10$ дБ и в канале присутствует шум AWGN, рассчитайте теоретическую доступную пропускную способность в полосе частот 3429 Гц.

в) Какое значение E_b/N_0 необходимо для получения в полосе 3429 Гц скорости передачи 28,8 Кбит/с?

9.18. На рис. 9.17 показано несколько совокупностей 16-ричных символов.

а) Для кольцевой совокупности (5, 11) рассчитайте минимальные радиальные расстояния r_1 и r_2, если минимальное расстояние между символами должно быть 1.

б) Рассчитайте среднюю мощность сигнала для кольцевой совокупности (5, 11) и сравните ее со средней мощностью квадратной совокупности 4×4 ($M = 16$) (при том же минимальном расстоянии между символами).

в) Почему квадратный набор может оказаться более практичным?

9.19. Рассмотрим систему решетчатого кодирования со степенью 2/3 из раздела 9.10.5, которая используется в двоичном симметричном канале (binary symmetric channel — BSC). Исходное состояние кодера предполагается равным 00. На выходе BSC принимается последовательность $\mathbf{Z} = (1\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 1$ остальные все $0)$.

а) Найдите максимально правдоподобный путь по решетчатой диаграмме и определите первые 6 декодированных информационных битов. Если появляется петля между двумя сливающимися путями, выбирайте верхнюю ветвь, входящую в определенное состояние.

б) Определите, были ли изменены в канале какие-либо биты $\mathbf{Z}$, и если это так, определите, какие именно.

в) Объясните, как вы решите задачу, если вместо канала BSC дан гауссов канал.

9.20. Найдите асимптотическую эффективность кодирования для схемы решетчатого кодирования (trellis-coded modulation — TCM) с 4 состояниями. Степень кодирования 2/3 получается с помощью кодера, конфигурация которого показана на рис. 9.29, где 50% информационных бит поданы на вход сверточного кодера со степенью кодирования 1/2, а оставшиеся 50% — непосредственно на выход. Кодовая модуляция — 8-PAM, как показано на верхней части рис. 9.31. Эталонным служит набор сигналов 4-PAM с амплитудами -16, -1, $+1$, $+16$. Не кажется ли вам, что полученный ответ не согласуется с теоремой Шеннона, которая предсказывает предел эффективности кодирования порядка 11–12 дБ? Будет ли кто-либо использовать эталонный набор, который был предложен здесь? Можно заметить, что эффективность кодирования для комбинированной схемы модуляции/кодирования слегка отличается от той, которая имеется в случае одного лишь кодирования. Объясните ваши результаты в этом контексте.

9.21. Найдите асимптотическую эффективность кодирования для схемы решетчатого кодирования с 8 состояниями. Кодовая модуляция — 8-PSK, а некодированный эталон — 4-PSK. Решетчатая структура между моментами t_k и t_{k+1} строится следующим образом: все состояния (от верхнего до нижнего) произвольно обозначаются от 1 до 8. Затем состояния 1, 3, 5 и 7 в момент t_k соединяются с состояниями 1–4 в момент t_{k+1}. Аналогично состояния 2, 4, 6 и 8 в момент t_k соединяются с состояниями 5–8 в момент t_{k+1}. Нарисуйте три секции (три интервала времени) решетчатой структуры. Сопоставьте ветви и сигналы и найдите кратчайший ошибочный путь.

Вопросы

9.1. Почему связь ширины полосы с эффективностью ее использования одинакова для ортогональных двоичной и четверичной частотных манипуляций (см. раздел 9.5.1)?

9.2. В схеме модуляции MPSK, *эффективность использования полосы частот* растет при увеличении размерности, а в схеме MFSK, наоборот, снижается. Объясните, почему так происходит (см. разделы 9.7.2 и 9.7.3).

9.3. Опишите преобразования скрытой энергии и скоростей при преобразовании информационных битов в канальные биты, затем — в символы и элементарные сигналы (см. раздел 9.7.7).

9.4. Резкое *увеличение боковых максимумов* в спектре MSK на рис. 9.15 показывает, почему схема MSK считается более спектрально эффективной, чем QPSK. Как в таком случае можно объяснить тот факт, что спектр QPSK имеет более *узкий основной максимум*, чем спектр MSK (см. раздел 9.8.2)?

9.5. В главе 4 было сказано, что двоичная фазовая манипуляция (binary phase shift keying — BPSK) и квадратичная фазовая манипуляция (quaternary phase shift keying — QPSK) имеют одинаковые соотношения для вероятности возникновения битовой ошибки (см. раздел 4.8.4). Можно ли утверждать то же самое для M-арной амплитудно-импульсной модуляции (M-ary pulse amplitude modulation — M-PAM) и M^2-арной квадратурной амплитудной модуляции (M^2-QAM), т.е. будут ли эти схемы иметь одинаковую вероятность возникновения битовой ошибки (см. раздел 9.8.3.1)?

9.6. Хотя схемы решетчатого кодирования (trellis-coded modulation — TCM) не требуют дополнительной полосы пропускания или мощности, в них все равно присутствует некоторый *компромисс*. За счет чего достигается эффективное кодирование в TCM (см. раздел 9.10)?

9.7. В чем смысл *состояния* в системе с конечным числом состояний (см. раздел 9.10)?

9.8. Какой *избыточности сигнала* при применении схемы TCM достаточно для получения выгод кодирования (снижение вероятности появления ошибки или повышение пропускной способности) (см. раздел 9.10.1.1)?

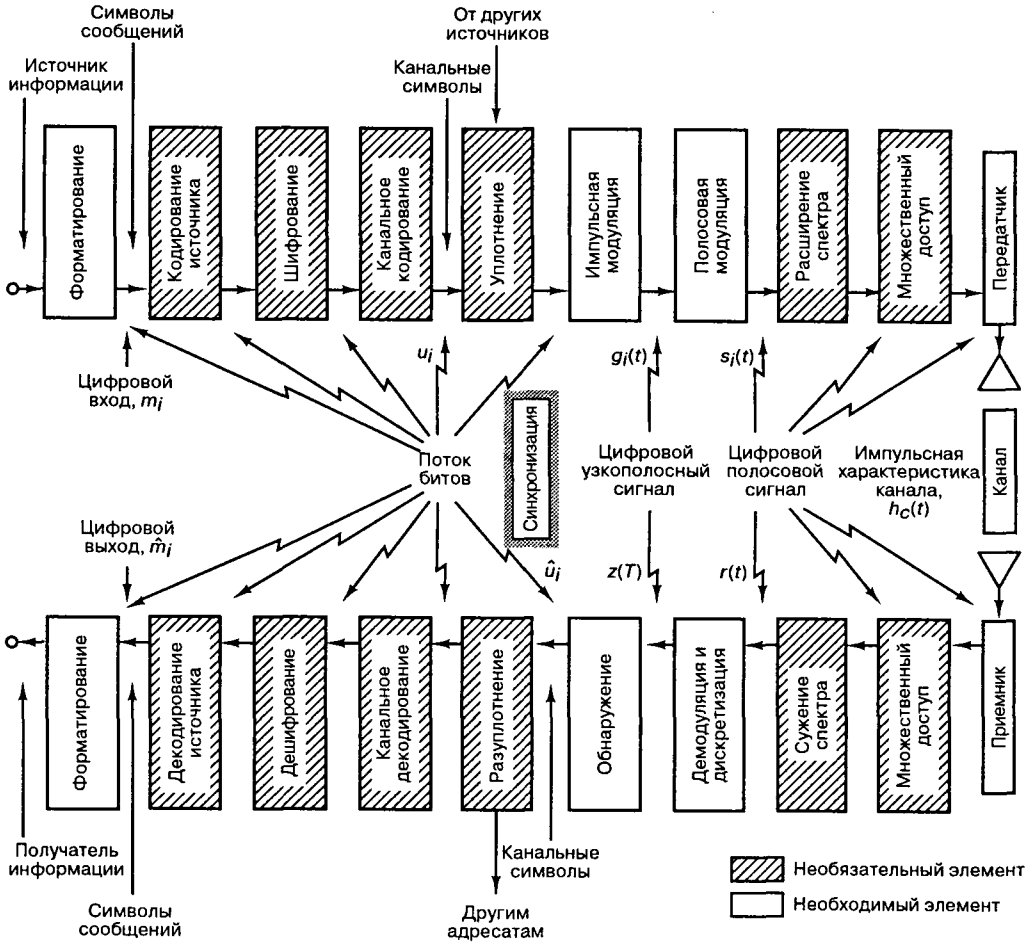
9.9. Для схем TCM дайте определение понятию *асимптотическая эффективность кодирования*, и из этого определения объясните, к чему нужно стремиться при построении кода TCM (см. раздел 9.10.3.2).

9.10. Когда на решетчатой диаграмме TCM нужны *параллельные пути* для удовлетворения правил разбиения Унгербоека? Чем грозит нарушение этих правил (см. раздел 9.10.4.1)?

ГЛАВА 10

Синхронизация

Морис Эй. Кинг мл. (Maurice A. King, Jr.)
The Aerospace Corporation
Эль Сегундо, Калифорния



10.1. Вступление

10.1.1. Виды синхронизации

Как правило, при рассмотрении производительности приемника или демодулятора предполагается наличие некоторого уровня синхронизации сигнала, хотя явно это высказывается не всегда. Например, при когерентной фазовой демодуляции (схема PSK) предполагается, что приемник может генерировать опорные сигналы, фаза которых идентична (возможно, с точностью до постоянного смещения) фазе элементов сигнального алфавита передатчика. Затем в процессе принятия решения относительно значения принятого символа (по принципу максимального правдоподобия) опорные сигналы сравниваются с поступающими.

При генерации подобных опорных сигналов приемник должен быть синхронизирован с принимаемой несущей. Это означает, что фаза поступающей несущей и ее копии в приемнике должны согласовываться. Другими словами, если в поступающей несущей не закодирована информация, поступающая несущая и ее копия в приемнике будут проходить через нуль одновременно. Этот процесс называется *фазовой автоподстройкой частоты* (это — условие, которое следует удовлетворить максимально близко, если в приемнике мы хотим точно демодулировать когерентно модулированные сигналы). В результате фазовой автоподстройки частоты местный гетеродин приемника синхронизируется по частоте и фазе с принятым сигналом. Если сигнал-носитель информации модулирует непосредственно не несущую, а поднесущую, требуется определить как фазу несущей, так и фазу поднесущей. Если передатчик не выполняет фазовой синхронизации несущей и поднесущей (обычно так и бывает), от приемника потребуется генерация копии поднесущей, причем управление фазой копии поднесущей производится отдельно от управления фазой копии несущей. Это позволяет приемнику получать фазовую синхронизацию как по несущей, так и по поднесущей.

Кроме того, предполагается, что приемник точно знает, где начинается поступающий символ и где он заканчивается. Эта информация необходима, чтобы знать соответствующий промежуток интегрирования символа — интервал интегрирования энергии перед принятием решения относительно значения символа. Очевидно, если приемник интегрирует по интервалу несоответствующей длины или по интервалу, захватывающему два символа, способность к принятию точного решения будет снижаться.

Можно видеть, что символьную и фазовую синхронизации объединяет то, что обе включают создание в приемнике копии части переданного сигнала. Для фазовой синхронизации это будет точная копия несущей. Для символьной — это меандр с переходом через нуль одновременно с переходом поступающего сигнала между символами. Говорят, что приемник, способный сделать это, имеет *символьную синхронизацию*. Поскольку на один период передачи символа обычно приходится очень большое число периодов несущей, этот второй уровень синхронизации значительно грубее фазовой синхронизации и обычно выполняется с помощью другой схемы, отличной от используемой при фазовой синхронизации.

Во многих системах связи требуется еще более высокий уровень синхронизации, который обычно называется *кадровой синхронизацией*. Кадровая синхронизация требуется, когда информация поставляется блоками, или сообщениями, содержащими фиксированное число символов. Это происходит, например, при использовании блочного кода для реализации схемы прямой защиты от ошибок или если канал связи имеет

временное разделение и используется несколькими пользователями (технология TDMA). При блочном кодировании декодер должен знать расположение границ между кодовыми словами, что необходимо для верного декодирования сообщения. При использовании канала с временным разделением нужно знать расположение границ между пользователями канала, что необходимо для верного направления информации. Подобно символьной синхронизации, кадровая равнозначна возможности генерации меандра на скорости передачи кадров с нулевыми переходами, совпадающими с переходами от одного кадра к другому.

Большинство систем цифровой связи, использующих когерентную модуляцию, требуют всех трех уровней синхронизации: фазовой, символьной и кадровой. Системы с некогерентной модуляцией обычно требуют только символьной и кадровой синхронизации; поскольку модуляция является некогерентной, точной синхронизации фазы не требуется. Кроме того, некогерентным системам необходима *частотная синхронизация*. Частотная синхронизация отличается от фазовой тем, что копия несущей, генерируемая приемником, может иметь произвольные сдвиги фазы от принятой несущей. Структуру приемника можно упростить, если не предъявлять требование относительно определения точного значения фазы поступающей несущей. К сожалению, как показано при обсуждении методов модуляции, это упрощение влечет за собой ухудшение зависимости достоверности передачи от отношения сигнал/шум. В следующем разделе будут рассмотрены различные относительные компромиссы, связанные с синхронизацией разных уровней, достоверностью передачи и универсальностью системы.

До настоящего момента в центре обсуждения находилась принимающая часть канала связи. Однако иногда передатчик играет более активную роль в синхронизации — он изменяет отчет времени и частоту своих передач, чтобы соответствовать ожиданиям приемника. Примером того является спутниковая сеть связи, где множество наземных терминалов направляют сигналы на единственный спутниковый приемник. В большинстве подобных случаев передатчик для определения точности синхронизации использует обратный канал связи от приемника. Следовательно, для успеха синхронизации передатчика часто требуется двусторонняя связь или сеть. По этой причине синхронизация передатчика часто называется *сетевой*. Этот тип синхронизации также будет рассмотрен далее в этой главе.

10.1.2. Плата за преимущества

Необходимость синхронизации приемника связана с определенными затратами. Каждый дополнительный уровень синхронизации подразумевает большую стоимость системы. Наиболее очевидное вложение денег — необходимость в дополнительном программном или аппаратном обеспечении для приемника, обеспечивающего получение и поддержание синхронизации. Кроме того, что менее очевидно, иногда мы платим временем, затраченным на синхронизацию до начала связи, или энергией, необходимой для передачи сигналов, которые будут использоваться в приемнике для получения и поддержания синхронизации. В данном случае может возникнуть вопрос, почему разработчик системы связи вообще должен рассматривать проект системы, требующий высокой степени синхронизации. Ответ: улучшенная производительность и универсальность.

Рассмотрим обычное коммерческое аналоговое АМ-радио, которое может быть важной частью системы широковещательной связи, включающей центральный передатчик и множество приемников. Данная система связи не синхронизирована. В то же время полоса пропускания приемника должна быть достаточно широкой, чтобы включать не только ин-

формационный сигнал, но и любые флуктуации несущей, возникающие вследствие эффекта Доплера или дрейфа опорной частоты передатчика. Это требование к полосе пропускания передатчика означает, что на детектор поступает дополнительная энергия шума, превышающая энергию, которая теоретически требуется для передачи информации. Несколько более сложные приемники, содержащие систему слежения за частотой несущей, могут включать узкий полосовой фильтр, центрированный на несущей, что позволит значительно снизить шумовую энергию и увеличить принятое отношение сигнал/шум. Следовательно, хотя обычные радиоприемники вполне подходят для приема сигналов от больших передатчиков на расстоянии несколько десятков километров, они могут оказаться недееспособными при менее качественных условиях.

Для цифровой связи компромиссы между производительностью и сложностью приемника часто рассматриваются при выборе модуляции. В число простейших цифровых приемников входят приемники, разработанные для использования с бинарной схемой FSK с некогерентным обнаружением. Единственные требования — битовая синхронизация и сопровождение частоты. Впрочем, если в качестве модуляции выбрать когерентную схему BPSK, то можно получить ту же вероятность битовой ошибки, но при меньшем отношении сигнал/шум (приблизительно на 4 дБ). Недостатком модуляции BPSK является то, что приемник требует точного отслеживания фазы, что может представлять сложную конструкторскую проблему, если сигналы обладают высокими доплеровскими скоростями[2] или для них характерно замирание (см. главу 15).

Еще один компромисс между ценой и производительностью затрагивает кодирование с коррекцией ошибок. В предыдущих главах было установлено, что при использовании подходящих методов защиты от ошибок возможно значительное улучшение производительности. В то же время цена, выраженная в сложности приемника, может быть высока. Для надлежащей работы блочного декодера требуется, чтобы приемник достигал блочной синхронизации, кадровой или синхронизации сообщений. Эта процедура является дополнением к обычной процедуре декодирования, хотя существуют определенные коды коррекции ошибок, имеющие встроенную блочную синхронизацию [1]. Сверточные коды также требуют некоторой дополнительной синхронизации для получения оптимальной производительности. Хотя при анализе производительности сверточных кодов часто делается предположение о бесконечной длине входной последовательности, на практике это не так. Поэтому для обеспечения минимальной вероятности ошибки декодер должен знать начальное состояние (обычно все нули), с которого начинается информационная последовательность, конечное состояние и время достижения конечного состояния. Знание момента окончания начального состояния и достижения конечного состояния эквивалентно наличию кадровой синхронизации. Кроме того, декодер должен знать, как сгруппировать символы канала для принятия решения при разветвлении. Это требование также относится к синхронизации.

Приведенное выше обсуждение компромиссов велось с точки зрения соотношения между производительностью и сложностью отдельных каналов и приемников. Стоит от-

---

[1]Отклонение частоты, воспринимаемой приемником, от частоты, переданной передатчиком, которое возникает вследствие относительного движения передатчика и приемника. Если пренебречь эффектами второго и более высоких порядков, смещение частоты Δf равно Vf_0/c, где V — относительная скорость (положительная, если расстояние между приемником и передатчиком сокращается), f_0 — номинальная частота, а c — скорость света.

[2]Скорость изменения доплеровского смещения частоты. Эта скорость накладывает ограничение на возможности системы слежения за частотой несущей.

метить, что способность синхронизировать также имеет значительные потенциальные последствия, связанные с эффективностью и универсальностью системы. Кадровая синхронизация позволяет использовать передовые, универсальные методы множественного доступа, подобные схемам множественного доступа с предоставлением каналов по требованию (demand assignment multiple access — DAMA). Кроме того, использование методов расширения спектра — как схем множественного доступа, так и схем подавления интерференции — требует высокого уровня синхронизации системы. (Методы расширения спектра подробно рассмотрены в главе 12.) Далее будет показано, что эти технологии предлагают возможность создания весьма разносторонних систем, что является очень важным свойством при изменении системы или при воздействии преднамеренных или непреднамеренных помех от различных внешних источников.

10.1.3. Подход и предположения

Со времени первого редактирования текста было сделано, по крайней мере, два значительных открытия в области синхронизации. Одно — использование методов работы с дискретными данными для обработки сигналов (в том числе — синхронизации). Другое — это публикация нескольких работ о синхронизации [2–4]. В данной главе мы не будем пытаться охватить весь материал, связанный с синхронизацией. Нашей задачей является выработка широкого интуитивного понимания данного вопроса, а не перечисление методов проектирования синхронизаторов. Следовательно, мы будем подразумевать использование традиционных аналоговых разработок, считая, что те же принципы применимы и к системам обработки дискретных данных, даже если реализация синхронизаторов будет отличаться. Схемы ФАПЧ коммерчески доступны в виде относительно небольших чипов или являются частью большего устройства обработки сигналов. Предполагается, что читатель, интересующийся современными реализациями описанных принципов, способен определить, как они применяются к дискретным данным.

10.2. Синхронизация приемника

Все системы цифровой связи требуют определенной синхронизации сигналов, поступающих в приемник. В данном разделе рассматриваются основы синхронизации различных уровней. Обсуждение начинается с рассмотрения основных уровней синхронизации, требуемых для когерентного приема, — частотной и фазовой — и краткого обсуждения структуры и принципов работы схем фазовой автоподстройки частоты (ФАПЧ). Затем рассматривается символьная синхронизация. В некоторой степени символьная синхронизация требуется всем цифровым операциям приема (когерентным и некогерентным). В заключение раздела описывается кадровая синхронизация приемника и методы ее получения и поддержания.

10.2.1. Частотная и фазовая синхронизация

Практически во всех схемах синхронизации имеется определенная разновидность контура фазовой автоподстройки частоты (ФАПЧ). В современных цифровых приемниках опознать этот контур может быть трудно, но его функциональный эквивалент присутствует практически всегда. Схематическая диаграмма основы контура ФАПЧ показана на рис. 10.1. Контуры ФАПЧ самоуправляемы, причем управляющим параметром является фаза локально генерируемой копии поступающего несущего сигнала. Контуры ФАПЧ состоят из трех основных компонентов: детектора фазы, контурного фильтра и генератора,

управляемого напряжением (ГУН). Детектор фазы — это устройство, измеряющее различия фаз поступающего сигнала и локальной копии. Если поступающий сигнал и его локальная копия изменяются относительно друг друга, то эта разность фаз (или рассогласование по фазе) — это зависимый от времени сигнал, поступающий на контурный фильтр. Контурный фильтр регулирует отклик контура ФАПЧ на эти изменения сигнала. Качественно спроектированный контур должен иметь возможность отслеживать изменения фазы поступающего сигнала и не должен быть чрезмерно восприимчив к шуму приемника. Генератор, управляемый напряжением, — это устройство, создающее копию несущей. Данный генератор, как можно догадаться из названия, является генератором синусоидального сигнала, частота которого управляется уровнем напряжения на входе устройства. На рис. 10.1 детектор фазы показан как умножитель, контурный фильтр описывается собственной импульсной характеристикой $f(t)$ и ее Фурье-образом $F(\omega)$ и также соответствующим образом обозначен генератор, управляемый напряжением.

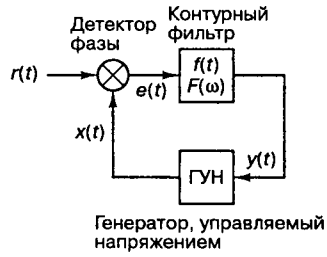


Рис. 10.1. Схема контура фазовой автоподстройки частоты

ГУН — это генератор, выходная частота которого является линейной функцией входного напряжения (в определенном рабочем диапазоне частот). Положительное входное напряжение приведет к тому, что выходная частота ГУН будет выше неуправляемого значения ω_0, тогда как отрицательное напряжение приведет к тому, что частота ГУН будет меньше этого значения. Синхронизация по фазе достигается путем подачи фильтрованной версии разности фаз (т.е. рассогласования по фазе) между входным сигналом $r(t)$ и выходным сигналом с ГУН $x(t)$, который возвращается на вход ГУН (на рис. 10.1 эта функция обозначена как $y(t)$).

Для современных цифровых приемников детектор рассогласования может быть сложнее математически, чем это показано на рис. 10.1. Например, детектор рассогласования может представлять собой набор корреляторов (согласованных фильтров), каждый из которых согласовывается с иным сдвигом фаз, с последующей подачей на вход ГУН взвешенной суммы (весовой функции) сигналов с выходов этих корреляторов. Выход весовой функции может представлять собой оценку рассогласования по фазе. Подобная функция может быть математически очень сложной, но ее легко аппроксимировать, используя современные цифровые технологии. Генератор, управляемый напряжением, не обязательно должен быть генератором синусоидального сигнала, он может быть реализован как постоянное запоминающее устройство, указатели которого определяют местный таймер и выход устройства оценки рассогласования по фазе. Контур обратной связи не обязательно должен быть непрерывным (как на рис. 10.1), а коррекция фазы может производиться только один раз на кадр или один раз на пакет, в зависимости от структуры сигнала. В информационный поток может вводиться специальный заголовок или известная последовательность символов, кото-

рые будут облегчать процесс синхронизации. И все же, несмотря на эти очевидные различия, основные элементы всех схем ФАПЧ сходны с показанными на рис. 10.1.

Рассмотрим нормированный входной сигнал следующего вида.

$$r(t) = \cos[\omega_0 t + \theta(t)] \tag{10.1}$$

Здесь ω_0 — номинальная несущая частота, а $\theta(t)$ — медленно меняющаяся фаза. Подобным образом рассмотрим нормированный выходной сигнал генератора, управляемого напряжением.

$$x(t) = -2\sin[\omega_0 t + \hat{\theta}(t)] \tag{10.2}$$

На выходе детектора фазы эти сигналы дадут выходной сигнал рассогласования следующего вида.

$$\begin{aligned} e(t) = x(t)r(t) &= 2\sin[\omega_0 t + \hat{\theta}(t)]\cos[\omega_0 t + \theta(t)] = \\ &= \sin[\theta(t) - \hat{\theta}(t)] + \sin[2\omega_0 t + \theta(t) + \hat{\theta}(t)] \end{aligned} \tag{10.3}$$

Пусть контурный фильтр является фильтром нижних частот, тогда второй член правой части выражения (10.3) будет отфильтрован и им можно пренебречь. (Предположение о фильтре нижних частот является разумным решением при проектировании контура.) Фильтр нижних частот дает сигнал рассогласования, являющийся функцией исключительно разности фаз между входом (формула (10.1)) и выходом ГУН (формула (10.2)). Это именно тот сигнал, который нам нужен. Выходная частота ГУН является производной по времени от аргумента синусоиды из уравнения (10.2). Если предположить, что ω_0 — это неуправляемая частота ГУН (частота на выходе при нулевом входном напряжении), отличие выходной частоты ГУН от ω_0 можно выразить как производную по времени от фазового члена $\hat{\theta}(t)$. Выходная частота ГУН является линейной функцией входного напряжения. Следовательно, поскольку выходное нулевое напряжение дает выходную частоту ω_0, отличие выходной частоты от ω_0 будет пропорционально значению выходного напряжения $y(t)$.

$$\begin{aligned} \Delta\omega(t) &= \frac{d}{dt}[\hat{\theta}(t)] = K_0 y(t) = \\ &= K_0 e(t) * f(t) \approx \\ &\approx K_0[\theta(t) - \hat{\theta}(t)] * f(t) \end{aligned} \tag{10.4}$$

Здесь $\Delta\omega(t)$ обозначает разность частот, знак * — свертку (см. приложение А), а при последнем преобразовании использовалось приближение малых углов (т.е. $e(t) = \sin[\theta(t) - \hat{\theta}(t)] \approx \theta(t) - \hat{\theta}(t)$). Приближение малых углов справедливо при малых значениях выходного рассогласования по фазе (контур близок к синхронизации по фазе). Все сказанное выше справедливо при нормально функционирующем контуре. Множитель K_0 — это усиление ГУН, а $f(t)$ — импульсная характеристика контурного фильтра. Данное линейное дифференциальное уравнение относительно $\hat{\theta}(t)$ (в котором использовано приближение малых углов) называется линеаризованным уравнением контура. Это, пожалуй, наиболее полезное соотношение при определении поведения контура при нормальной работе (когда мало рассогласование по фазе).

Пример 10.1. Линеаризованное уравнение контура

Покажите, что при надлежащем выборе K_0 и $f(t)$ линеаризованное уравнение контура (10.4) имеет тенденцию к синхронизации фазы, т.е. вне зависимости от начальных условий разность фаз между входным сигналом и выходом ГУН будет снижаться.

Решение

Пусть фаза входного сигнала $\theta(t)$ медленно меняется со временем. Можно видеть, что если разность фаз в правой части уравнения (10.4) положительна (т.е. $\theta(t) > \hat{\theta}(t)$), то надлежащим выбором K_0 и $f(t)$ производную по времени от $\hat{\theta}(t)$ можно сделать больше нуля, так что $\hat{\theta}(t)$ будет расти со временем, что приведет к снижению разности $|\theta(t) - \hat{\theta}(t)|$. С другой стороны, если разность фаз отрицательна, $\hat{\theta}(t)$ будет уменьшаться со временем, что также приведет к снижению разности фаз. И наконец, если $\theta(t) = \hat{\theta}(t)$, из уравнения (10.4) видно, что $\hat{\theta}(t)$ не будет меняться со временем и условие $\theta(t) = \hat{\theta}(t)$ будет выполняться всегда.

Рассмотрим преобразование Фурье уравнения (10.4).

$$i\omega\hat{\Theta}(\omega) = K_0[\Theta(\omega) - \hat{\Theta}(\omega)]F(\omega) \tag{10.5}$$

Здесь функции от ω, обозначенные буквами верхнего регистра, являются Фурье-образами соответствующих функций от t, обозначенных в уравнении (10.4) буквами нижнего регистра. Иными словами, $\hat{\Theta}(\omega) \leftrightarrow \hat{\theta}(t)$, $\Theta(\omega) \leftrightarrow \theta(t)$ и $F(\omega) \leftrightarrow f(t)$. После преобразования уравнения (10.5) получаем следующий результат.

$$\frac{\hat{\Theta}(\omega)}{\Theta(\omega)} = \frac{K_0 F(\omega)}{i\omega + K_0 F(\omega)} = H(\omega) \tag{10.6}$$

Член $H(\omega)$ называется передаточной функцией замкнутого контура ФАПЧ. Этот термин очень полезен при описании переходной характеристики контура ФАПЧ. Порядок контура ФАПЧ определяется старшим порядком $i\omega$ в знаменателе $H(\omega)$. Уравнение (10.6) показывает, что этот порядок всегда на единицу больше порядка контурного фильтра $F(\omega)$. Это объясняется тем, что $F(\omega)$ аналитически выражается как $F(\omega) = N(\omega)/D(\omega)$, знаменатель $H(\omega)$, при записи в виде полинома от $i\omega$, будет содержать член $i\omega D(\omega)$, который по $i\omega$ дает член, на один порядок больший члена максимального порядка в $D(\omega)$. Порядок контура ФАПЧ очень важен при определении стационарной характеристики контура при стационарном входе. Подробно этот вопрос рассматривается в следующем разделе.

10.2.1.1. Характеристики стационарного состояния

После преобразования уравнения (10.6) можно получить следующее выражение для Фурье-образа рассогласования по фазе.

$$\begin{aligned} E(\omega) &= \mathfrak{F}\{e(t)\} = \\ &= \Theta(\omega) - \hat{\Theta}(\omega) = \\ &= [1 - H(\omega)]\Theta(\omega) = \\ &= \frac{i\omega\Theta(\omega)}{i\omega + K_0 F(\omega)} \end{aligned} \tag{10.7}$$

Для определения характеристики установившейся ошибки контура при разнообразных выходных характеристиках можно использовать уравнение (10.7) и теорему об окончательном значении преобразования Фурье. Установившаяся ошибка — это остаточная ошибка после завершения всех переходных процессов, поэтому данная ошибка определяет, насколько контур способен справиться с различными типами изменений на входе. Теорема об окончательном значении формулируется следующим образом.

$$\lim_{t\to\infty} e(t) = \lim_{i\omega\to 0} i\omega E(\omega) \tag{10.8}$$

Объединяя уравнения (10.7) и (10.8), получаем следующее.

$$\lim_{t\to\infty} e(t) = \lim_{i\omega\to 0} \frac{(i\omega)^2\Theta(\omega)}{i\omega + K_0 F(\omega)} \tag{10.9}$$

Пример 10.2. Реакция на скачок фазы

Рассмотрите отклик контура, находящегося в стационарном состоянии, на скачок фазы на входе контура.

Решение

Предположим, что изначально контур ФАПЧ синхронизирован по фазе с входным сигналом, а скачок фазы вывел его из этого состояния. Причем после резкого изменения входная фаза снова стала стабильной. Вообще, это самый простой случай, с которым способен справиться контур ФАПЧ. Итак, Фурье-образ скачка фазы равен следующему.

$$\Theta(\omega) = \mathfrak{F}\{\Delta\phi\, u(t)\} = \\ = \frac{\Delta\phi}{i\omega} \tag{10.10}$$

Здесь $\Delta\phi$ — величина скачка, а $u(t)$ — единичная ступенчатая функция.

$$u(t) = \begin{cases} 1 & \text{для} > 0 \\ 0 & \text{для} < 0 \end{cases} = \\ = \int_{-\infty}^{t} \delta(\tau)\, d\tau$$

В последнем выражении $\delta(\tau)$ — дельта-функция Дирака. Из формул (10.9) и (10.10) получаем

$$\lim_{t\to\infty} e(t) = \lim_{i\omega\to 0} \frac{i\omega\Delta\phi}{i\omega + K_0 F(\omega)} = 0$$

в предположении, что $F(0) \neq 0$. Таким образом, при любом скачке фазы, происшедшем на входе, контур со временем синхронизируется, если характеристика контурного фильтра имеет ненулевую постоянную составляющую. Это означает, что для любого контурного фильтра, обладающего свойством $F(\omega) = N(\omega)/D(\omega)$ и $N(0) \neq 0$, контур ФАПЧ автоматически восстановит фазовую синхронизацию, если входной сигнал заменить сигналом с произвольной постоянной фазой. Очевидно, что это свойство контура является очень полезным.

Пример 10.3. Реакция на скачок частоты

Рассмотрите отклик контура, находящегося в стационарном состоянии, на скачок частоты на входе.

Решение

Посредством скачка частоты можно аппроксимировать последствия доплеровского смещения частоты входного сигнала вследствие относительного движения передатчика и приемника. Следовательно, данный пример важен для систем с мобильными терминалами. Поскольку фаза является интегралом частоты, при постоянном сдвиге входной частоты входная фаза (как функция времени) будет меняться линейно. Фурье-образ фазовой характеристики — это Фурье-образ интеграла частотной характеристики. Поскольку частотная характеристика — это ступенчатая функция, а образ интеграла — это образ подынтегрального выражения, деленного на параметр $i\omega$, можем записать

$$\Theta(\omega) = \frac{\Delta\omega}{(i\omega)^2}, \tag{10.11}$$

где $\Delta\omega$ — величина скачка частоты. Подстановка уравнения (10.11) в уравнение (10.9) дает следующий результат.

$$\lim_{t\to\infty} e(t) = \lim_{i\omega\to 0} \frac{\Delta\omega}{i\omega + K_0 F(\omega)} = \frac{\Delta\omega}{K_0 F(0)} \tag{10.12}$$

В данном случае стационарный результат зависит не только от ненулевой постоянной составляющей в характеристике, но и от других свойств контурного фильтра. Если фильтр является "всепропускающим" (широкополосным с полосой, равной бесконечности), то

$$F_{ap}(\omega) = 1. \tag{10.13}$$

Если фильтр является фильтром нижних частот, то

$$F_{lp}(\omega) = \frac{\omega_1}{i\omega + \omega_1}. \tag{10.14}$$

Если фильтр является стабилизирующим, то

$$F_{ll}(\omega) = \left(\frac{\omega_1}{\omega_2}\right)\frac{i\omega + \omega_2}{i\omega + \omega_1}. \tag{10.15}$$

Уравнение (10.12) показывает, что контур отследит изменение входной фазы с установившейся ошибкой, величина которой зависит от члена K_0 и величины скачка частоты. Подстановка любого из значений $F_{ap}(\omega)$, $F_{lp}(\omega)$ или $F_{ll}(\omega)$ в уравнение (10.12) дает следующий результат.

$$\lim_{t\to\infty} e(t) = \frac{\Delta\omega}{K_0}$$

Отметим, что произведение нескольких фильтров с характеристиками, подобными указанным в формулах (10.13), (10.14) или (10.15), по-прежнему будет давать желаемый результат. Стационарная ошибка, называемая *ошибкой по скорости*, будет существовать вне зависимости от порядка фильтра, если только знаменатель $F(\omega)$ не будет содержать $i\omega$ в виде множителя ($\omega_1 = 0$ в знаменателе уравнений (10.14) или (10.15) при соответствующей перенормировке числителей). Наличие $i\omega$ в виде множителя в $D(\omega)$ равносильно наличию идеального интегратора в контурном фильтре. Построить идеальный интегратор невозможно, но его можно достаточно хорошо аппроксимировать либо цифровым образом, либо с помощью активных интегральных схем [5]. Следовательно, если структура системы требует отслеживания доплеровского смещения при нулевой стационарной ошибке, контурный фильтр должен быть близок к идеальному интегратору. Следует отметить, что даже при ненулевой ошибке по скорости частота по-прежнему отслеживается: существуют важные системы, где стремление к нулевой фазовой ошибке не важно. В качестве примера можно привести не-

когерентную передачу сигналов, например сигналов с модуляцией FSK. Для некогерентной передачи действительно важным является отслеживание частоты, а абсолютное значение фазы не важно.

Пример 10.4. Реакция на линейное изменение частоты

Рассмотрите отклик контура, находящегося в стационарном состоянии, на линейное (по времени) изменение частоты на входе.

Решение

Ситуация, описанная в данном примере, соответствует ступенчатому изменению производной по времени от входной частоты. Это может, например, аппроксимировать изменение скорости доплеровского смещения, что позволило бы смоделировать ускорение относительного движения спутника (или самолета) и наземного приемника. В данном случае Фурье-образ фазовой характеристики дается следующим выражением.

$$\Theta(\omega) = \frac{\Delta\dot{\omega}}{(i\omega)^3} \tag{10.16}$$

Здесь $\dot{\omega}$ — скорость изменения частоты. В данном случае использование уравнения (10.9) дает следующий результат.

$$\lim_{t\to\infty} e(t) = \lim_{i\omega\to 0} \frac{\Delta\dot{\omega}/i\omega}{i\omega + K_0 F(\omega)} = \lim_{i\omega\to 0} \frac{\Delta\dot{\omega}}{i\omega K_0 F(\omega)} \tag{10.17}$$

Если контур имеет ненулевую ошибку по скорости (т.е. если правая часть уравнения (10.12) не равна нулю), уравнение (10.17) показывает, что стационарная фазовая ошибка становится неограниченной вследствие линейного изменения частоты. Это означает, что контур ФАПЧ с контурными фильтрами, характеристики которых описываются уравнениями (10.13)–(10.15), не сможет отследить линейное изменение частоты. Чтобы все-таки отследить это изменение, знаменатель преобразования контурного фильтра $D(\omega)$ должен в качестве множителя иметь $i\omega$. Из уравнения (10.17) видно, что контурный фильтр с передаточной функцией вида $F(\omega) = N(\omega)/i\omega D_1(\omega)$ позволит контуру ФАПЧ отследить линейное изменение частоты с постоянным рассогласованием по фазе. Из этого вытекает, что для отслеживания сигнала с линейно меняющимся доплеровским сдвигом (постоянным относительным ускорением) приемник должен содержать контур ФАПЧ второго или более высокого порядка. Для отслеживания линейного изменения частоты с нулевым рассогласованием по фазе потребуется контурный фильтр с передаточной функцией, имеющей в знаменателе множитель $(i\omega)^2$: $F(\omega) = N(\omega)/(i\omega)^2 D_2(\omega)$. Из этого следует, что контур ФАПЧ должен быть третьего или более высокого порядка. Следовательно, в высокоэффективных самолетах, которые должны точно отслеживать фазу при различных маневрах, могут требоваться контуры ФАПЧ третьего или более высокого порядка. Во всех случаях синхронизация частоты получается с помощью контура на один порядок ниже, чем необходимо для синхронизации фазы. Итак, анализ стационарной ошибки является полезным показателем требуемой сложности контурных фильтров.

На практике подавляющее большинство контуров ФАПЧ имеет второй порядок. Это объясняется тем, что контур второго порядка можно спроектировать безусловно устойчивым [5]. Безусловно устойчивые контуры всегда будут пытаться отследить входной сигнал. Никакие входные условия не приведут к тому, что контур будет реагировать на изменения входа в ненадлежащем направлении. Контуры второго порядка отследят последствия скачка частоты (доплеровского смещения); кроме того, они относительно просто анализируются, поскольку аналитические выражения, полученные для контуров первого порядка, являются хорошей аппроксимацией для контуров второго порядка. Контуры третьего порядка применяются в некоторых специальных областях, например некоторые навигационные приемники систем GPS (Global Positioning System — глобальная система навигации и определения положения) и некоторые

авиационные приемники. В то же время характеристики таких контуров относительно сложно определить, кроме того, контуры третьего и более высоких порядков являются только условно устойчивыми. Если же вследствие динамики сигнала для когерентной демодуляции потребуются контуры третьего и более высоких порядков, то вместо этого используется некогерентная демодуляция.

10.2.1.2. Производительность при шуме

При анализе стационарного состояния в предыдущем разделе подразумевалось, что входной сигнал не зашумлен. В некоторых случаях это может быть справедливо, но в общем случае анализа связи воздействие шума все же следует учитывать.

Вернемся к нормированному входному сигналу, приведенному в формуле (10.1) и изображенному на рис. 10.1. При включении нормированного узкополосного аддитивного гауссового шума $n(t)$ выражение для входного сигнала принимает следующий вид.

$$r(t) = \cos(\omega_0 t + \theta) + n(t) \tag{10.18}$$

Здесь входной сдвиг фазы θ пока считаем константой. Предполагается, что процесс шума $n(t)$ является узкополосным гауссовым процессом с нулевым средним и его можно разложить по квадратурным составляющим несущей [6].

$$n(t) = n_c(t)\cos\omega t + n_s(t)\sin\omega_0 t \tag{10.19}$$

Здесь $n_c(t)$ и $n_s(t)$ — случайные, независимые между собой, гауссовы процессы с нулевым средним. Теперь выход детектора фазы можно записать следующим образом (см. уравнение (10.3)).

$$\begin{aligned} e(t) = x(t)r(t) &= \sin(\theta - \hat{\theta}) + n_c(t)\cos\hat{\theta} + n_s(t)\sin\hat{\theta} + \\ &+ (\text{члены с частотой, равной удвоенной несущей}) \end{aligned} \tag{10.20}$$

Как и выше, контурный фильтр отсекает члены с частотой, равной удвоенной несущей. Обозначим второй и третий члены уравнения (10.20) следующим образом.

$$n'(t) = n_c(t)\cos\hat{\theta} + n_s(t)\sin\hat{\theta} \tag{10.21}$$

Легко доказать, что дисперсия $n'(t)$ равна дисперсии $n(t)$. Далее эта дисперсия обозначается как σ_n^2.

Рассмотрим автокорреляционную функцию от $n'(t)$

$$\begin{aligned} R(t_1, t_2) &= \mathbf{E}\{n'(t_1)n'(t_2)\} = \\ &= \mathbf{E}\{n_c(t_1)n_c(t_2)\}\cos^2\hat{\theta} + \mathbf{E}\{n_s(t_1)n_s(t_2)\}\sin^2\hat{\theta} + \\ &+ [\mathbf{E}\{n_c(t_1)n_s(t_2)\} + \mathbf{E}\{n_s(t_1)n_c(t_2)\}]\sin\hat{\theta}\cos\hat{\theta}, \end{aligned} \tag{10.22}$$

где $\mathbf{E}\{\cdot\}$ обозначает математическое ожидание. Перекрестные произведения в правой части уравнения (10.22) равны нулю, поскольку n_c и n_s взаимно независимы и имеют нулевые средние [6]. Если принять предположения о стационарности процесса в широком смысле [7], получим

$$R(\tau) = R_c(\tau)\cos^2\hat{\theta} + R_s(\tau)\sin^2\hat{\theta}, \tag{10.23}$$

где $\tau = t_1 - t_2$. Если применить преобразование Фурье к обеим частям уравнения (10.21), то спектральную плотность мощности $n'(t)$ можно будет записать в следующем виде.

$$G(\omega) = \mathfrak{F}[R(t)] = \\ = G_c(\omega)\cos^2\hat{\theta} + G_s(\omega)\sin^2\hat{\theta} \tag{10.24}$$

Здесь G_c и G_s — Фурье-образы R_c и R_s. Из уравнения (10.19) видно, что спектры G_c и G_s составлены из смещенных версий спектра исходного процесса шума $n(t)$. Таким образом, вследствие выбранной структуры [8],

$$G_s(\omega) = G_c(\omega) = G_n(\omega_0 - \omega) + G_n(\omega_0 + \omega),$$

где $G_n(\omega)$ — спектральная плотность исходного широкополосного процесса шума $n(t)$. Уравнение (10.24) можно переписать следующим образом.

$$G(\omega) = G_n(\omega_0 - \omega) + G_n(\omega_0 + \omega) \tag{10.25}$$

Для частного случая белого шума имеем $G_n(\omega) = N_0/2$ Вт/Гц, где N_0 — односторонняя спектральная плотность белого шума. Следовательно, из уравнения (10.25) для этого важного частного случая получаем следующее.

$$G(\omega) = N_0 \tag{10.26}$$

Важность полученного результата состоит в том, что для того же приближения малых углов, которое было принято в предыдущем разделе, спектральная плотность фазы ГУН, $G_{\hat{\theta}}$, связана со спектральной плотностью процесса шума через передаточную функцию контура (уравнение (10.6)). Иными словами,

$$G_{\hat{\theta}}(\omega) = G(\omega)\,|H(\omega)|^2, \tag{10.27}$$

где $G(\omega)$ выражено в формуле (10.25), а $H(\omega)$ определено в (10.6). Таким образом, дисперсия выходной фазы равна следующему.

$$\sigma_{\hat{\theta}}^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)\,|H(\omega)|^2\,d\omega \tag{10.28}$$

Для частного случая белого шума

$$\sigma_{\hat{\theta}}^2 = \frac{N_0}{2\pi}\int_{-\infty}^{\infty} |H(\omega)|^2\,d\omega \tag{10.29}$$

Интеграл в уравнении (10.29) (нормированный на собственную частоту) называется *двусторонней полосой контура* W_L. *Односторонняя полоса контура* обозначается как B_L. Определяются эти величины следующим образом.

$$W_L = 2B_L = \frac{1}{2\pi}\int_{-\infty}^{\infty} |H(\omega)|^2\,d\omega \text{ Гц} \tag{10.30}$$

Следовательно, если процесс шума является белым и, кроме того, принято приближение малых углов (другими словами, контур успешно отслеживает входную фазу), дисперсия фазы дается следующим выражением.

$$\sigma_{\hat{\theta}}^2 = 2N_0 B_L \tag{10.31}$$

Дисперсия фазы — это мера неустойчивой синхронизации на выходе генератора, управляемого напряжением, вследствие шума на входе. Уравнения (10.31) и (10.7) описывают один из множества компромиссов в теории связи. Очевидно, что величину $\sigma_{\hat{\theta}}^2$ хотелось бы сделать как можно меньше; при данном уровне шума это подразумевает меньшую полосу контура B_L, а из уравнения (10.30) следует более узкая функция $H(\omega)$. В то же время из уравнения (10.7) можно сделать вывод, чем уже эффективная полоса $H(\omega)$, тем хуже способность контура к отслеживанию изменения фазы поступающего сигнала $\Theta(\omega)$. Следовательно, при проектировании контура должен достигаться определенный баланс между параметрами, связанными с шумом, и желаемой реакцией на изменения входной фазы. Перед разработчиком стоит задача: разработать контур, который бы надлежащим образом реагировал на изменения входного сигнала, но при этом не был бы слишком чувствителен к кажущимся изменениям, которые на самом деле являются следствиями процесса шума.

10.2.1.3. Анализ нелинейного контура

Обсуждение контура ФАПЧ, приведенное в предыдущих разделах, основывалось на линеаризованной модели контура ФАПЧ. Схематически эта модель показана на рис. 10.2. В данной модели использовано приближение малых углов.

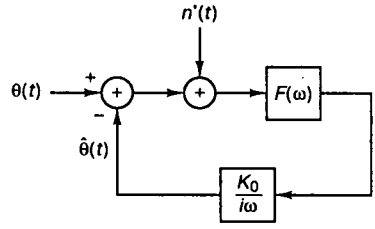


Рис. 10.2. Схема линеаризованной модели контура ФАПЧ

$$\sin(\theta - \hat{\theta}) \approx \theta - \hat{\theta} \tag{10.32}$$

Данное приближение справедливо, когда контур синхронизирован и функционирует желаемым образом (т.е. с небольшими рассогласованиями по фазе). Очевидно, эти условия формируют только часть общей картины. Полный анализ производительности контура ФАПЧ должен исходить из предположения, что уравнение (10.32) справедливо не всегда. Когда приближение малых углов становится неточным, подходящей моделью является изображенная на рис. 10.3. С помощью формул (10.4), (10.20) и (10.21) и рис. 10.3 модель можно описать следующим дифференциальным уравнением.

$$\frac{d}{dt}[\hat{\theta}(t)] = K_0 f(t) * \sin[\theta(t) - \hat{\theta}(t)] + K_0 f(t) * n'(t) \tag{10.33}$$

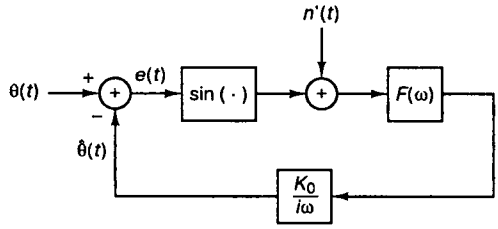


Рис. 10.3. Схема нелинейной модели контура ФАПЧ

Здесь, как и ранее, знак * обозначает операцию свертки. Несмотря на значительные усилия исследователей, общее решение данного дифференциального уравнения не удается найти на протяжении многих лет. Впрочем, Витерби (Viterbi) [8] вывел аналитическое решение для одного важного частного случая.

Рассмотрим следующий случай. Пусть входная фаза $\theta(t)$, которая, вообще-то, является функцией времени, равна константе θ. Определим теперь новую фазовую переменную

$$\phi(t) = [\theta - \hat{\theta}(t)] \text{ по модулю } 2\pi. \tag{10.34}$$

Поскольку θ — это константа, уравнение (10.33) можно переписать следующим образом.

$$\frac{d}{dt}[\phi(t)] = K_0 f(t) * \sin\phi(t) + K_0 f(t) * n'(t) \tag{10.35}$$

Поскольку из уравнения (10.35) $\phi(t)$ является функцией случайного процесса $n'(t)$, сама $\phi(t)$ также есть случайным процессом. Так как фаза $\phi(t)$ определена по модулю 2π, можно показать [5], что $\phi(t)$ стационарна в пределе, по прошествии всех переходных процессов (т.е. θ — константа). Витерби [8] определил, что для контура ФАПЧ первого порядка (т.е. контурный фильтр — это просто цепь короткого замыкания, или, что эквивалентно, $f(t) = \delta(t)$) функция плотности вероятности ϕ имеет следующий вид.

$$p(\phi) = \frac{\exp(\rho\cos\phi)}{2\pi I_0(\rho)} \text{ для } |\phi| \le \pi \tag{10.36}$$

Здесь $\rho = 1/\sigma_\theta^2$ (см. уравнение (10.31)) — нормированное (на энергию единичного сигнала) отношение сигнал/шум контура, а $I_0(\rho)$ — модифицированная функция Бесселя первого рода нулевого порядка, взятая в точке ρ. Дисперсию фазы по модулю 2π теперь можно вычислить с использованием уравнения (10.36). Полученное значение дисперсии фазы будет точным для контуров первого порядка и весьма хорошим приближением для многих контуров второго порядка [5]. В работе [9] было показано, что это выражение также справедливо для контуров высоких порядков при несколько видоизмененном определении ρ.

Замена переменной с фазы, которая может принимать любое действительное значение, на фазу по модулю 2π приводит к необходимости введения понятия проскальзывания цикла контура. Проскальзывание цикла происходит, когда величина исходного рассогласования по фазе $|\theta - \hat{\theta}(t)|$ превышает 2π радиан. Это приводит к внезапному изменению значения ϕ (уравнение (10.34)) с 2π на 0. Данное явление можно рассматривать как мгновенную потерю синхронизации с практически немедленным

ее восстановлением. Статистика проскальзываний цикла может быть таким же важным показателем производительности контура ФАПЧ, как и дисперсия фазы — особенно при низких отношениях сигнал/шум в контуре, когда проскальзывание цикла может происходить довольно часто.

Витерби, используя выражения, полученные для распределения фаз, вывел [8] выражения для среднего времени до первого проскальзывания цикла T_m, отсчитываемого от некоторого произвольного эталонного времени.

$$T_m = \frac{\pi^2 \rho I_0^2(\rho)}{2B_L} \tag{10.37}$$

При больших ρ это выражение можно приближенно записать следующим образом.

$$T_m \approx \frac{\pi \exp(2\rho)}{4B_L} \tag{10.38}$$

Как и для функции плотности вероятности в уравнении (10.36), полученные результаты выведены для контуров первого порядка, но они являются полезной аппроксимацией для контуров второго порядка и описывают верхнюю границу производительности циклов второго порядка при средних и больших отношениях сигнал/шум в контуре. Кроме того, компьютерное моделирование и лабораторные измерения [5] показывают, что время T между проскальзываниями цикла имеет экспоненциальное распределение.

$$P(T) = 1 - \exp\left(-\frac{T}{T_m}\right) \tag{10.39}$$

Иными словами, вероятность того, что в течение промежутка времени T при нулевом текущем рассогласовании по фазе произойдет проскальзывание цикла, описывается выражением (10.39).

10.2.1.4. Схемы подавления несущей

До настоящего момента при обсуждении контуров ФАПЧ предполагалось, что входящая несущая — это достаточно устойчивая синусоида с некоторой известной средней положительной энергией. В системе связи с фазовой модуляцией несущая частота будет переносить положительную энергию, если дисперсия фазы несущей, вследствие модуляции, меньше $\pi/2$ радиан. В этом случае говорят, что в системе имеется остаточная составляющая несущей. Все обсуждение разработки контуров ФАПЧ, приведенное выше, применимо непосредственно к этой остаточной составляющей. Диаграмма сигнального пространства для системы бинарной фазовой модуляции с остаточной составляющей несущей показана на рис. 10.4 для угла модуляции $\gamma \leq \pi/2$. Одно время подобным образом разрабатывалось большинство систем с фазовой модуляцией. В то же время остаточная составляющая несущей является в некотором смысле бесполезно растрачиваемой энергией — энергия на остаточной несущей используется не для передачи информации, а только для передачи самой несущей. Поэтому большинство современных систем фазовой модуляции являются системами с подавлением несущей. Это означает, что на несущей частоте не имеется никакой средней передаваемой энергии. Вся передаваемая энергия уходит на модуляцию. К сожалению, это означает, что не существует сигнала, составляющего основу для отслеживания с помощью простого контура ФАПЧ, показанного на рис. 10.1.

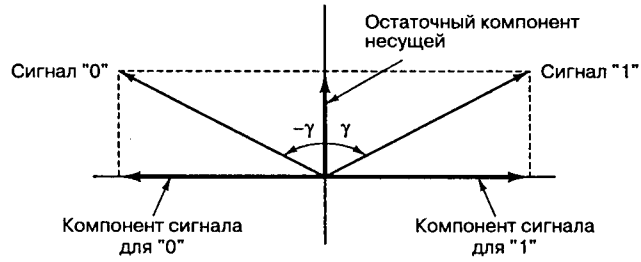


Рис. 10.4. Бинарная фазовая модуляция с остаточной несущей

Рассмотрим в качестве примера сигнал с модуляцией BPSK

$$r(t) = m(t) \sin(\omega_0 t + \theta) + n(t), \tag{10.40}$$

где $m(t)$ с равной вероятностью равен ±1. Данный пример — это передача с подавлением несущей; средняя энергия на угловой частоте ω_0 равна нулю. Графически это представлено на рис. 10.4, где $\gamma = \pi/2$. Из рисунка видно, что в данном случае горизонтальный компонент несущей исчезает. Для отслеживания и синхронизации фазы несущей последствия модуляции необходимо устранить. Это можно сделать путем возведения сигнала в квадрат.

$$\begin{aligned} r^2(t) &= m^2(t) \sin^2(\omega_0 t + \theta) + n^2(t) + 2n(t)m(t) \sin(\omega_0 t + \theta) = \\ &= 1/2 - 1/2 \cos(2\omega_0 t + 2\theta) + n^2(t) + 2n(t)m(t) \sin(\omega_0 t + \theta) \end{aligned} \tag{10.41}$$

Выше использовано $m^2(t) = 1$. Второй член в правой части уравнения (10.41) зависит от несущей (от удвоенной частоты несущей) и может быть отслежен с помощью простого контура ФАПЧ, показанного на рис. 10.1. Соответствующая схема показана на рис. 10.5. При возведении входного сигнала с подавленной несущей в квадрат получаемый компонент, зависящий от удвоенной несущей, можно выделить и отследить с помощью стандартного контура ФАПЧ. Изучение уравнения (10.41) позволяет предсказать некоторые потенциальные проблемы такой схемы. Одна из них — это просто удвоение всех фазовых углов. Следовательно, фазовый шум и случайное смещение фазы также удваиваются, и дисперсия фазовой ошибки (связанная с возведенным в квадрат фазовым шумом) в 4 раза больше по сравнению с исходным сигналом. Этот удваивающийся угол нейтрализуется схемой деления на 2 на выходе ГУН и, следовательно, не влияет непосредственно на точность выходного сигнала контура, используемого для демодуляции данных. В то же время эта большая внутренняя дисперсия приведет к тому, что контур ФАПЧ потребует для поддержания фазовой синхронизации на 6 дБ большего отношения сигнал/шум, чем система с остаточной несущей. Кроме того, вследствие взаимной корреляции между шумом и сигналом в уравнении (10.41) теперь существует два эффективных члена шума, который мешает работе контура. Для сред или контуров с низким отношением сигнал/шум данные два члена шума еще больше снизят номинальное отношение сигнал/шум по сравнению с исходным немодулированным сигналом. Эти дополнительные потери, обусловленные членами произведения сигнал-шум и шум-шум, называются *потерями вследствие возведения в квадрат* и обозначаются S_L. Гарднер (Gardner) [5] показал, что если входной процесс шума $n(t)$ является узкополосным гауссовым шумом с шириной полосы B_i, то потери вследствие возведения в квадрат ограничены сверху следующей величиной.

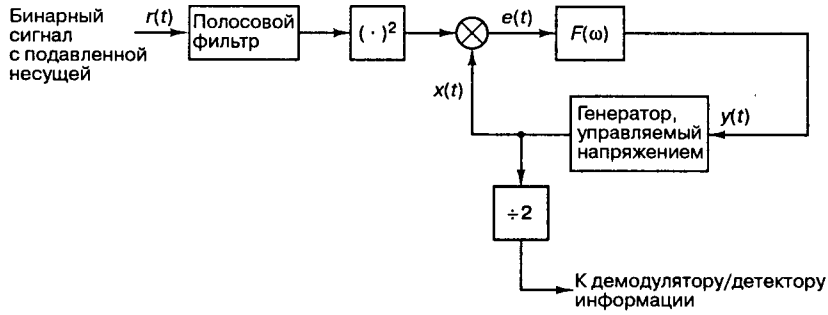


Рис. 10.5. Схема контура возведения в квадрат

$$S_L \leq 1 + N_0 B_i \tag{10.42}$$

Здесь, как и выше, N_0 — односторонняя спектральная плотность мощности предварительно фильтрованного нормированного процесса белого гауссового шума. Уравнение (10.42) представляет собой верхнюю границу, поскольку подразумевается, что ширина полосы фильтра B_i достаточно велика для неискаженной передачи сигнала. В реальных системах (как показано в [10]) потери вследствие возведения в квадрат можно устранить за счет некоторого искажения сигнала.

Поскольку нормирование в уравнении (10.42) выполняется относительно мощности сигнала, второй член пропорционален отношению сигнал/шум.

$$\rho_i = \frac{1}{2N_0 B_i} \tag{10.43}$$

Здесь ρ_i — отношение сигнал/шум на входе фильтра. Для больших отношений сигнал/шум в контуре выходную дисперсию фазы можно записать следующим образом.

$$\sigma_{\hat{\theta}}^2 = 2N_0 B_L S_L = 2N_0 B_L \left(1 + \frac{1}{2\rho_i}\right) \tag{10.44}$$

Видим, что главный член в правой части уравнения (10.44) идентичен главному члену в уравнении (10.31), дисперсии фазы стандартного контура ФАПЧ. Кроме того, для больших входных отношений сигнал/шум второй член в выражении для потерь вследствие возведения в квадрат исчезает и остается только дисперсия фазы стандартного контура ФАПЧ.

Еще одна потенциальная серьезная проблема, связанная преимущественно с контурами подавления несущей, — это *ложная синхронизация*, которая может затруднить синхронизацию и восстановление синхронизации фазы несущей. Взаимодействие информационного потока с нелинейностями контура (особенно схемы возведения в квадрат) и контурными фильтрами будет порождать боковые полосы в спектре, поступающем на вход детектора фазы. Эти боковые полосы могут содержать компоненты с устойчивыми частотами. Необходимо следить, чтобы эти устойчивые компоненты не захватывались контуром слежения. Если контур захватит подобную частоту, может создаться впечатление, что он функционирует нормально; управляющий сигнал ГУН $y(t)$ будет небольшим, но выход ГУН будет смещен по частоте от истинной несущей. Описанная ситуация называется ложной синхронизацией. Контур отслеживает компонент боковой полосы частот, а контурный фильтр отфильтровывает действительную несущую. Эта проблема до-

вольно часто определяет нижний предел полосы контурных фильтров. Поскольку фильтры контуров остаточных несущих содержат меньше нелинейных компонентов, ложная синхронизация не является для них серьезной проблемой.

10.2.1.5. Синфазно-квадратурные схемы

Важной разновидностью контуров подавления несущей является *синфазно-квадратурная схема* (Costas loop), схематически изображенная на рис. 10.6. Эта схема важна, поскольку она позволяет избежать применения устройства возведения в квадрат, реализация которого на несущих частотах может быть затруднительной. Вместо этого в контур вводится умножитель и относительно простые фильтры нижних частот. Хотя внешне схемы на рис. 10.5 и 10.6 достаточно различны, можно показать [5], что их теоретические производительности равны. Основной проблемой реализации синфазно-квадратурных схем является то, что для получения теоретической оптимальной производительности два фильтра нижних частот должны быть идеально согласованы. Этого можно достичь в любой аналоговой аппаратной реализации. Если фильтры реализуются цифровым образом, то проблем с поддержанием их согласованности не возникнет, но разработчик сталкивается с обычными проблемами разработки схем, оперирующих дискретными данными. Таким образом, решение о том, какой контур использовать — классический (рис. 10.5) или синфазно-квадратурный (рис. 10.6), — эквивалентно выбору между сложностью реализации устройства возведения в квадрат и сложностью реализации идеально согласованных фильтров. Это проектное решение будет зависеть от параметров и требований конкретной принимающей системы, поэтому универсального совета мы дать не можем.

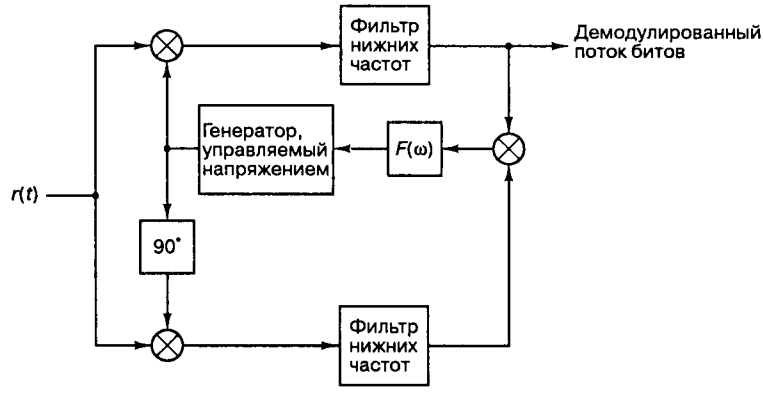


Рис. 10.6. Синфазно-квадратурная схема

10.2.1.6. Схемы подавления несущей высших порядков

Двоичная фазовая манипуляция (binary phase-shift keying — BPSK) — это не единственная модуляция с подавлением несущей. Фактически, если предположить, что *априорно* все сигналы равновероятны, любая модуляция, средняя амплитуда которой, усредненная по сигнальному множеству, равна нулю, не будет иметь средней энергии на передаваемой несущей. Возможно, самой распространенной недвоичной модуляцией с подавлением несущей является квадратурная фазовая манипуляция (quadrature phase-shift keying — QPSK). При возведении сигнала QPSK в квадрат результат выглядит подобно сигналу BPSK. Следовательно, для равновероятных сигналов QPSK несущая по-прежнему подавляется. В то же время повторное возведение сигнала в квад-

рат — что равносильно возведению исходного сигнала в четвертую степень — дает член, где частота несущего компонента в 4 раза больше частоты переданной несущей. Как и при двоичной модуляции, пропускание входного сигнала через устройство возведения в степень дает перекрестные произведения шума и сигнала и вводит эквивалент "потерь вследствие возведения в квадрат". Если предположить, что ширина полосы шумов достаточна для пропускания сигнала без искажения, потери в контурах возведения в четвертую степень будут ограничены сверху следующей величиной [5].

$$S_L \le 1 + \frac{9}{\rho_i} + \frac{6}{\rho_i^2} + \frac{3}{2\rho_i^3} \tag{10.45}$$

Как и в схеме возведения в квадрат, при значительных входных отношениях сигнал/шум ρ_i из уравнения (10.45) видно, что дополнительные члены потерь исчезают и производительность данного контура сравнивается с производительностью обычного контура. Как и для контуров второго порядка, существуют синфазно-квадратурные схемы, эквивалентные контурам четвертого порядка [5, 14, 15], реализация которых может иметь определенные аппаратные преимущества. Впрочем, их теоретическая производительность аналогична производительности обычных контуров четвертого порядка.

Пример 10.5. Пределы потерь вследствие возведения в квадрат

Сравните верхние границы потерь вследствие возведения в квадрат S_L, приведенные в уравнениях (10.42) и (10.45) для контуров второго и четвертого порядков. Входное отношение сигнал/шум ρ_i считать равным 10 дБ.

Решение

Из уравнений (10.42)–(10.44) для схемы возведения в квадрат получаем следующий результат.

$$S_L = 1 + \frac{1}{2\rho_i} = 1{,}05 = 0{,}2 \text{ дБ}$$

Из уравнения (10.45) для контура возведения в четвертую степень получаем следующее.

$$S_L = 1 + 0{,}9 + 0{,}06 + 0{,}0015 = 1{,}9615 = 2{,}9 \text{ дБ}$$

Следовательно, если входного отношения сигнал/шум, равного 10 дБ, достаточно для поддерживания небольших потерь в контуре возведения в квадрат, то же отношение может приводить к значительным потерям в контуре возведения в четвертую степень.

10.2.1.7. Начальная синхронизация

Ранее при обсуждении большинства вопросов предполагалось, что контур ФАПЧ изначально синхронизирован. Это оправдано, если рассогласование по фазе $|\theta - \hat{\theta}|$ мало. В то же время иногда контур должен достигать синхронизации, т.е. его нужно синхронизировать. Начальная синхронизация может выполняться с помощью внешних схем или сигналов (принудительная синхронизация) либо посредством автосинхронизации [5].

По сути, синхронизация — это нелинейная операция; следовательно, общий ее анализ затруднителен. Впрочем, некоторые интуитивно приемлемые результаты можно получить при рассмотрении свободного от шумов контура первого порядка. Подобный контур изображен на рис. 10.3, где $n'(t) = 0$ (отсутствие шумов) и $F(\omega) = 1$ (первый порядок). Запишем входную фазу

$$\theta(t) = \omega_i t$$

и выходную фазу

$$\hat{\theta}(t) = \omega_0 t + \int_0^t K_0 \sin e(t)\, dt + \hat{\theta}(0)\,, \tag{10.46}$$

где ω_i и ω_0 — угловая частота входного и выходного сигналов. Следовательно, рассогласование по фазе дается следующим выражением.

$$\begin{aligned} e(t) &= \theta(t) - \hat{\theta}(t) = \\ &= (\omega_i - \omega_0)t + \int_0^t K_0 \sin e(t)\, dt + \hat{\theta}(0) \end{aligned} \tag{10.47}$$

Дифференцируя обе части предыдущего выражения и полагая $\Delta\omega = \omega_i - \omega_0$, получаем следующее.

$$\frac{de}{dt} = \Delta\omega - K_0 \sin e \tag{10.48}$$

Здесь для простоты записи опущен аргумент (время) функции $e(t)$. Данное дифференциальное уравнение описывает поведение свободного от шумов контура ФАПЧ первого порядка. Условие синхронизации записывается следующим образом.

$$\frac{de}{dt} = 0 \tag{10.49}$$

Уравнение (10.49) является необходимым, но не достаточным условием фазовой синхронизации. Это можно проверить, изучив диаграмму фазовой плоскости на рис. 10.7. На данном рисунке отображены результаты деления обеих частей уравнения (10.48) на K_0. Сначала рассмотрим точку a. Если рассогласование по фазе приведет к небольшому смещению точки, описывающей состояние контура, вправо или влево от a, знак производной обеспечит смещение фазовой ошибки e к точке a. Следовательно, точка a — это устойчивая точка системы; точка, где можно получить фазовую синхронизацию и где эта синхронизация будет поддерживаться. Рассмотрим теперь точку b. Если рассогласование по фазе e находится точно в точке b, уравнение (10.49) будет удовлетворено. В то же время, если e несколько сместится от точки b, то знак производной обусловит дальнейшее смещение от b. Следовательно, b — точка, где уравнение (10.49) удовлетворяется, но решение не является устойчивым.

Время, необходимое контуру для синхронизации, может быть важным параметром при проектировании системы. Изучая уравнение (10.48), можно видеть, что требования уравнения (10.49) к фазовой синхронизации не могут удовлетворяться, если не выполнено следующее условие.

$$\frac{|\Delta\omega|}{K_0} \le 1 \tag{10.50}$$

Это объясняется тем, что максимальная амплитуда синусоидальной функции получается при аргументе, равном единице. Этот диапазон разности частот $-K_0 < \Delta\omega < K_0$ иногда называют диапазоном синхронизации контура. Предполагая, что условие (10.50) удовлетворяется для времени, требуемого для синхронизации контура, Гарднер [5] предложил

эвристическую величину $3/K_0$ секунд. Реальные значения из уравнения (10.47) можно получить аналитически (для однозначно определенных наборов начальных условий) или с помощью компьютерного моделирования. Из графика на рис. 10.7 видно, что необходимое время сильно меняется как функция первоначального рассогласования по фазе. Для значений e, близких к точке b, управляющий фактор $(de/dt)/K_0$ будет очень мал. Поэтому в наихудшем случае фазовая ошибка будет долго находится в окрестности точки b. Это явление называется зависанием конечного цикла [16] и может представлять серьезную проблему в системах с автосинхронизацией.

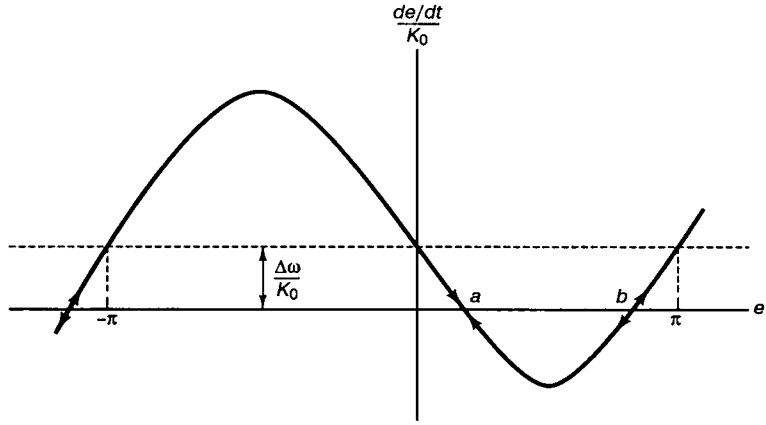


Рис. 10.7. Изображение контура первого порядка на фазовой плоскости

Возможно, важнейшим операционным различием контуров первого и высших порядков является способность последних "выскакивать" из разностей частот, не входящих в диапазон синхронизации. Контур первого порядка с рассогласованием частоты, превышающим частоты диапазона синхронизации, будет стремиться к нужному диапазону, но никогда это не будет происходить быстро. Почему? Контуры второго и высших порядков могут входить в синхронизацию вследствие их более сложных фазовых характеристик. (Читателям, интересующимся этим вопросом, можно посоветовать работы [5, 8, 9, 17–19].)

Изучение автосинхронизации для контуров ФАПЧ представляет преимущественно академический интерес. Гарднер [5] утверждает, что контуры автосинхронизации, дающие требуемый результат за разумное время, могут создаваться только при весьма благоприятных условиях. К сожалению, на практике такие условия встречаются крайне редко.

Принудительная синхронизация — это перенос рабочей точки контура в область фазового пространства, где предположительно находится область синхронизации, посредством некоторого внешнего направляющего сигнала. Это является наиболее распространенным методом получения синхронизации. Внешняя помощь может быть реализована путем простой подачи линейного изменения напряжения на вход ГУН. Этот направляющий сигнал приведет к тому, что выходная частота ГУН будет линейно изменяться во времени. Как показывалось ранее (уравнение (10.17)), схемы с контурными фильтрами, знаменатели передаточных функций которых не содержат множителя $i\omega$, не смогут отследить линейное изменение частоты с конечным рассогласованием по фазе. Следовательно, если поиск частоты должен реализовываться на контуре первого или второго порядка без этой особенности передаточной функции,

скорость изменения частоты должна быть достаточно малой, чтобы после синхронизации контура наличие синхронизации по фазе могло быть обнаружено и поисковый сигнал был удален до того, как он выведет контур из синхронизации. Для контуров, содержащих в $D(\omega)$ множитель $i\omega$, удалять поисковый сигнал не обязательно, поскольку (по крайней мере, теоретически) контур сможет отследить линейное изменение частоты. В любом случае частота сканирования не должна быть слишком большой, иначе контур будет проскакивать мимо точки синхронизации так быстро, что ее будет невозможно достичь. Для контура второго порядка с передаточной функции (см. уравнение (10.6))

$$H(\omega) = \frac{1}{-(i\omega/\omega_n)^2 + 2\zeta(i\omega/\omega_n) + 1} \tag{10.51}$$

Гарднер [5] показал, что максимальная скорость сканирования $\Delta\omega$ должна быть близка к следующей величине.

$$\Delta\omega \approx \frac{1}{2}\omega_n^2 (1 - 2\sigma_{\hat{\theta}}) \tag{10.52}$$

Здесь $\sigma_{\hat{\theta}}$ определено в выражении (10.31), а ω_n, неявно определенное в формуле (10.51), называется *собственной частотой* контура ФАПЧ второго порядка и связано с шириной полосы контура B_L и декрементом затухания контура ζ следующим соотношением.

$$\omega_n = \frac{8\zeta}{4\zeta^2 + 1} B_L$$

Более подробное исследование принудительной синхронизации приведено в работе [17].

10.2.1.8. Ошибки сопровождения фазы и производительность канала

Если контур не способен отследить все фазовые ошибки, вероятность ошибки в принятом символе будет больше теоретически достижимой. Анализ, который требуется провести для определения объема ухудшения, весьма сложен, но для большинства стандартных схем когерентной передачи сигналов эта работа уже сделана [14, 15, 20]. На рис. 10.8 приведен пример зависимости производительности для контура остаточной несущей, работающего с сигналами в модуляции BPSK при аддитивном белом гауссовом шуме. Видно, что для средних значений отношения сигнал/шум небольшое рассогласование по фазе приводит к незначительному ухудшению производительности. Ухудшение становится значительным только тогда, когда среднеквадратическое отклонение рассогласования по фазе начинает превышать 0,3. Это означает, что собственным ухудшением производительности качественных контуров, работающих в благоприятных условиях, можно, в общем случае, пренебрегать. Приведенный график также показывает, что если дисперсия фазы велика, то увеличение отношения сигнала к гауссовому шуму может быть неэффективной мерой по снижению вероятности обнаруженной ошибки. Следует отметить, что наличие неустранимой ошибки в этих ситуациях характерно для схем остаточной несущей с постоянным отношением сигнал/шум в контуре ρ_i. Схемы с подавлением несущей не имеют тенденции к возникновению неустранимых ошибок, поскольку увеличение отношения информационного сигнала к шуму повышает отношение сигнал/шум в контуре сопровождения подавленной несущей, что приводит к уменьшению ошибки сопровождения.

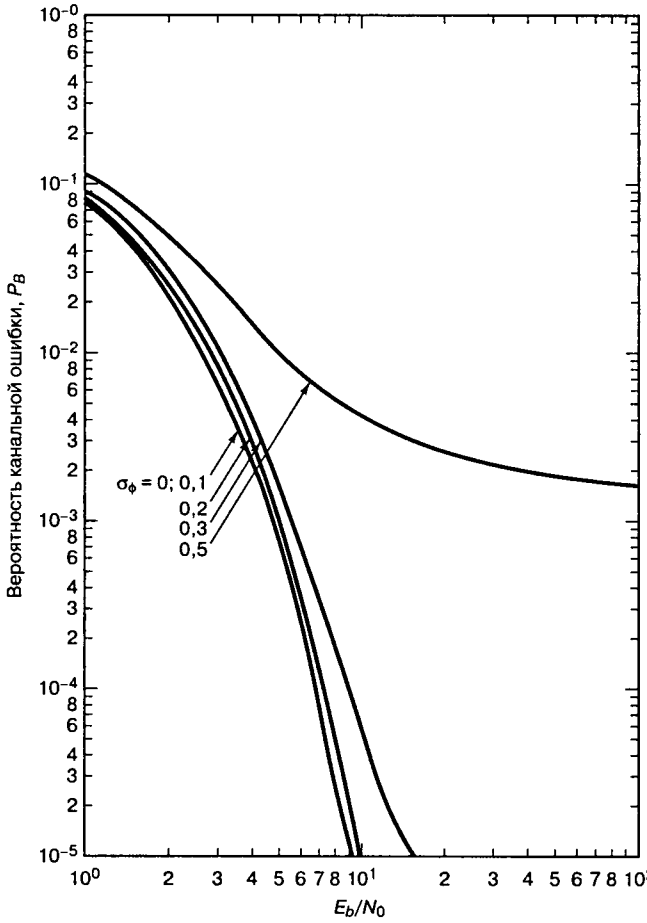


Рис. 10.8. Зависимость вероятности битовой ошибки от E_b/N_0 для модуляции BPSK при неидеальной синхронизации несущей. (Перепечатано с разрешения автора из J. J. Stiffler. Theory of Synchronous Communications. *Prentice-Hall, Inc., Englewood Cliffs, N. J., Fig. 9.1, p. 270.)*

Пример 10.6. Отношение сигнал/шум в контуре ФАПЧ

Выведите интегральное выражение для влияния медленно меняющейся ошибки сопровождения фазы на вероятность битовой ошибки для канала с остаточной несущей. При передаче сигналов применяется модуляция BPSK. Используя рис. 10.8, сравните результаты для нормированных отношений сигнал/шум ($\rho = 1/\sigma_{\tilde{\theta}}^2$), равных 20 и 10 дБ, при желательной вероятности битовой ошибки 10^{-5}.

Решение

Из главы 4 для канала с модуляцией BPSK при аддитивном белом гауссовом шуме теоретическая зависимость вероятности битовой ошибки от односторонней спектральной плотности N_0 Вт/Гц дается выражением

$$P_B = Q\left(\sqrt{\frac{2E_b}{N_0}}\right),$$

где E_b — энергия, принятая за время передачи одного бита. Если внимательно проследить вывод этого выражения для вероятности ошибки, то можно показать, что при медленно меняющейся (относительно скорости передачи данных) ошибке сопровождения фазы, β радиан, вероятность ошибки будет равна следующему.

$$P_B(\beta) = Q\left(\sqrt{\frac{2E_b \cos\beta}{N_0}}\right)$$

Теперь, если ошибка рассогласования по фазе β является результатом ошибок сопровождения, вызванных системным шумом, β будет стохастически описываться некоторой функцией плотности вероятности $p(\beta)$. Далее ожидаемая вероятность битовой ошибки дается следующим выражением.

$$P_B = \int_0^{2\pi} P_B(\beta) p(\beta)\, d\beta$$

Для частного случая контура первого порядка функция плотности вероятности описывается выражением (10.36). Следовательно, окончательное выражение для вероятности битовой ошибки выглядит следующим образом.

$$P_B = \int_0^{2\pi} Q\left(\sqrt{\frac{2E_b \cos\beta}{N_0}}\right) \frac{\exp(\rho\cos\beta)}{2\pi I_0(\rho)}\, d\beta$$

Отношение сигнал/шум в контуре (ρ_i), равное 20 дБ, будет соответствовать среднеквадратическому отклонению фазового шума $\sigma_{\hat{\theta}} = 0{,}1$ рад. Из рис. 10.8 видно, что этот небольшой фазовый шум не сильно ухудшает вероятность битовой ошибки. В то же время контур с $\rho_i = 10$ дБ соответствует среднеквадратическому отклонению фазового шума $\sigma_{\hat{\theta}} = 0{,}32$ рад. Из рис. 10.8 видно, что для вероятности битовой ошибки 10^{-5} это среднеквадратическое отклонение фазового шума потребует отношения сигнал/шум, несколько превышающего 11 (10,4 дБ), а не 9,1 (9,6 дБ), как при идеальном сопровождении фазы. Следовательно, данное отношение сигнал/шум в контуре приведет к росту требований более чем на 0,8 дБ при вероятности ошибки 10^{-5}. Следует отметить, что для отношений сигнал/шум, меньших 10 дБ, ухудшение происходит очень быстро. Поэтому при проектировании систем с остаточной несущей величины порядка 10 дБ обычно не рассматриваются. При описанных условиях лучше работают системы с подавлением несущей, не имеющие проблем с неустранимыми ошибками.

10.2.1.9. Методы анализа спектра

Рассмотренные выше методы относятся к классу *методов спектральной линии*. В данных методах основным при определении ошибок является либо использование существующей спектральной линии на несущей частоте, либо создание такой линии на несущей частоте или частоте, кратной несущей. Существует иной набор методов, особенно полезных при оценке или сопровождении частоты несущей, в котором используется форма спектра пропускания сигнала. Эти методы основаны на теории максимального правдоподобия [4], но они также привлекательны и в общих чертах будут описаны ниже.

Возможно, наиболее привлекательным методом этого класса является использование блока согласованных фильтров, каждый из которых согласовывается с ожидаемым

сигналом с определенным сдвигом несущей частоты. Подобный блок фильтров может реализовываться непосредственно или может быть реализован как операция взвешивания и сложения на выходе быстрого преобразования Фурье. В любом случае фильтр с максимальным выходом будет соотнесен со сдвигом частоты сигнала. Символически подобный детектор частоты показан на рис. 10.9. В зависимости от структуры сигнала и его чувствительности к отклонениям частоты, а также от плотности сдвигов частоты, в качестве прямой оценки частоты может быть принят наибольший выходной сигнал либо произведена дополнительная обработка для уточнения оценки. В любом случае очевидно, что блок фильтров, охватывающий диапазон возможных сдвигов частот, может быть спроектирован, и подобная схема будет давать быструю и надежную оценку сдвига несущей частоты.

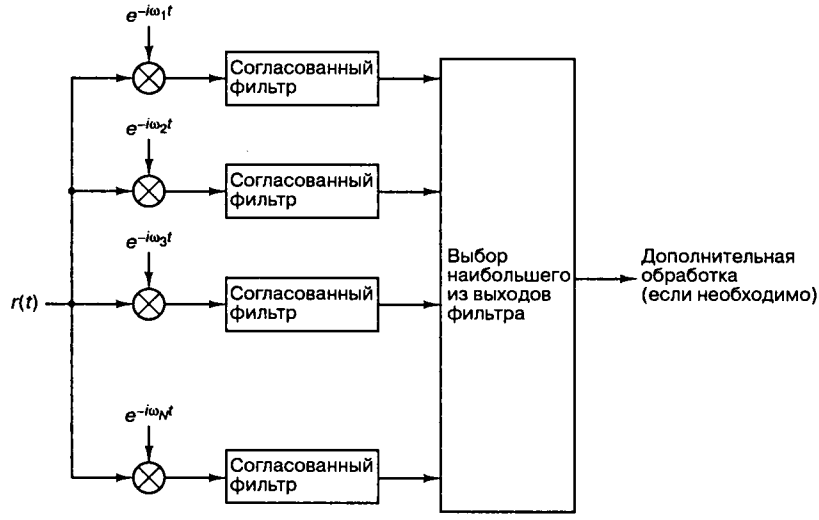


Рис. 10.9. Оценка частоты путем использования блока согласованных фильтров

Преимуществом рассмотренного выше подхода с использованием блока фильтров является возможность снижения неопределенности по частоте до любого требуемого значения. Недостаток заключается в неравномерности первоначальной оценки. Еще один спектральный метод, иногда называемый *фильтрацией краев полосы пропускания*, может давать значительно более точную оценку за счет снижения возможностей в определении неопределенности по частоте. Принцип работы метода легко понять с помощью графического примера.

На верхнем графике, приведенном на рис. 10.10, спектр полосового сигнала показан в виде широкой затененной области, центрированной на номинальной несущей частоте ω_0. Кроме того, там показаны два более узких полосовых фильтра, расположенных на краях спектра сигнала. Если (второй график) обнаруженный сигнал равен на обоих фильтрах, спектр сигнала будет центрирован между ними и ошибка по номинальной несущей частоте будет равна нулю. В то же время, если (третий и четвертый графики) спектр входного сигнала смещен относительно фильтров края полосы пропускания, то один фильтр будет иметь более обнаружимый сигнал, поэтому на основе данного отличия можно выработать меру ошибки. Эта мера может использовать-

ся для направления контура управления или же она может применяться непосредственно для вычисления требуемой коррекции частоты. Основным преимуществом методов этого типа является отсутствие необходимости в нелинейностях, вносящих дополнительный шум. Недостаток состоит в необходимости знаний о спектре сигнала и реализации двух узкополосных фильтров с идеально согласованными полосовыми характеристиками. Создать узкополосные, идеально согласованные фильтры может быть затруднительно (или дорого), если это выполняется на аналоговых схемах, но теоретически это можно легко сделать при использовании цифровых технологий.

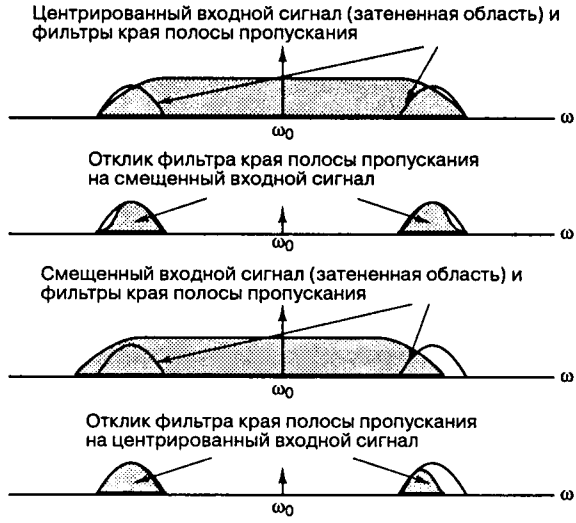


Рис. 10.10. Фильтр края полосы пропускания

10.2.2. Символьная синхронизация — модуляции дискретных символов

Для оптимальной демодуляции все цифровые приемники должны синхронизироваться с переходами поступающих цифровых символов. Ниже рассматривается несколько основных проектов символьных синхронизаторов. В центре обсуждения будет (для простоты записи и используемой терминологии) находиться случайный двоичный узкополосный сигнал, но расширение на недвоичные узкополосные сигналы должно быть очевидно.

При изложении материала в данном разделе предполагается, что о реальной информационной последовательности ничего не известно. Класс синхронизаторов, используемых в подобном случае, называется синхронизаторами без применения данных (non-data-aided — NDA). Существует еще один класс символьных синхронизаторов, которые используют известную информацию об информационном потоке. Эта информация может извлекаться из переданных по обратной связи решений относительно принятых данных или из введенной в информационный поток известной последовательности. В настоящее время более важными и доминирующими при выборе модуляций, эффективно использующих полосу, становятся методы с использованием данных (data-aided — DA). Эти методы рассматриваются в следующем разделе.

Рассматриваемые символьные синхронизаторы можно разделить на две основные группы. Первая группа состоит из разомкнутых синхронизаторов. Данные схемы выделяют копию выхода генератора тактовых импульсов передатчика непосредственно

из поступающего информационного потока. Вторая группа — это замкнутые синхронизаторы; они синхронизируют локальный генератор тактовых импульсов с поступающим сигналом посредством сличения локального и поступающего сигналов. Замкнутые синхронизаторы, как правило, точнее, но при этом сложнее и дороже.

10.2.2.1. Разомкнутые символьные синхронизаторы

Разомкнутые символьные синхронизаторы также иногда называют нелинейными синхронизаторами на фильтрах [20]; данное название говорит само за себя. Синхронизаторы этого класса генерируют частотный компонент со скоростью передачи символов, пропуская поступающий узкополосный сигнал через последовательность фильтра и нелинейного устройства. Работа данного устройства аналогична восстановлению несущей в контуре сопровождения с подавленной несущей. В данном случае желательный частотный компонент, передаваемый со скоростью передачи символов, изолируется с помощью полосового фильтра, после чего насыщающий усилитель с высоким коэффициентом насыщения придает ему нужную форму. В результате восстанавливается прямоугольный сигнал генератора тактовых импульсов.

На рис. 10.11 приведены три примера разомкнутых битовых синхронизаторов. В первом примере (рис. 10.11, *а*) поступающий сигнал $s(t)$ фильтруется с использованием согласованного фильтра. Выход этого фильтра — автокорреляционная функция исходного сигнала. Например, для передачи с помощью прямоугольных импульсов, на выходе имеем сигнал, состоящий из равнобедренных треугольников. Затем полученная последовательность спрямляется с помощью некоторой нелинейности четного порядка, например квадратичного устройства. Полученный сигнал будет содержать пики положительной амплитуды, которые, с точностью до временной задержки, соответствуют переходам входных символов. Последовательность описанных процессов изображена на рис. 10.12. Таким образом, сигнал с выхода четного устройства будет содержать Фурье-компонент на собственной частоте тактового генератора. Данная частотная составляющая изолируется от остальных гармоник с помощью полосового фильтра (bandpass filter — BPF), и ей придается форма посредством насыщающего усилителя с передаточной функцией следующего вида.

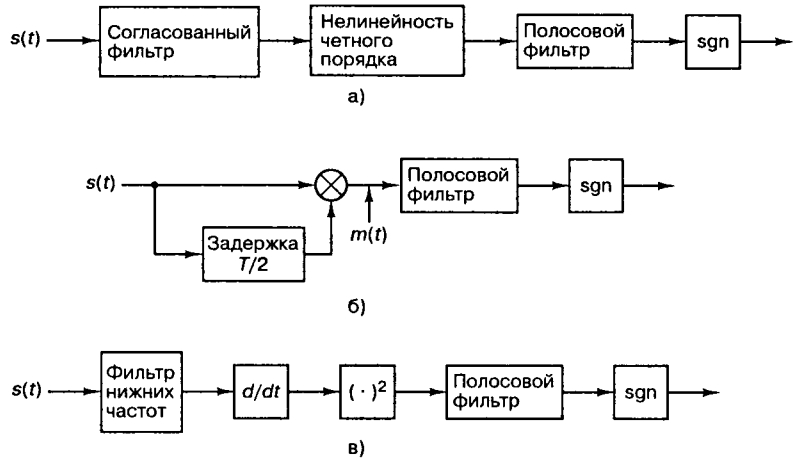


Рис. 10.11. Три типа разомкнутых битовых синхронизаторов

Рис. 10.12. Иллюстрация процессов, проходящих в разомкнутом битовом синхронизаторе

$$\operatorname{sgn} x = \begin{cases} 1 & \text{для } x > 0 \\ -1 & \text{для других } x \end{cases} \tag{10.53}$$

Во втором примере (рис. 10.11, *б*) Фурье-компонент на частоте тактового генератора создается посредством задержки и умножения. Длительность задержки, показанной на рис. 10.11, *б*, равна половине периода передачи бита, и это значение является оптимальным, поскольку оно дает наибольший Фурье-компонент [20]. Сигнал $m(t)$ всегда будет положительным во второй половине любого периода передачи бита, но будет иметь отрицательную первую половину, если во входном потоке битов $s(t)$ произошло изменение состояния. Это дает прямоугольный сигнал, спектральные компоненты и все гармоники которого совпадают с теми, что были у сигнала в схеме на рис. 10.11, *а*. Как и ранее, нужный спектральный компонент может быть отделен с помощью полосового фильтра, и ему будет придана нужная форма.

Последний пример (рис. 10.11, *в*) соответствует контурному детектору. Основными операциями здесь являются дифференцирование и спрямление (посредством использования квадратичного устройства). Если на вход поступает сигнал прямоугольной формы, дифференциатор дает положительные или отрицательные пики на всех переходах символов. При спрямлении получаемая последовательность положительных импульсов будет давать Фурье-компонент на скорости передачи информационных символов. Потенциальной проблемой данной схемы является то, что дифференциаторы обычно весьма чувствительны к широкополосному шуму. Это делает необходимым введение перед дифференциатором фильтра нижних частот (low-pass filter — LPF), как показано на рис. 10.11, *в*. В то же время данный фильтр удаляет высокочастотные составляющие информационных символов, что приводит к потере сигналом исходной прямоугольной формы. Это, в свою очередь, приводит к тому, что результирующий дифференциальный сигнал будет иметь конечные времена нарастания и спада и уже не будет последовательностью импульсов.

Очевидно, что с этапами обработки сигналов, изображенными на рис. 10.11, будет связана некоторая аппаратная задержка. В работе [12] показано, что для полосового фильтра, эффективно усредняющего K входных символов (ширина полосы = $1/KT$), величина среднего сбоя времени (задержки) приблизительно описывается следующим выражением.

$$\frac{\overline{|\varepsilon|}}{T} \approx \frac{0{,}33}{\sqrt{KE_b / N_0}} \quad \text{для } \frac{E_b}{N_0} > 5 \quad K \geq 18 \tag{10.54}$$

Здесь T — период передачи символа, E_b — обнаруженная энергия на бит, а N_0 — односторонняя спектральная плотность мощности принятого шума. Там же показано, что при высоких отношениях сигнал/шум отношение среднеквадратического отклонения временной ошибки дается следующим выражением.

$$\frac{\sigma_\varepsilon}{T} \approx \frac{0{,}411}{\sqrt{KE_b / N_0}} \text{ для } \frac{E_b}{N_0} > 1 \qquad (10.55)$$

Таким образом, если для данного полосового фильтра принятое отношение сигнал/шум достаточно велико, все методы, приведенные на рис. 10.11, приведут к точной битовой синхронизации.

10.2.2.2. Замкнутые символьные синхронизаторы

Основным недостатком разомкнутых символьных синхронизаторов является наличие неустранимой ошибки сопровождения с ненулевым средним. Эту ошибку можно снизить при больших отношениях сигнал/шум, но поскольку форма сигнала синхронизации зависит непосредственно от поступающего сигнала, устранить ошибку не удастся никогда.

Замкнутые символьные синхронизаторы сравнивают входной сигнал с локально генерируемым с последующей синхронизацией локального сигнала с переходами во входном сигнале. По сути, процедура ничем не отличается от используемой в разомкнутых синхронизаторах.

Среди наиболее популярных замкнутых символьных синхронизаторов можно выделить синхронизатор с опережающим и запаздывающим стробированием (early/late-gate synchronizer). Пример такого синхронизатора схематически изображен на рис. 10.13. Его работа заключается в выполнении двух отдельных интегрирований энергии входного сигнала по двум различным промежуткам символьного интервала длительностью $(T-d)$ секунд. Первое интегрирование (опережающее) начинается в момент, определенный как начало периода передачи символа (условно — момент времени 0), и заканчивается через $(T-d)$ секунд. Второе интегрирование (запаздывающее) начинается с задержкой на d секунд и заканчивается в конце периода передачи символа (условно — момент времени T). Разность абсолютных значений выходов описанных интеграторов y_1 и y_2 является мерой ошибки синхронизации символов приемника и может подаваться обратно для последующей коррекции приема.

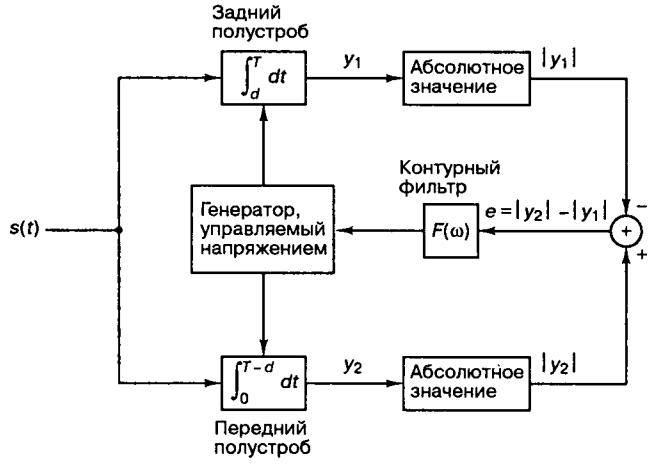


Рис. 10.13. Синхронизатор с опережающим и запаздывающим стробированием

Работа синхронизатора с опережающим и запаздывающим стробированием представлена на рис. 10.14. При идеальной синхронизации (рис. 10.14, *а*) показано, что оба периода стробирования попадают в интервал передачи символа. В этом случае оба интегратора получат одинаковый объем энергии сигнала и разность соответствующих сигналов (сигнал рассогласования *e* на рис. 10.13) будет равна нулю. Следовательно, если устройство синхронизировано, оно стабильно; нет тенденции к самопроизвольному выходу из синхронизации. На рис. 10.14, *б* показан пример для приемника, генератор тактовых импульсов которого функционирует с опережением по отношению к входному сигналу. В данном случае начало интервала опережающего интегрирования попадает на предыдущий интервал передачи бита, тогда как запаздывающее интегрирование по-прежнему выполняется в пределах текущего символа. При запаздывающем интегрировании энергия накапливается за интервал времени $(T - d)$, как и в случае, изображенном на рис. 10.14, *а*; но опережающее интегрирование накапливает энергию всего за время $[(T - d) - 2\Delta]$, где Δ — часть интервала опережающего интегрирования, приходящаяся на предыдущий интервал передачи бита. Следовательно, для этого случая сигнал рассогласования будет равен $e = -2\Delta$, что приведет к снижению входного напряжения ГУН на рис. 10.13. Это, в свою очередь, приведет к снижению выходной частоты ГУН и замедлит отсчет времени приемника для согласования с входными сигналами. Используя рис. 10.14 как образец, можно видеть, что если таймер приемника опаздывает, объемы энергии, накопленные при опережающем и запаздывающем интегрировании, будут обратны к полученным ранее и, соответственно, поменяется знак сигнала рассогласования. Таким образом, запаздывание таймера приемника приведет к увеличению напряжения ГУН, что вызовет увеличение выходной частоты генератора и приближение скорости таймера приемника к скорости входного сигнала.

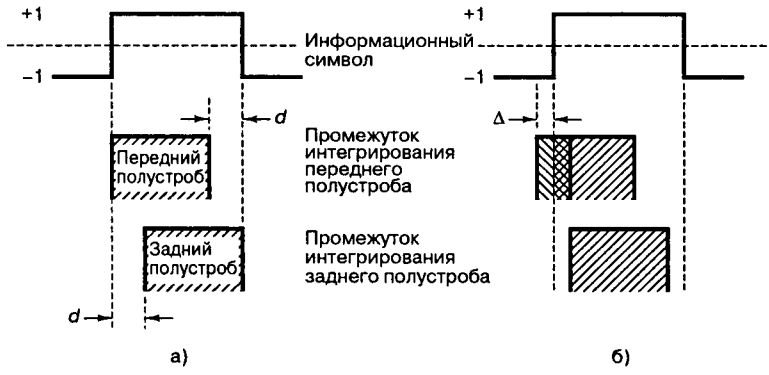


Рис. 10.14. Символьная синхронизация: а) точная синхронизация приемника; б) синхронизация с опережением

В примере, проиллюстрированном на рис. 10.14, неявно подразумевалось, что до и после рассматриваемого символа происходит изменение информационного состояния. Если переходов нет, можно видеть, что опережающее и запаздывающее интегрирование приведет к одинаковым результатам. Следовательно, если не происходит изменения информационного состояния, сигнал рассогласования не генерируется. Это всегда следует иметь в виду при использовании любых символьных синхронизаторов. Вернемся к рис. 10.13. Создать два абсолютно одина-

ковых интегратора невозможно. Следовательно, сигналы из двух ветвей контура будут сдвинуты относительно друг друга, даже если теоретически они должны быть идентичны. Данный сдвиг будет небольшим для качественно спроектированных интеграторов, но он приведет к постепенному уходу от синхронизации при наличии продолжительных последовательностей одинаковых информационных символов. Во избежание этого можно либо, что, вероятно, наиболее очевидно, форматировать данные так, чтобы гарантированно не было достаточно длительных интервалов без перехода, либо модифицировать структуру схемы таким образом, чтобы она содержала один интегратор. Примером структур такого типа является контур сглаживания, рассмотренный в связи с синхронизацией систем расширенного спектра в главе 12.

Еще один момент, связанный с проектированием контура, — это интервалы интегрирования. В примере, приведенном на рис. 10.14, интегрирование охватывает примерно три четверти периода передачи символа. В действительности величина этого интервала может быть от половины до практически всего периода передачи символа. Почему не меньше половины? Компромисс достигается между объемом проинтегрированного шума и интерференцией в стробе, с одной стороны, и длительностью сигнала, с другой. Как было справедливо для нелинейной модели контуров фазовой автоподстройки частоты, схемы этого типа трудно анализировать; определение производительности обычно выполняется с помощью компьютерного моделирования. Особенно это актуально для перекрывающихся интервалов интегрирования, подобных показанным на рис. 10.14, поскольку выборки шума в двух стробах будут коррелировать. Гарднер (Gardner) [5] показал, что для нормированного входного сигнала в 1 В, аддитивного белого гауссового шума, случайной последовательности данных (вероятность перехода $\frac{1}{2}$), опережающего и запаздывающего интегрирования, продолжительностью половина интервала передачи бита, и для больших отношений сигнал/шум в контуре относительное случайное смещение синхронизации приблизительно описывается следующим выражением.

$$\frac{\sigma_e^2}{T^2} = 2N_0B_L \tag{10.56}$$

Здесь N_0 — (нормированная) спектральная плотность мощности, T — интервал передачи символа, а B_L — ширина полосы контура.

10.2.2.3. Ошибки символьной синхронизации и вероятность символьной ошибки

Влияние ошибки символьной синхронизации на вероятность битовой ошибки для сигнала с модуляцией BPSK при аддитивном белом гауссовом шуме показано на рис. 10.15. Из графика видно, что для относительного случайного смещения синхронизации, меньшего 5%, ухудшение отношения сигнал/шум меньше 1 дБ. Сравнивая воздействие ошибки символьной синхронизации с влиянием фазового шума (см. рис. 10.8), видим, что ошибка символьной синхронизации, взятая относительно длительности передачи символа, не так сильно влияет на характеристики системы, как фазовый шум, взятый относительно цикла. Впрочем, в обоих случаях ухудшение характеристики повышается с ростом ошибки.

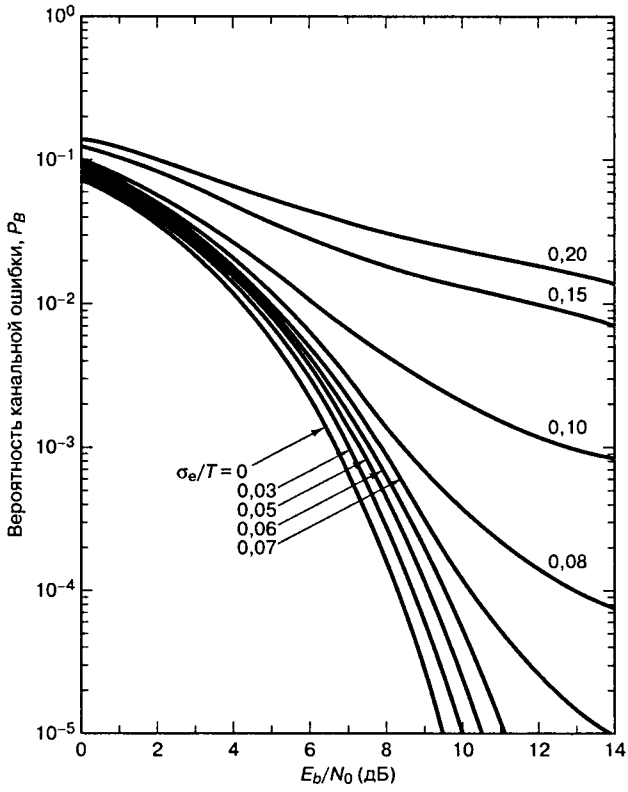


Рис. 10.15. Зависимость вероятности битовой ошибки от E_b/N_0 при использовании в качестве параметра среднеквадратического отклонения ошибки символьной синхронизации σ_ε. (Перепечатано с разрешения авторов из Lindsey W. C. and Simon M. K. Telecommunication Systems Engineering, *Prentice-Hall. Inc., Englewood Cliffs, N. J., 1973.)*

Пример 10.7. Влияние случайного смещения синхронизации

С помощью рис. 10.15 определите влияние 10%-ного случайного смещения синхронизации на систему, в которой требуется поддерживать вероятность ошибки 10^{-3}.

Решение

Из рис. 10.15 видно, что вероятность битовой ошибки 10^{-3} требует отношения SNR порядка 6,7 дБ при отсутствии любого случайного смещения синхронизации. Из того же рисунка видно, что при относительном случайном смещении синхронизации 10% ($\sigma_\varepsilon/T = 0{,}1$) необходимо отношение SNR порядка 12,9 дБ. Следовательно, способность выдерживать такое большое случайное смещение синхронизации потребует на 6,2 дБ большего отношения сигнал/шум, чем нужно для поддержания вероятности ошибки 10^{-3} при отсутствии случайного смещения. Данный пример показывает, как можно использовать графики, приведенные на рис. 10.15. В то же время на практике никакая система связи не будет проектироваться с четырехкратным запасом мощности для возможности работы при большой ошибке символьной синхронизации. В таких случаях обычно применяется другой подход, например перепроектирование системных фильтров с целью увеличения K в уравнении (10.55), что приведет к уменьшению случайного смещения символьной синхронизации.

10.2.3. Синхронизация при модуляциях без разрыва фазы

10.2.3.1. Основы

Модуляции без разрыва фазы (Continuous-Phase Modulation — CPM) появились при исследовании методов передачи сигналов, эффективно использующих полосу. По мере того как полоса становилась дороже, повышалась важность этих схем. С появлением этих модуляций возникли новые вопросы в области синхронизации, особенно символьной. Эффективность использования полосы схемой CPM достигается за счет сглаживания сигнала во временной области. Это сглаживание приведет к концентрации энергии сигнала в узкой полосе, что обеспечит уменьшение ширины полосы, требуемой для передачи сигнала, и размещение соседних сигналов плотнее друг к другу. В то же время, вследствие сглаживания сигнала во временной области, проявляется тенденция к уничтожению символьных переходов, от которых зависит работа множества схем синхронизации. Имеется и другая, родственная проблема — при использовании схемы CPM сложно различить последствия ошибки фазы несущей и ошибки символьной синхронизации, что делает взаимозависимыми задачи сопровождения фазы и синхронизации. В защиту сглаживания в схеме CPM говорит то, что в большинстве случаев, представляющих практический интерес, характеристики приемников относительно нечувствительны к средним ошибкам синхронизации [3].

В комплексной форме записи нормированный сигнал CPM имеет следующий вид.

$$s(t) = \exp\{i[\omega_0 t + \theta + \psi(t - \tau, \alpha)]\} \tag{10.57}$$

Здесь ω_0 — несущая частота, θ — фаза несущей (измеряемая относительно фазы приемника), а $\psi(t, \alpha)$ — *избыточная фаза* сигнала $s(t)$. Именно $\psi(t, \alpha)$ и является носителем информации сигнала. Кроме того, $\psi(t, \alpha)$ определяет, какая ширина полосы требуется сигналу; требуемая ширина полосы иногда называется *занятостью полосы* сигнала. При рассмотрении уменьшения или минимизации требуемой ширины полосы с точки зрения теории Фурье можно видеть, что компоненты относительно высокой частоты связаны с относительно резкими скачками сигнала во временной области [22]. Следовательно, для снижения или устранения высокочастотных компонентов следует сгладить все острые углы или резкие скачки сигнала во временной области. При передаче сигналов с использованием схемы CPM это выполняется путем объединения трех методов.

1. Использование сигнальных импульсов, имеющих непрерывные производные нескольких порядков.
2. Отдельным сигнальным импульсам разрешается занимать множественные интервалы передачи сигнала (т.е. намеренно вводится некоторая межсимвольная интерференция).
3. Снижение максимального разрешенного изменения фазы в символьном интервале.

Не все схемы CPM используют все перечисленные выше методы, но в каждой схеме применяются хотя бы некоторые из них. Для схем CPM следует отметить, что в начале каждого интервала передачи символа избыточная фаза $\psi(t, \alpha)$ является Марковским процессом [4], поскольку она зависит только от фазы в начале символа и значения текущего символа. Значение фазы в начале символа является следствием некоторого числа предыдущих символов. Следовательно, для частного случая конечного числа возможных состояний фазы получается канал с конечным числом состояний. Таким образом, избыточную фазу можно определить следующим образом.

$$\psi(t, \alpha) = \eta(t, \mathbf{C}_k, \alpha_k) + \Phi_k \quad kT \le t \le (k+1)T, \tag{10.58}$$

где

$$\eta(t, \mathbf{C}_k, \alpha_k) = 2\pi h \sum_{i=k-L+1}^{k} \alpha_i q(t - iT) \tag{10.59}$$

Здесь $\mathbf{C}_k$ — корреляционное состояние, k — временной индекс, а α_k — k-й информационный символ, взятый из алфавита $\{\alpha_k\} = \{\pm 1, \pm 3, \ldots, \pm(M-1)\}$. Данный алфавит в общем случае допускает M-арную (а не только бинарную) передачу сигналов. Параметр h — коэффициент модуляции, а $q(t)$ — *фазовая характеристика модуляции*, которая определяется вне области $0 < t < LT$ следующим образом.

$$q(t) = \begin{cases} 0 & \text{для } t \le 0 \\ 1/2 & \text{для } t \ge LT \end{cases} \tag{10.60}$$

В данном случае L является радиусом корреляции. Радиус корреляции — это число периодов передачи информационных символов, длительностью T секунд, на которые влияет отдельный информационный символ. Это мера объема умышленной межсимвольной интерференции. При $L = 1$ говорят, что передача сигналов идет *с полным откликом*. При обсуждении модуляции в предыдущих главах предполагался именно такой тип передачи. При этом каждый импульс замкнут в собственных временных рамках. В то же время при $L > 1$ говорят, что передача сигналов производится *с частичным откликом*. Это означает, что каждый импульс не ограничен собственным интервалом, а "размыт" на $L-1$ соседний интервал передачи символа. Этот тип передачи применяется во многих схемах CPM для умышленного введения управляемой межсимвольной интерференции, что приводит к увеличению эффективности использования полосы. Одна из ранних схем CPM, классическая манипуляция с минимальным сдвигом (minimum-shift-keying — MSK) (см. главу 9), не использует множественные интервалы передачи символа на импульс. Следовательно, классическая схема MSK — это пример передачи сигналов с полным откликом. Изучая уравнение (10.60), можно заметить, что при $q(LT) = \frac{1}{2}$ максимальное возможное изменение фазы на промежутке LT равно $(M-1)\pi h$, как можно видеть из уравнений (10.58) и (10.59).

Вектор $\mathbf{C}_k$, называемый *корреляционным состоянием*, представляет собой последовательность информационных символов $\{\alpha_k\}$, начинающихся с наиболее раннего момента, когда возможно влияние на фазу сигнала в текущий момент времени k.

$$\mathbf{C}_k = (\alpha_{k-L+1}, \ldots, \alpha_{k-2}, \alpha_{k-1})$$

Член Φ_k в уравнении (10.58) называется *фазовым состоянием* и выражается следующим образом.

$$\Phi_k = \pi h \sum_{i=0}^{k-L} \alpha_i \quad \text{по модулю } 2\pi \tag{10.61}$$

Фазовое состояние — это одна из набора дискретных фаз, которые может иметь сигнал при данных значениях предыдущих символов. Необходимое условие непрерывности фазы заключается в следующем: фаза должна переходить в следующий символ

только с фазового состояния. В контексте решетчатой диаграммы Φ_k можно рассматривать как исходное состояние или узел, а $\mathbf{C}_k$ — как определение пути к одному из других узлов. Характеристики любой модуляции определяются $q(t)$ в интервале ($0 < t < LT$). Схема MSK имеет следующие параметры: $h = \frac{1}{2}$, $L = 1$, $M = 2$ и $q(t) = t/(2T)$ в промежутке ($0 < t < T$). Частотная характеристика, определяемая как $g(t) \stackrel{\text{def}}{=} \frac{dq(t)}{dt}$, имеет для схемы MSK прямоугольную форму.

$$g(t) = \begin{cases} 1/2T & 0 \le t \le T \\ 0 & t < 0, t > T \end{cases} \tag{10.62}$$

Гауссова манипуляция с минимальным частотным сдвигом (Gaussian MSK — GMSK) — еще один пример схемы CPM — определяется как схема, частотная характеристика которой является сверткой описанного выше прямоугольника с гауссоидой.

Многие способы синхронизации, описанные в предыдущих разделах, основаны на *специально разработанных* методах. Большинство этих методов понятно интуитивно. К сожалению, за исключением нескольких случаев, для схемы CPM не существует подобных интуитивных подходов. Здесь большинство методов основано на принципах классической теории оценок, причем наиболее популярной была оценка по методу максимального правдоподобия. Принципы, использованные в этих случаях, аналогичны разработанным для обнаружения сигнала по методу максимального правдоподобия.

Оценка по методу максимального правдоподобия, основанная на теории Байеса [7], включает максимизацию условных вероятностей. Пусть $s(t, \gamma)$ представляет сигнал с набором неизвестных параметров γ. Параметрами могут быть: фаза несущей, значение смещения символьной синхронизации, значения переданных информационных символов или, возможно, другие параметры. Пусть

$$r(t) = s(t, \gamma) + n(t) \tag{10.63}$$

представляет принятый сигнал, где $n(t)$ — некоторый аддитивный шум приемника. Допустим, $R(t)$ — это реализация процесса $r(t)$. Тогда оценкой по методу максимального правдоподобия для набора неизвестных параметров γ является значение γ, максимизирующее правдоподобие $p[r(t) = R(t)|\gamma]$ по всем γ. Как показывалось в главе 3, для известного сигнала реализация детектора, работающего по принципу максимального правдоподобия, — это фильтр, согласованный с этим сигналом. Для схем CPM это решение приводит к структуре, изображенной на рис. 10.16

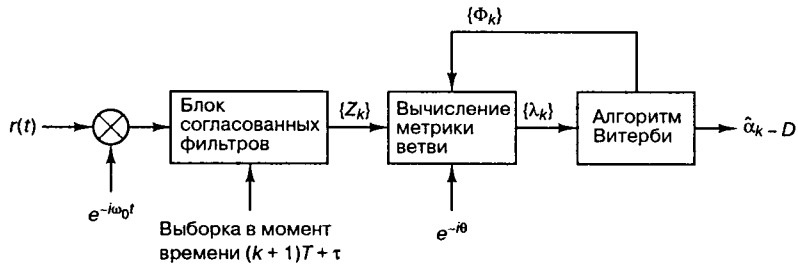


Рис. 10.16. Приемник схемы CPM. (Примечание: $\hat{\alpha}_{k-D}$ — это k-й выходной символ с задержкой вследствие обработки D.)

При первичном обнаружении сигнала частота несущей ω_0, фаза несущей θ и сбой символьной синхронизации τ предполагаются известными. Принимающая структура — это блок согласованных фильтров, каждый из которых согласован с L-символьной реализацией сигнала, после чего следует аппаратная реализация алгоритма Витерби. Число фильтров равно M^L, а число узлов в вычислении метрики ветви — PM^{L-1}, где P — число фазовых состояний $\{\Phi_k\}$. Эти числа могут быть достаточно большими, что может создавать неудобства при реализации, поэтому на практике обычно используются более простые приемники [3, 4, 22]; впрочем, в качестве основы синхронизации данная структура все же является полезной.

Используя приведенное выше описание схемы CPM, запишем импульсную характеристику отдельных фильтров блока.

$$h^{(l)}(t) \stackrel{\text{def}}{=} \begin{cases} e^{-i\eta_l(T-t,\mathbf{C}_0^{(l)}\alpha_0^{(l)})} & 0 \le t \le T \\ 0 & \text{для других } t \end{cases} \tag{10.64}$$

Здесь через ($l = 1, 2, ..., M^L$) обозначена L-символьная строка $(\mathbf{C}_0^{(l)}, \alpha_0^{(l)}) = (\alpha_{-L+1}^{(l)}, ..., \alpha_{-1}^{(l)}, \alpha_0^{(l)})$, причем каждое $\alpha_k^{(l)}$ выбирается из алфавита сигналов, а l обозначает конкретный путь (последовательность символов) во множестве M^L возможных путей. Согласно использованной ранее форме записи, получаем следующее.

$$\eta_l(t, \mathbf{C}_0^{(l)}, \alpha_0^{(l)}) = 2\pi h \sum_{i=-L+1}^{0} \alpha_i^{(l)} q(t - iT) \tag{10.65}$$

Из рис. 10.16 видно, что выход отдельного фильтра описывается следующим выражением.

$$Z_k^{(l)}(\mathbf{C}_k, \alpha_k, \tau) \stackrel{\text{def}}{=} \int_{\tau + kT}^{\tau + (k+1)T} r(t) h^{(l)}(t - \tau - kT) e^{-i\omega_0 t} dt \tag{10.66}$$

Данный набор выходов $\{Z_k\}$, оценка фазы несущей $\hat{\theta}$ и фазовое состояние $\{\Phi_k\}$ используются для вычисления метрики пути и, в конечном итоге, решения на выходе алгоритма Витерби.

10.2.3.2. Синхронизация с использованием данных

Методы синхронизации приемников CPM можно разделить на зависящие и независящие от знаний об информационных символах. Первые называются методами с использованием данных (data-aided — DA), вторые — методами без использования данных (non-data-aided — NDA). Очевидно, что подобное разделение методов можно применить ко всем модуляциям, но методы с использованием данных особенно полезны и популярны при схеме CPM. Существует два пути получения знаний об информационных символах: либо рассматриваемый символ является частью известного заголовка или настроечной последовательности, введенных в информационный поток, либо решения с выхода алгоритма Витерби по обратной связи возвращаются на вход процесса синхронизации. Если обратная связь по принятию решения не реализуется, очевидно, решения должны быть весьма надежными; следовательно, приемник должен быть весьма близок к синхронизации.

Если считать, что за некоторый промежуток наблюдения L_0 известен поток переданных символов, индекс l в уравнении (10.66) можно опустить. Если принять обычные предположения — гауссов процесс шума, сигналы с равными энергиями — функция правдоподобия $\Lambda(R|\hat{\theta},\hat{\tau})$, связанная с θ и τ неизвестным сдвигом фазы и неизвестным сдвигом времени, выражается следующим образом [3].

$$\Lambda(R\,|\,\hat{\theta},\hat{\tau}) = \exp\left\{\sum_{k=0}^{L_0-1} \mathrm{Re}\left[Z_k(\mathbf{C}_k,\alpha_k,\hat{\tau})e^{-i(\hat{\theta}+\Phi_k)}\right]\right\} \tag{10.67}$$

Здесь были опущены несущественные постоянные множители, а $\mathrm{Re}\{\cdot\}$ обозначает действительную часть комплексного аргумента. Очевидно, что правая часть выражения (10.67) достигает максимума при максимальном значении суммы. Следовательно, если взять от суммы частные производные по $\hat{\theta}$ и $\hat{\tau}$ и приравнять результаты к нулю, получим следующие соотношения.

$$\sum_{k=0}^{L_0-1} \mathrm{Im}\left[Z_k(\mathbf{C}_k,\alpha_k,\hat{\tau})e^{-i(\hat{\theta}+\Phi_k)}\right] = 0 \tag{10.68}$$

и

$$\sum_{k=0}^{L_0-1} \mathrm{Re}\left[Z_k(\mathbf{C}_k,\alpha_k,\hat{\tau})e^{-i(\hat{\theta}+\Phi_k)}\right] = 0 \tag{10.69}$$

Здесь $Y_k = \partial Z_k / \partial\hat{\tau}$, а $\mathrm{Im}\{\cdot\}$ обозначает мнимую часть комплексного аргумента. В работе [3] показано, что левую часть уравнения (10.69) можно получить двумя способами: либо путем взятия производной "в лоб", либо посредством реализации набора "дифференцирующих фильтров". В каждом конкретном случае выбирается наиболее предпочтительный вариант.

К сожалению, уравнения (10.68) и (10.69) не имеют какого-либо интуитивного решения; кроме того, не существует известных аналитических решений. Уравнения приходится решать численно, используя некоторую итеративную процедуру для $\hat{\theta}$ и $\hat{\tau}$. В той же работе [3] предложена итеративная процедура, где последовательные члены каждой суммы используются для создания членов ошибки последовательных приближений.

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \gamma_P l_P(k-1) \tag{10.70}$$

$$\hat{\tau}_{k+1} = \hat{\tau}_k + \gamma_T l_T(k-1) \tag{10.71}$$

Здесь l_P и l_T — члены старшего порядка левых частей уравнений (10.68) и (10.69), а γ_P и γ_T — "коэффициенты усиления", которые выбираются для обеспечения сходимости процесса. Очевидно, данную итеративную процедуру проще реализовать с помощью обратной связи по принятию решения, чем посредством настроечной последовательности фиксированного размера.

10.2.3.3. Синхронизация без использования данных

Один из первых принципов теории информации заключается в том, что иметь больше информации лучше, чем иметь меньше. В контексте текущего обсуждения это

означает, что знание последовательности символов позволяет лучше оценить фазу несущей и символьную синхронизацию. Впрочем, возможны варианты, когда использование настроечной последовательности непрактично или неудобно и процесс принятия решения не достаточно надежен для организации обратной связи. В этих случаях применяется процесс синхронизации без использования данных (non-data-aided — NDA). Ниже будут рассмотрены два универсальных метода и один степенной метод, который может использоваться во многих случаях.

Первый метод — это прямое развитие метода, описанного в предыдущем разделе. Очевидно, если последовательность символов (C_k, α_k) неизвестна, новую функцию правдоподобия, подобную приведенной в уравнении (10.67), можно записать следующим образом.

$$\Lambda(R\,|\,\hat{C}_k, \hat{\alpha}_k \hat{\theta}, \hat{\tau}) = \exp\left\{\sum_{k=0}^{L_0-1} \mathrm{Re}\left[Z_k(\hat{C}_k, \hat{\alpha}_k, \hat{\tau}) e^{-i(\hat{\theta}+\Phi_k)}\right]\right\} \tag{10.72}$$

Поскольку функция правдоподобия пропорциональна условной вероятности, к выражению функции правдоподобия, зависящей от $\hat{\tau}$ и $\hat{\theta}$, можно применить цепное правило условных вероятностей, которое утверждает следующее [7].

$$p(r(t) = R(t)\,|\,\gamma) = \int_{\text{по всем } \beta} p[r(t) = R(t)\,|\,\gamma, \beta]\, p(\beta)\, d\beta \tag{10.73}$$

Из этого вытекает, что искомая функция правдоподобия имеет следующий вид.

$$\Lambda'(R\,|\,\hat{\theta}, \hat{\tau}) = \frac{1}{M^L} \sum_{\text{по всем } (\hat{C}_k, \hat{\alpha}_k)} \Lambda(R\,|\,\hat{C}_k, \hat{\alpha}_k, \hat{\theta}, \hat{\tau}) \tag{10.74}$$

Здесь было сделано предположение о равновероятности всех последовательностей символов. Функцию правдоподобия в правой части уравнения (10.74) теперь можно продифференцировать, в результате чего получим два уравнения, аналогичные (10.68) и (10.69). Очевидно, данный результат вычислить значительно сложнее, чем полученный в уравнениях (10.68) и (10.69). В работе [3] рассмотрены некоторые аппроксимации, которые дают несколько более простую оценку $\hat{\tau}$.

Второй метод основан на использовании (близкой к оптимальной) структуры приемника с фильтрами Лорана [23, 24]. В данной ситуации сигнал CPM аппроксимируется набором налагающихся сигналов с импульсно-кодовой модуляцией (pulse code modulation — PCM). При рассмотрении первого члена этого ряда получим следующее выражение.

$$e^{i\psi(t,\alpha)} \approx \sum_i a_{0,i} h_0(t - iT) \tag{10.75}$$

Здесь $\psi(t, \alpha)$ определено в уравнении (10.58), а коэффициенты $a_{0,i}$ являются *псевдосимволами*. Псевдосимволы, значения которых зависят от предыдущего и последующего информационных символов, определяются следующим образом.

$$\alpha_{0,i} = \exp\left(i\pi h \sum_{l=0}^{i} \alpha_l\right) \tag{10.76}$$

Здесь коэффициент модуляции h может иметь любое неотрицательное значение. Для важного частного случая модуляции MSK, где $h=\frac{1}{2}$, выражение (10.75) точно совпадает с функцией фильтра, имеющей следующий вид.

$$h_0(t)=\begin{cases}\sin\left(\dfrac{\pi t}{2T}\right) & 0\le t\le 2T\\ 0 & \text{для других } t\end{cases} \tag{10.77}$$

Для других модуляций аппроксимация может быть более или менее точной, и $h_0(t)$ будет иметь иной вид [23]. В любом случае, не учитывая пока процесс шума, можно записать нормированный сигнал в следующем виде.

$$s(t)\approx e^{i(\omega_0 t+\theta)}\sum a_{0,i}h_0(t-iT-\tau) \tag{10.78}$$

Из данного выражения очевидно, что стандартные методы фазовой и символьной синхронизации, разработанные в предыдущих разделах для линейных модуляций, могут применяться и к данной аппроксимации. В работе [3] подчеркивалось, что при использовании этого подхода следует быть очень внимательным, поскольку фильтр, в действительности согласованный с $h_0(t)$, может давать импульс очень плохой формы. Подробно этот вопрос рассмотрен в работе [25].

И последнее, в частных случаях, когда коэффициент модуляции является рациональным, $h=k_1/k_2$, где (k_1, k_2) — целые, может применяться степенной метод [22]. В этом случае уравнение (10.57) можно переписать следующим образом.

$$s(t)=\exp\left\{i\left[\omega_0 t+\theta+2\pi\frac{k_1}{k_2}\sum_{i=k-L+1}^{k}\alpha_i q(t-iT)\right]\right\} \tag{10.79}$$

Здесь для простоты, Φ_k из уравнения (10.58) было включено в θ. Возведение $s(t)$ в степень k_2 дает следующее.

$$[s(t)]^{k_2}=\exp\left\{i\left[k_2(\omega_0 t+\theta)+2\pi k_1\sum_{i=k-L+1}^{k}\alpha_i q(t-iT)\right]\right\} \tag{10.80}$$

Член $\omega_0 t+\theta$ в правой части, очевидно, является высокочастотным и будет отфильтрован. Крайний правый член — это k_1-я степень информационной части сигнала. Из уравнений (10.57)–(10.60) видно, что данный последний член повторяется с периодом, не превышающим LT. В зависимости от точной природы фазовой характеристики $q(t)$, могут создаваться компоненты ряда Фурье, кратные $2\pi k_1/(LT)$ радиан. По крайнем мере, теоретически эти компоненты можно отделить и отследить. Даже если спектральные линии недоступны, но можно отделить спектр, кратный истинному спектру сигнала, то для оценки частоты, кратной скорости передачи символов, могут применяться методы фильтрации краев полосы пропускания (описанные в разделе 10.2.1.9). Фазовый член θk_2 также можно отделить. При использовании данной процедуры возникает несколько практических проблем. Период передачи символов будет иметь (k_1/L)-альтернативную неопределенность, а оценка фазы — k_2-альтернативную неопределенность, которые

нужно как-то разрешить. В зависимости от природы $q(t)$, Фурье-компоненты могут быть достаточно слабыми и могут быть расположены близко друг к другу, что затрудняет их обособление. И последнее, как и для всех степенных методов, шум приемника растет непропорционально, возможно, снижая эффективное отношение сигнал/шум детектора до непригодного для использования уровня. Этот метод не имеет такого преимущества, как возможность использования какого-либо интуитивного решения. Он предлагает прямое соединение с методами спектральных линий, рассмотренными ранее. В данных методах для восстановления чистой спектральной линии на интересующей частоте или на известной частоте, кратной несущей, применяются нелинейности — обычно степенные устройства. Тот же подход использован и здесь. Предполагаемая рациональная природа коэффициента модуляции h используется для создания спектральных линий на частотах, кратных скорости передачи символов и несущей частоте. Данные линии могут применяться для получения и поддержания символьной синхронизации, а также для сопровождения частоты и фазы несущей.

10.2.4. Кадровая синхронизация

Практически все потоки цифровых данных имеют некоторую кадровую структуру. Другими словами, поток данных разбит на равные группы бит. Если поток данных — это оцифрованный телесигнал, каждый пиксель в нем представляется словом из нескольких бит, которые группируются в горизонтальные растровые развертки, а затем в вертикальные растровые развертки. Компьютерные данные обычно разбиваются на слова, состоящие из некоторого числа 8-битовых байт, которые, в свою очередь, группируются в образы перфокарт, пакеты, кадры или файлы. Любая система, использующая кодирование с защитой от блочных ошибок, в качестве основы кадра должна брать длину кодового слова. Оцифрованная речь обычно передается пакетами или кадрами, неотличимыми от других цифровых данных.

Чтобы входной поток данных имел смысл для приемника, приемник должен синхронизироваться с кадровой структурой потока данных. Кадровая синхронизация обычно выполняется с помощью некоторой специальной процедуры передатчика. Данная процедура может быть как простой, так и довольно сложной, в зависимости от среды, в которой должна функционировать система.

Вероятно, простейшим методом, используемым для облегчения кадровой синхронизации, является введение маркера (рис. 10.17). Маркер кадра — это отдельный бит или краткая последовательность бит, периодически вводимая передатчиком в поток данных. Приемник должен знать эту последовательность и период ее введения. Приемник, достигший синхронизации данных, сопоставляет (проверяет корреляцию) эту известную последовательность с потоком поступающих данных в течение известного периода введения. Если приемник не синхронизирован с кадровой последовательностью, корреляция будет слабой. При синхронизации приемника с кадровой структурой, корреляция будет практически идеальной, повредить которую может только случайная ошибка обнаружения.

Преимуществом маркера кадра является его простота. Для маркера может быть достаточно даже одного бита, если перед принятием решения, находится ли система в состоянии кадровой синхронизации, было выполнено достаточное число корреляций. Основной недостаток состоит в том, что данное достаточное число может быть очень большим; следовательно, большим может быть и время, требуемое для достижения синхронизации. Таким образом, наибольшую пользу маркеры кадров представляют в системах, непрерывно

передающих данные, подобно многим телефонным и компьютерным каналам связи, и не подходят для систем, передающих отдельные пакеты, или систем, требующих быстрого получения кадровой синхронизации. Еще одним недостатком маркера кадра является то, что введенный бит (биты) может повысить громоздкость структуры потока данных.

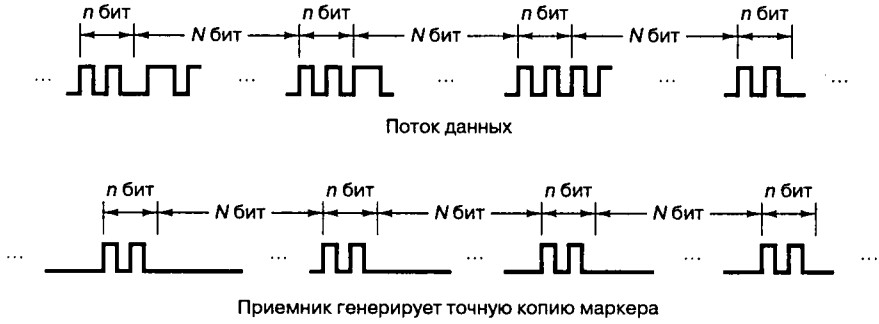


Рис. 10.17. Использование маркера кадра

В качестве примера можно привести линию T1, разработанную Bell Labs и широко используемую в североамериканских телефонных системах. Структура T1 включает использование маркера кадра размером 1 бит, вводимого после каждого набора из 24 8-битовых байт (каждый байт представляет один из 24 возможных потоков речевой информации). При таком подходе возникает информационная структура, кратная 193 бит, — неудобное число с точки зрения большинства интегральных схем.

В системах с неустойчивыми или пульсирующими передачами либо в системах с необходимостью быстрого получения синхронизации рекомендуется использовать синхронизирующие кодовые слова. Обычно такие кодовые слова передаются как часть заголовка сообщения. Приемник должен знать кодовое слово и постоянно искать его в потоке данных, возможно, используя для этого коррелятор на согласованных фильтрах. Обнаружение кодового слова укажет известную позицию (обычно — начало) информационного кадра. Преимуществом подобной системы является то, что кадровая синхронизация может достигаться практически мгновенно. Единственная задержка — отслеживание кодового слова. Недостаток — кодовое слово, выбираемое для сохранения низкой вероятности ложных обнаружений, может быть длинным, по сравнению с маркером кадра. Здесь стоит отметить, что сложность определения корреляции пропорциональна длине последовательности, поэтому при использовании кодового слова коррелятор может быть относительно сложным.

Хорошим синхронизирующим кодовым словом является то, которое имеет малое абсолютное значение "побочных максимумов корреляции". Побочный максимум корреляции — это значение корреляции кодового слова с собственной смещенной версией. Следовательно, данное значение побочного максимума корреляции для сдвига на k символов N-битовой кодовой последовательности $\{X_i\}$ описывается следующим выражением.

$$C_k = \sum_{j=1}^{N-k} X_j X_{j+k} \tag{10.81}$$

Здесь X_i $(1 \le i \le N)$ — отдельный кодовый символ, принимающий значения ±1, а соседние информационные символы (соотнесенные со значениями индекса $i > N$) предпо-

лагаются равными нулю. Пример вычисления побочного максимума корреляционной функции приведен на рис. 10.18. 5-битовая последовательность в данном примере имеет неплохие корреляционные свойства: наибольший побочный максимум в пять раз меньше основного, C_0. Последовательности, в которых, как на рис. 10.18, максимальный побочный максимум равен 1, называются последовательностями или словами Баркера (Barker word) [26]. Не существует известного конструктивного метода поиска слов Баркера, и в настоящее время известно всего 10 уникальных слов, наибольшее из которых состоит из 13 символов. Известные слова Баркера перечислены в табл. 10.1. После небольшого размышления становится понятно, что исчерпывающий перечень известных слов будет включать последовательности, порождаемые инверсией знака символов, и последовательности, порождаемые изменением направления хода времени в последовательностях символов, приведенных в табл. 10.1

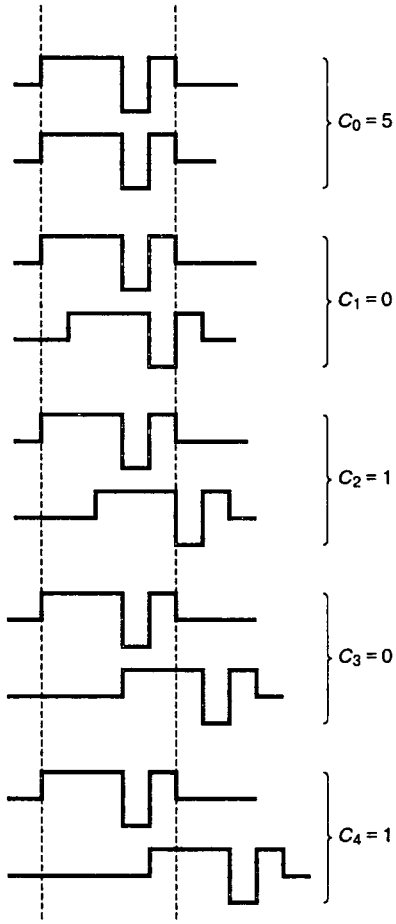


Рис. 10.18. Вычисление побочного максимума корреляционной функции

Таблица 10.1. Синхронизирующие кодовые слова Баркера

N	Последовательности Баркера
1	+
2	+ + или + –
3	+ + –
4	+ + + – или + + – +
5	+ + + – +
7	+ + + – – + –
11	+ + + – – – + – – + –
13	+ + + + + – – + + + – +

Свойства кодовых слов Баркера основываются на предположении о нулевом значении соседних символов. Это приближение к равновероятным случайным двоичным данным, когда символы, соседствующие со словом Баркера, принимают значения ±1. К сожалению, последовательности Баркера слишком коротки, чтобы это приближение во всех случаях давало лучшее кодовое слово при передаче случайной двоичной информации. Используя компьютерное моделирование, Уиллард (Willard) [27] нашел последовательности для случайных соседних символов, имеющие такую же длину, как и слова Баркера, но лучшие с точки зрения минимальной вероятности ложной синхронизации. Последовательности Уилларда приведены в табл. 10.2.

Таблица 10.2. Синхронизирующие кодовые слова Уилларда

N	Последовательности Уилларда
1	+
2	+ –
3	+ + –
4	+ + – –
5	+ + – + –
7	+ + + – + – –
11	+ + + – + + – + – – –
13	+ + + + + – – + – + – – –

Система, использующая синхронизирующее слово, описывается двумя вероятностями — вероятность пропущенного обнаружения и вероятность ложной тревоги. Очевидно, разработчик системы должен максимально уменьшить обе вероятности. К сожалению, это требование противоречиво. Для того чтобы уменьшить вероятность пропуска, система должна допускать неидеальную корреляцию входного синхронизирующего слова, т.е. слово должно приниматься даже в том случае, если оно содержит небольшое число ошибок. В то же время это увеличивает число последовательностей символов, которые будут приняты; следовательно, увеличивается вероятность ложной тревоги. Вероятность пропуска для N-битового слова, если допустимым является k или меньше ошибок, описывается следующим выражением.

$$P_m = \sum_{j=k+1}^{N} \binom{N}{j} p^j (1-p)^{N-j} \tag{10.82}$$

Здесь p — вероятность битовой ошибки, допущенной детектором. Вероятность ложной тревоги, вызванная N битами случайной последовательности данных, описывается следующим выражением.

$$P_{FA} = \sum_{j=0}^{k} \frac{\binom{N}{j}}{2^N} \tag{10.83}$$

Видно, что при малых p P_m растет с увеличением k, приблизительно как степенная функция. К сожалению, с увеличением k P_{FA} уменьшается, приблизительно как степенная функция. Для одновременного получения приемлемых значений P_m и P_{FA} при данном значении p разработчику системы часто требуются значения N, большие тех, которые дают последовательности Баркера и Уилларда. К счастью, в литературе приводится довольно много примеров подходящих длинных последовательностей. Большинство из них было найдено в результате обстоятельного компьютерного поиска. Спилкер (Spilker) [20] перечисляет последовательности с N до 24, найденные Ньюманом (Newman) и Хофманом (Hofman) [28], и упоминает, что в их оригинальной работе указаны последовательности с N до 100. By (Wu) [29] дает перечень последовательностей Мори-Стайлза (Maury-Styles) длиной до $N = 30$ и перечень последовательностей Линдера (Linder) длиной до 40. Кроме того, он приводит довольно полное обсуждение синхронизирующих последовательностей, в том числе конструктивных методов нахождения разумных, но неоптимальных последовательностей, а также рассматривает процедуры кадровой синхронизации некоторых спутниковых систем цифровой связи.

10.3. Сетевая синхронизация

Для систем, использующих методы когерентной модуляции, одностороннюю связь, такую как в широковещательных каналах, или одноканальную связь, как в большинстве микроволновых или оптоволоконных систем, оптимальный подход — это возложить все задачи синхронизации на приемник. Для систем связи, использующих методы некогерентной модуляции, или систем, где множество пользователей получают доступ к одному центральному узлу, например во многих системах спутниковой связи, функцию синхронизации часто имеет смысл возложить (полностью или частично) на оконечные устройства. Это означает, что для получения синхронизации модифицируются параметры передатчиков оконечных устройств, а не приемника центрального узла. Этот подход применим в системах, использующих множественный доступ с временным разделением (time-division multiple access — TDMA). В схеме TDMA каждому пользователю выделяется сегмент времени, в течение которого он может передавать информацию. Передатчик оконечного устройства должен синхронизироваться с системой, чтобы переданные им пакеты данных прибывали на центральный узел в тот момент, когда узел готов принимать данные. Синхронизация передатчика также имеет смысл в системах, объединяющих обработку сигналов на центральном узле с множественным доступом с частотным разделением (frequency-division multiple access — FDMA). Если оконечные устройства предварительно синхронизируют свои передачи с центральным узлом, узел может использовать конечный набор фильтров каналов и единое эталонное время для обслуживания всех каналов. В противном случае узел будет требовать возможности захвата и сопровождения длительности и частоты каждого

входного сигнала; кроме того, придется учитывать возможность интерференции сигналов из соседних каналов. Очевидно, что синхронизация передатчика оконечного устройства является более разумным решением при синхронизации сети.

Процедуру синхронизации передатчика можно отнести либо к открытой (без обратной связи), либо к замкнутой (с обратной связью). Открытые методы не зависят от измерения каких-либо параметров сигнала на центральном узле. Оконечное устройство заранее регулирует свою передачу, используя для этого знания о параметрах канала, которые предоставляются извне, но, возможно, могут модифицироваться при наблюдениях сигнала, приходящего с центрального узла. Открытые методы зависят от точности и предсказуемости параметров канала связи. Лучше всего их применять в системах с практически фиксированной архитектурой, где каналы непрерывно проработали достаточно длительный промежуток времени после установки/настройки. Эти методы достаточно трудно использовать эффективно, если геометрия канала связи не является статической или оконечные устройства нерегулярно получают доступ к системе.

Основными преимуществами открытых методов является быстрое получение синхронизации (метод может работать без обратного канала связи) и малый объем требуемых вычислений в реальном времени. Недостаток состоит в том, что требуется наличие внешнего источника знаний о требуемых параметрах канала связи; кроме того, этот источник должен быть относительно неизменным. Отсутствие каких бы то ни было измерений характеристик системы в реальном времени означает, что система не может быстро приспособиться к любому незапланированному изменению условий.

С другой стороны, замкнутые методы требуют незначительных *априорных* знаний о параметрах канала; эти знания помогут снизить время, требуемое для достижения синхронности, но они не обязательно должны быть такими точными, как в случае открытых методов. Замкнутые методы включают измерения точности синхронизации передач от оконечных устройств, поступающих на центральный узел, и возврата результатов этих измерений посредством обратного канала связи. Таким образом, замкнутые методы требуют обратного канала, обеспечивающего отклик на передачу, возможности распознавания, на что был этот отклик, и возможности соответствующей модификации характеристик передатчика, основываясь при этом на полученном отклике. Из этих требований вытекает необходимость довольно значительной обработки в реальном времени, производимой на оконечном устройстве, и двустороннего канала связи каждого оконечного устройства с центральным узлом. Итак, недостатком замкнутых методов является требование значительной обработки в реальном времени, производимой на оконечном устройстве, двусторонний канал связи каждого оконечного устройства с центральным узлом и то, что получение синхронизации требует относительно длительного промежутка времени. Преимущество состоит в том, что для работы системы не требуется внешнего источника знаний, а отклик по обратному каналу связи позволяет системе быстро и легко приспосабливаться к изменению геометрии системы и условий связи.

10.3.1. Открытая синхронизация передатчиков

Открытые системы можно разделить на те, которые используют информацию, полученную по каналу обратной связи, и те, которые не используют подобной информации. Последние являются наиболее простыми из возможных (с точки зрения требований к обработке в реальном времени), но качество связи в этом случае весьма сильно зависит от устойчивости характеристик канала.

Во всех схемах синхронизации передатчиков предварительно пытаются скорректировать отсчет времени и частоту передачи сигнала так, чтобы сигнал прибывал на приемник с ожидаемой частотой и в ожидаемый момент времени. Итак, для предварительного согласования времени передатчик делит расстояние до приемника на скорость света (что дает время передачи), после чего прибавляет к полученной величине время действительного начала передачи. При своевременной передаче сигнал поступит на приемник в соответствующее время. Время поступления сигнала определяется следующим выражением.

$$T_A = T_t + \frac{d}{c} \tag{10.84}$$

В данном случае T_t — действительное время начала передачи, d — расстояние передачи, c — скорость света. Подобным образом для предварительного согласования частоты передачи передатчик должен вычислить доплеровское смещение, происходящее вследствие относительного движения передатчика и приемника. Угловая частота передачи должна определяться следующим образом.

$$\omega = \left(1 - \frac{V}{c}\right)\omega_0 \tag{10.85}$$

Здесь c — скорость света, V — относительная скорость (больше нуля при уменьшении расстояния между приемником и передатчиком), а ω_0 — номинальная угловая частота передачи.

К сожалению, на практике ни предварительное согласование времени, ни предварительное согласование частоты точно выполнить невозможно. Даже спутники на геостационарных орбитах несколько изменяют свое положение относительно точки на земной поверхности, а поведение временных и частотных эталонов на оконечном устройстве и центральном узле невозможно предсказать идеально точно. Следовательно, всегда будет существовать некоторая ошибка предварительного согласования частоты и времени. Временные сбои можно записать следующим образом.

$$T_e = \frac{r_e}{c} + \Delta t \tag{10.86}$$

В данной ситуации r_e — ошибка в определении расстояния, а Δt — разность между эталонным временем терминала и эталонным временем приемника. Ошибку по частоте можно выразить следующим образом.

$$\omega_e = \frac{V_e \omega_0}{c} + \Delta\omega \tag{10.87}$$

Здесь V_e — ошибка в измеренной или предсказанной относительной скорости передатчика и приемника (доплеровская ошибка), а $\Delta\omega$ — разность между эталонными частотами приемника и передатчика. Помимо указанных, существует множество других источников временных и частотных ошибок, но, как правило, они менее важны. В работе [20] приводится полный список источников временных и частотных ошибок для спутниковых систем.

Члены Δt и $\Delta\omega$ обычно возникают вследствие случайных флуктуаций эталонных частот. Эталонное время для передатчика или приемника обычно получается посред-

ством подсчета периодов частотного эталона, так что ошибки точности измерения времени и частоты взаимосвязаны. Флуктуации эталонной частоты очень сложно описать статистически, хотя спектральная плотность мощности флуктуаций аппроксимируется последовательностью степенных сегментов [15]. Частотные эталоны часто характеризуются максимальным относительным изменением частоты за день.

$$\delta = \frac{\Delta\omega}{\omega_0} \text{ Герц/(Герц за день)} \tag{10.88}$$

Типичные значения δ находятся в диапазоне от 10^{-5} до 10^{-6} для недорогих кварцевых генераторов, от 10^{-9} до 10^{-11} — для высококачественных кварцевых генераторов; до 10^{-12} — для рубидиевых и 10^{-13} — для цезиевых. Следствием задания системного эталона частоты через максимальную относительную частоту является то, что при отсутствии внешнего воздействия номинальная частота ω_0 может линейно расти со временем.

$$\Delta\omega(T) = \omega_0 \int_0^T \delta\, dt + \Delta t(0) = \omega_0 \delta T + \Delta\omega(0) \text{ Герц} \tag{10.89}$$

Для эталонного времени, определяемого подсчетом периодов, суммарный сдвиг времени связан с суммарной фазовой ошибкой эталонной частоты.

$$\begin{aligned} \Delta t(T) &= \int_0^T \frac{\Delta\omega(t)}{\omega_0} dt + \Delta t(0) = \\ &= \int_0^T \delta t\, dt + \int_0^T \frac{\Delta\omega(0)}{\omega_0} dt + \Delta t(0) = \\ &= \tfrac{1}{2}\delta T^2 + \frac{\Delta\omega(0)T}{\omega_0} + \Delta t(0) \end{aligned} \tag{10.90}$$

Следовательно, при отсутствии внешнего воздействия ошибка эталонного времени может квадратично расти со временем. Для систем открытой синхронизации передатчиков данный квадратичный рост временной ошибки часто определяет, насколько часто должна поставляться информация извне для обновления знаний оконечного устройства о ходе времени в приемнике или для сброса эталонного таймера приемника и передатчика до номинальных значений. Рост квадратичной ошибки часто означает, что ошибка синхронизации — это большая проблема, чем частотные ошибки, хотя, вообще-то, это зависит еще и от структуры системы.

Если передатчик не обладает информацией об измерениях, поступающей по каналу обратной связи, сдвиги частоты и времени, моделируемые согласно уравнениям (10.86)–(10.90), позволят разработчику системы определить максимальную длительность времени между сеансами передачи информации извне. Повторная калибровка временного и частотного эталонов часто представляет собой обременительную процедуру; она должна выполняться как можно реже.

Если оконечное устройство имеет доступ к обратному каналу от центрального узла и возможность проводить сравнительные измерения локального эталона и параметров поступающего сигнала, промежуток времени между повторными калибровками можно сделать больше. Большие станции управления спутниками могут измерять и моделировать параметры орбит геостационарных спутников с

точностью до нескольких сантиметров в пространстве и до нескольких метров в секунду по скорости относительно наземного терминала. Таким образом, для важного частного случая синхронных спутников первым членом правой части уравнений (10.86) и (10.87) обычно можно пренебречь. Если это справедливо, разность между параметрами поступающего сигнала и сигнала, генерируемого с использованием эталонных частоты и времени терминала, будет приблизительно равна Δt и $\Delta\omega$. Данные векторы ошибок, измеряемые в обратном канале, могут применяться для вычисления соответствующей коррекции передачи в прямом канале. С другой стороны, если известно, что частотный и временной эталоны точны, но под вопросом находится геометрия канала — возможно, потому что оконечное устройство мобильно или спутник находится не на геостационарной орбите — некоторые измерения в обратном канале могут использоваться для определения неопределенности по скорости или координате. Данные измерения расстояния или относительной скорости могут затем применяться для предварительной коррекции частоты и отсчета времени в канале "оконечная станция-центральный узел".

Если оконечное устройство может использовать измерения, произведенные над сигналом из обратного канала, это иногда называется квазизамкнутой синхронизацией приемника. Квазизамкнутые системы, очевидно, обладают большей способностью приспосабливаться к неопределенностям в системе связи, чем открытые. Для корректной работы чистые открытые системы требуют полного *априорного* знания всех важных параметров канала связи. Непредвиденных изменений в канале допускать нельзя. Квазизамкнутые системы, с другой стороны, требуют *априорного* знания всех (кроме одного) важных параметров как для синхронизации времени, так и для синхронизации частоты, а оставшийся параметр можно определить из наблюдения обратного канала. Это как усложняет оконечное устройство, так и позволяет адаптироваться к некоторым типам незапланированных изменений в канале, что может значительно снизить частоту требуемых калибровок системы.

10.3.2. Закрытая синхронизация передатчиков

Закрытая синхронизация передатчиков включает передачу специальных синхронизирующих сигналов, которые используются для определения временной или частотной ошибки сигнала относительно желаемой частоты или отсчета времени поступления сигнала на приемник. Затем полученные результаты по обратной связи подаются на передатчик. Определение ошибок синхронизации может быть явным или неявным. Если центральный узел имеет достаточные возможности для обработки, он может выполнять действительное измерение ошибки. Результатом подобного измерения может быть указание величины и направления сдвига или, возможно, только направления. Данная информация будет отформатирована и возвращена на передатчик по обратному каналу. Если центральный узел имеет недостаточные возможности для обработки, особый синхронизирующий сигнал может просто возвращаться на передатчик по обратному каналу. В этом случае интерпретацией сигнала занимается передатчик. Отметим, что создание специального синхронизирующего сигнала, который легко и однозначно интерпретировать, может оказаться довольно сложной задачей.

Относительные преимущества и недостатки закрытых систем обоих типов связаны с расположением средств обработки сигнала и эффективностью использования канала. Основным преимуществом обработки на центральном узле является то, что ре-

зультатом измерений ошибки, произведенных на узле, может быть короткая цифровая последовательность. Подобное эффективное использование обратного канала может быть важным, если обратный канал является единственным на большое количество терминалов, использующих уплотнение с временным разделением. Еще одно потенциальное преимущество состоит в том, что средство измерения ошибки на центральном узле может совместно использоваться всеми терминалами, которые связываются через этот узел. Это, в свою очередь, может значительно снизить потребление ресурсов системы. Принципиальным потенциальным преимуществом обработки на терминале является то, что связь с центральным узлом не всегда является легкой задачей, а из соображений надежности, возможно, центральный узел должен быть максимально простым. Описанная ситуация — это, например, использование в роли центрального узла космического спутника. Еще одним потенциальным преимуществом обработки на терминале является то, что результат может быть получен быстрее, поскольку при использовании центрального узла всегда имеется некоторая задержка. Это может быть важно, если параметры канала меняются очень быстро. Основные недостатки заключаются в неэффективном использовании обратного канала и в том, что обратные сигналы может оказаться сложно интерпретировать. Сложность возникает, когда центральный узел является не просто ретранслятором, а выполняет функцию принятия решения относительно значений символов и передает эти решения по обратному каналу. Возможность принятия решения относительно значений символов может значительно снизить вероятность появления ошибки при передаче между терминалами; кроме того, это усложняет процедуру синхронизации. Это объясняется тем, что сдвиги частоты и отсчета времени неявно присутствуют в обратном сигнале, т.е. постольку, поскольку они влияют на процесс принятия решения относительно значения символов. Рассмотрим в качестве примера передачу сигналов в модуляции BFSK на центральный узел, принимающий некогерентные двоичные решения. Решения будут зависеть от энергии обнаруженного сигнала в детекторах метки и паузы. (Напомним, что "метка" (mark) — это название двоичной единицы, а "пауза" (space) — двоичного нуля.) Если переданный сигнал — это последовательность чередующихся меток и пауз, сигнал на центральном узле можно смоделировать следующим образом.

$$r(t) = \begin{cases} \sin[(\omega_0 + \omega_s + \Delta\omega)t + \theta] & 0 \le t \le \Delta t \\ \sin[(\omega_0 + \Delta\omega)t + \theta] & \Delta t < t \le T \end{cases} \tag{10.91}$$

Здесь T — интервал передачи символов, ω_0 — частота одного символа, $(\omega_0 + \omega_s)$ — частота другого символа, $\Delta\omega$ — ошибка по частоте на центральном узле, Δt — ошибка времени поступления сигнала на центральный узел, а θ — произвольная фаза. Теперь, если

$$x = \frac{1}{T}\int_0^T r(t)\cos\omega_0 t\, dt \tag{10.92}$$

и

$$y = \frac{1}{T}\int_0^T r(t)\sin\omega_0 t\, dt \tag{10.93}$$

представляют квадратурные компоненты детектора, то энергию обнаруженного сигнала можно записать следующим образом.

$$z^2 = x^2 + y^2$$
$$= \left(\frac{\sin[(\omega_s + \Delta\omega)\Delta t/2]}{(\omega_s + \Delta\omega)T}\right)^2 + \left(\frac{\sin[\Delta\omega(T-\Delta t)/2]}{\Delta\omega T}\right)^2 + \tag{10.94}$$
$$+ \frac{\cos(\Delta\omega\Delta t) + \cos[\Delta\omega T - (\omega_s + \Delta\omega)\Delta t] - \cos(\Delta\omega T) - \cos(\omega_s \Delta t)}{2\Delta\omega(\omega_s + \Delta\omega)T^2}$$

В частном случае нулевой ошибки времени Δt уравнение (10.94) упрощается до следующего вида.

$$z^2 = \left[\frac{\sin(\Delta\omega T/2)}{\Delta\omega T}\right]^2 \tag{10.95}$$

При нулевой ошибке по частоте, получаем следующее.

$$z^2 = \left(\frac{T-\Delta t}{2T}\right)^2 + \left[\frac{\sin(\omega_s \Delta t/2)}{\omega_s T}\right]^2 \tag{10.96}$$

Относительно выражений (10.94)–(10.96) следует сделать одно важное замечание: любая ошибка времени, частотный сдвиг или их комбинация снизит энергию принятого сигнала в детекторе, согласованном с истинным сигналом, и увеличит энергию в другом детекторе. Это приведет к уменьшению эффективного расстояния между сигналами в сигнальном пространстве и повышению вероятности ошибки. В то же время измерения вероятности ошибки (единственное, что доступно по обратному каналу) не позволяют определить, вызвана ли ошибка в результате сбоя времени или частоты (или их комбинации). Следовательно, передача обычных сигналов не дает отклика, который можно было бы использовать для синхронизации.

Полезным методом точной предварительной коррекции частоты для нашего примера передачи сигналов с модуляцией BFSK является передача постоянного тона, частота которого равна среднему от двух символьных частот. Подобный тон должен создавать случайную двоичную последовательность в обратном канале с равным числом меток и пауз. Смещение частоты со среднего значения приведет к доминированию пауз или меток. Нахождение центральной частоты описанным методом позволяет провести точную предварительную коррекцию частоты сигналов. После нахождения точной частоты передатчик может передавать последовательность чередующихся пауз и меток с целью определения точного отсчета времени. Изменяя отсчет времени при передаче (в пределах половины интервала передачи символа), передатчик может искать отсчет времени, дающий *максимальное* число ошибок. Если передача поступает на центральный узел со смещением относительно истинного отсчета времени на половину интервала передачи символа, оба детектора получают равную энергию и последовательность в обратном канале будет случайной. Определив время, когда переданные и полученные сигналы декоррелируют, передатчик вычисляет точное время передачи. Отметим, что данная процедура дает лучшие результаты, чем попытка найти точку с *минимальным* числом ошибок. Любая качественно разработанная система будет обладать достаточной энергией передачи, допускающей незначительные погрешности синхронизации времени; так что безошибочный обратный сигнал может быть получен и при неидеальной синхронизации. Фактически, чем больше отношение сигнал/шум, тем хуже работает процедура нахождения оптимума. В то же время процедура нахождения наихудшего варианта будет хорошо работать в любой качественной системе, а

ее потенциальная точность повышается с увеличением отношения сигнал/шум. Это можно понять интуитивно, поскольку увеличение отношения сигнал/шум позволяет системе справляться с большими погрешностями синхронизации; так что уменьшение вероятности ошибки при уменьшении погрешности отсчета времени от половины времени передачи символа будет более быстрым при большом отношении сигнал/шум. Таким образом, это позволит точнее определить смещение отсчета времени на половину интервала передачи символа.

10.4. Резюме

В данной главе рассмотрены фундаментальные проблемы и вопросы, связанные с синхронизацией в цифровой связи. Компромиссы обычно заключаются между стоимостью и сложностью, с одной стороны, и вероятностью ошибки, с другой. В главе обсуждались синхронизация приемника и контуры фазовой автоподстройки частоты (phase-lock loop — PLL, ФАПЧ), в частности. Обычно более активную роль в обеспечении синхронизации канала связи играет именно приемник. Даже в тех случаях, когда предполагается, что более активную роль играет передатчик, как в некоторых спутниковых каналах связи, процесс часто облегчается за счет обратного канала, по которому терминал получает информацию с приемника. Таким образом, более важное значение имеет синхронизация приемника. Контуры ФАПЧ и их разновидности — это основные схемы управления, используемые для сопровождения (отслеживания) изменений фазы поступающего сигнала. Математическое описание реакции контура ФАПЧ на данный входной сигнал включает решение нелинейного дифференциального уравнения. Было показано, впрочем, что при стационарных условиях линеаризованная модель дает достаточно хорошее приближенное описание системы. Для случая, когда линеаризованная модель неприменима, были представлены результаты Витерби (Viterbi) [8], полученные для контуров первого порядка. Строго, данные результаты справедливы только для контуров первого порядка, но было показано [5], что и для контуров более высоких порядков они являются полезным приближением.

В этой главе был рассмотрен крайне важный частный случай схем подавления несущей. Данные схемы необходимы для сопровождения фазы входного сигнала, не имеющего средней энергии на несущей частоте. Распространенный пример подобного сигнала — модулированный с использованием обычной антиподной схемы BPSK. В данной ситуации гармоника подавления несущей создается посредством применения нелинейности и далее отслеживается.

Следующий уровень синхронизации — символьная. Здесь были рассмотрены основные классы символьной синхронизации. Открытые синхронизаторы работают непосредственно с модулированным сигналом, отмечая символьные переходы. Замкнутые синхронизаторы используют управляющий контур обратной связи для нахождения и сопровождения символьных переходов.

Наивысший из рассмотренных уровней синхронизации — кадровая. Для получения данных в удобной форме приемник должен определить, какие символы и к каким кадрам принадлежат. Данное знание эквивалентно наличию кадровой синхронизации, что обычно выполняется путем включения в поток информации о некоторой характерной последовательности битов, известной приемнику. Приемник исследует входные данные, пока не обнаружит данную последовательность. Проверка синхронизации — это, например, проверка периодичности появления данной последовательности.

В данной главе были обозначены основные важные проблемы, вопросы и результаты, связанные с синхронизацией систем цифровой связи. Читатель, интересующийся данным вопросом, может обратиться к представленной ниже литературе, где обстоятельно описываются все важные моменты.

Литература

1. Peterson W. W. and Weldon E. J. *Error-Correcting Codes.* The MIT Press, Cambridge, Mass., 1972.
2. Lee E. A. and Messerschmitt D. G. *Digital Communications.* Kluwer Academic Publications, Boston, 1988.
3. Mengali U. and D'Andrea A. N. *Synchronization Technique for Digital Receivers.* Plenum Press, New York, 1997.
4. Meyr H., Moeneclaey M. and Fechtel S. A. *Digital Communication Receivers.* John Wiley & Sons, Inc., New York, 1998.
5. Gardner F. M. *Phaselock Techniques.* 2nd ed., John Wiley & Sons, Inc., New York, 1979.
6. Davenport W. B. and Root W. L. *Random Signals and Noise.* McGraw-Hill Book Company, New York, 1958.
7. Papoulis A. *Probabillity, Random Variables, and Stochastic Processes.* McGraw-Hill Book Company, New York, 1965.
8. Viterbi A. J. *Principles of Coherent Communications.* McGraw-Hill Book Company, New York, 1966.
9. Lindsey W. C. *Synchronization Systems in Communication and Control.* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1972.
10. Lindsey W. C. and Simon M. K. *Detection of Digital FSK and PSK Using a First-Order Phase-Locked Loop.* IEEE Trans. Commun., vol. COM25, n. 2, February, 1977, pp.200–214.
11. Develet J. A., Jr. *The Influence of Time Delay on Second-Order Phase Lock Loop Acquisition Range.* Int. Telem. Conf., London, 1963.
12. Johnson W. A. *A General Analysis of the False-Lock Problem Associated with the Phase-Lock Loop.* The Aerospace Corp., Rep. TOR-269(4250-45)-1, NASA Accession N64-13776, 1963.
13. Tausworthe R. C. *Acquisition and False-Lock Behavior of Phase-Locked Loops with Noise Inputs.* Jet Propulsion Laboratory, JPL SPS 37–46, vol. 4, 1967.
14. Franks L. E. *Synchronization Subsystems: Analysis and Design*; in K. Feher, Digital Communications, Satellite/Earth Station Engineering, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1981, Chap. 7.
15. Simon M. K. and Yuen J. H. *Receiver Design and Performance Characteristics*; in J. H. Yuen, ed., Deep Space Telecommunications Systems Engineering, Plenum Press, New York, 1983.
16. Gardner F. M. *Hang-up in Phase-Lock Loops.* IEEE Trans. Commun., COM25, October 1977.
17. Blanchard A. *Phase-Locked Loops.* John Wiley & Sons, Inc., New York, 1976.
18. Holmes J. K. *Coherent Spread Spectrum Systems.* John Wiley & Sons, Inc., New York, 1976.
19. Lindsey W. C. and Simon M. K., eds. *Phase Locked Loops and Their Applications.* IEEE Press, New York, 1977.
20. Spilker J. J., Jr. *Digital Communications by Satellite.* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1977.
21. Wintz P. A and Luecke E. J. *Performance of Optimum and Suboptimum Synchronizer.* IEEE Trans. Commun. Technol., June, 1969, pp.380–389.
22. Anderson J. B., Aulin T. and Sundberg C. E. *Digital Phase Modulation.* Plenum Press, New York, 1986.
23. Laurent P. A. *Exact and Approximate Construction of Digital Phase Modulations by Superposition of Amplitude Modulated Pulses.* IEEE Trans. Commun., COM–34, n. 2, pp. 150–160, February, 1986.
24. Lui G. L. *Threshold Detection Performance of GMSK Signal with BT = 0.5.* IEEE MILCOM 98 Proceedings, vol. 2, Boston, October, 18–21, 1998, pp. 515–519.
25. Kaleh G. *Differentially Coherent Detection of Binary Partial Response Continuous Phase Modulation with Index 0.5.* IEEE Trans. Commun., COM–39, pp. 1335–40, September, 1991.
26. Barkey R. H. *Group Synchronization of Binary Digital Systems*; in W. Jackson, ed., Communication Theory, Academic Press, Inc., New York, 1953.
27. Willard M. W. *Optimum Code Patterns for PCM Synchronization.* Proc. Natl. Telem. Conf., 1962, paper 5–5.

28. Newman F. and Hofman L. *New Pulse Sequences with Desirable Correlation Properties.* Proc. Natl. Telem. Conf., 1971, pp. 272–282.
29. Wu W. W. *Elements of Digital Satellite Communications.* Vol. 1, Computer Science Press, Inc., Rockville, Md., 1984.

Задачи

10.1. Передатчик (маяк) посылает немодулированный тон постоянной энергии к удаленному приемнику. Приемник и передатчик движутся друг относительно друга так, что $d(t) = D[1 - \sin(mt)] + D_0$, где $d(t)$ — расстояние между передатчиком и приемником (данное выражение может, например, описывать самолет, выписывающий "восьмерки" над наземной станцией), а D, m и D_0 — некоторые константы. Данное относительное движение приведет к доплеровскому смещению принятой частоты передатчика

$$\Delta\omega_D(t) = \frac{\omega_0 V(t)}{c},$$

где $\Delta\omega_D$ — доплеровское смещение, ω_0 — номинальная несущая частота, $V(t) = d(t)$ — относительная скорость приемника относительно передатчика, а c — скорость света. Пусть используется линеаризованное уравнение контура, а контур ФАПЧ приемника синхронизирован (нулевое рассогласование по фазе) в момент времени $t = 0$. Покажите, что контур первого порядка подходящей структуры может поддерживать синхронизацию по частоте.

10.2. Рассмотрим передатчик и приемник, движущиеся один относительно другого, как описано в задаче 10.1. Снова предположим, что используется линеаризованное уравнение контура. Определите (при таком предположении) рассогласование по фазе контура ФАПЧ как функцию времени для широкополосного фильтра и фильтра нижних частот (см. формулы (10.13) и (10.14)). Покажите, что правомочность использования уравнений линеаризованного контура зависит от значения коэффициента K_0.

10.3. Высокоэффективный летательный аппарат передает немодулированный несущий сигнал на наземный терминал. Изначально терминал синхронизирован с сигналом. Аппарат выполняет маневр, динамика которого описывается значением ускорения $a(t) = At^2$, где A — константа. Предполагая использование линеаризованного уравнения контура, определите минимальный порядок контура ФАПЧ, необходимого для сопровождения сигнала от данного аппарата.

10.4. Покажите, что ширина полосы контура ФАПЧ первого порядка записывается в виде $B_L = K_0/4$, где K_0 — коэффициент усиления контура.

10.5. Контур ФАПЧ второго порядка содержит следующий фильтр нижних частот.

$$F(\omega) = \frac{\omega_1}{i\omega + \omega_1}$$

Коэффициент усиления контура равен K_0. Предполагая, что $K_0 \geq \omega_1/4$, покажите, что ширина полосы контура ФАПЧ определяется выражением $B_L = K_0/8$. (*Подсказка*[3]:

$$\int_{-\infty}^{\infty} \frac{dx}{R} = \frac{\pi \cos(h/2)}{2cq^3 \sin h} \quad \text{для } 4ac > b,$$

где $R = a + bx^2 + cx^4$, $q = \sqrt[4]{a/c}$ и $\cos h = -b/2\sqrt{ac}$).

10.6. Контур ФАПЧ первого порядка с усилением K_0 возмущается аддитивным белым гауссовым шумом с нормированной (на энергию единичного сигнала) двусторонней спектральной плотностью мощности $N_0/2$ Вт/Гц. Определите требуемое соотношение между спектральной плотностью мощности шума и коэффициентом усиления контура, если проскальзывание цикла происходит не чаще одного раза в сутки.

---

[3] Gradshteyn I. S. and Ryzhik I. M. *Table of Integrals, Series and Products.* New York: Academic Press, 1965, 2.161.1.

10.7. Витерби [8] показал, что функция плотности вероятности выходной фазы контура ФАПЧ первого порядка, возмущенная белым гауссовым шумом, описывается следующим выражением.

$$p(\phi) = \frac{\exp(\rho \cos \phi)}{2\pi I_0(\rho)}, \quad |\phi| \leq \pi, \ \rho \geq 0$$

Покажите, что приведенное выше $p(\phi)$ действительно является функцией плотности вероятности, и вычислите среднее и дисперсию ϕ.

10.8. Компьютерное моделирование и лабораторные измерения показали, что времена между проскальзываниями цикла распределены экспоненциально, т.е. функция распределения времени между проскальзываниями цикла T выглядит следующим образом.

$$p(T) = 1 - \exp\left(-\frac{T}{T_m}\right)$$

Используя данную функцию распределения, найдите среднее время между проскальзываниями цикла и дисперсию как функцию от T_m. Если среднее между проскальзываниями цикла равно 1 день, чему равна вероятность проскальзывания цикла менее чем через час после предыдущего? Более чем через 3 дня?

10.9. Рассмотрим контур ФАПЧ второго порядка с фильтром нижних частот.

$$F(\omega) = \frac{\omega_1}{i\omega + \omega_1}$$

В процессе принудительной синхронизации желательно, чтобы контур сканировался по всей области неопределенности (1000 радиан) за 1 с. Соотношение между усилением контура и константой фильтра постоянно, $K_0 = 2\omega_1$. Определите требуемое соотношение между усилением контура и односторонней спектральной плотностью мощности аддитивного белого гауссова шума, N_0. Найдите максимальное приемлемое значение N_0.

10.10. Рассмотрим работу открытого символьного синхронизатора; ширина полосы полосового фильтра этого синхронизатора равна $0{,}1/T$ Герц, где T — период передачи символа. Если отношение энергии бита к спектральной плотности мощности шума (E_b/N_0) равно 10 дБ, чему приблизительно будут равны среднее и дисперсия относительной ошибки сопровождения? Вычислите верхнюю границу вероятности того, что ошибка сопровождения превышает утроенное приближенное относительное среднее. (*Подсказка*: рассмотрите неравенство Чебышева [7].)

10.11. Система связи используется для передачи команд со скоростью 100 бит/с. Каждая команда предваряется N-битовым заголовком, идентифицирующим ее в потоке данных. Предполагая, что (возможно, за исключением заголовка) биты появляются случайным образом [P(1) = P(0) = 1/2], определите минимальную длину заголовка, при которой ожидаемая частота ложных тревог — одна за год. Предполагая, что вероятность ошибки в канальном бите равна 10^{-5}, определите вероятность пропуска заголовка. Чему равна вероятность пропуска, если вероятность ошибки в канальном бите равна 2×10^{-2}? Если система изменяется так, что разрешает использование заголовка с двумя ошибками, чему равна минимальная требуемая длина заголовка, дающего ожидаемую частоту ложных тревог — одну за год? Чему равна вероятность пропуска заголовка в этой новой системе при вероятности ошибки в канальном бите 2×10^{-2}?

10.12. Зонд для исследования дальнего космоса удаляется от земли со скоростью 15 000 м/с, с неточностью определения скорости ±3 м/с. Эталонная частота зонда откалибрована так, чтобы ее скорость ухода не превышала 10^{-9} Герц/(Герц в день). Номинальная частота передачи зонда равна 8 ГГц. После месяца (30 дней) молчания зонд начинает запланированные передачи на наземную станцию, которая использует цезиевые часы. Какую частоту центрирования и ширину полосы поиска частоты следует использовать наем-

ной станции? Пусть расстояние до зонда точно известно на начало месяца, а неопределенность в определении времени и частоты зонда равна нулю [$\Delta t(0) = 0$, $\Delta\omega(0) = 0$]. Определите неопределенность во времени поступления сигнала от зонда.

10.13. Канал связи раз в сутки в течение небольшого периода времени работает на частоте 10 ГГц. Приемник использует контур ФАПЧ второго порядка с погрешностью частоты получения синхронизации ±1 КГц. Пусть контур самосинхронизируется, и приемник и передатчик используют однотипные эталоны частоты. Определите тип данного эталона частоты.

10.14. В некоторый момент времени ($t = 0$) выходной сигнал генератора тактовых импульсов имеет ошибку -4×10^{-3} относительно эталонного генератора. В этот момент времени генератор дает сигнал на точной частоте f_r, но далее он начинает спешить со скоростью 2 на 10^{10} в день.

а) Через сколько дней выходной сигнал генератора тактовых импульсов будет иметь нулевую ошибку?

б) Если генератору позволить работать 30 дней после получения нулевой ошибки, какой станет ошибка за это время?

10.15. При обычных предположениях (шум AWGN с нулевым средним, сигналы равных энергий) подтвердите справедливость утверждения, что правая часть уравнения (10.67) имеет вид функции правдоподобия для оценки фазы несущей и синхронизации символов.

10.16. Рассмотрим передачу сигналов с модуляцией MSK с полным откликом, где синхронизационная настроечная последовательность — это последовательность чередующихся единиц и нулей (т.е. $\alpha_k = 1$ — для четных k и -1 — для нечетных).

а) Покажите, что в данном случае существует всего два различных фазовых состояния $\{\Phi_k\}$.

б) Выведите для данного случая импульсную характеристику фильтра $h^{(l)}(t)$(определенную в (10.64)).

в) Используя результаты п. б, получите уравнения (10.68) и (10.69).

10.17. Дайте разумное объяснение причин успеха (или неуспеха) итеративной процедуры, предложенной для решения уравнений (10.70) и (10.71).

Вопросы для самопроверки

10.1. Каково определение *синхронизации* в контексте систем цифровой связи и почему она важна (см. раздел 10.1.1)?

10.2. Почему системы синхронизации, хорошо работающей в домашнем радиоприемнике, может быть *недостаточно* на высокоэффективном самолете? Какой модификации обычно требует подобная система (см. раздел 10.1.2)?

10.3. *Линеаризованное уравнение контура* зависит от приближения. Какое это приближение, почему оно подходит для синхронизированных или почти синхронизированных контуров и почему его нельзя использовать для анализа получения синхронизации (см. раздел 10.2.1)?

10.4. Контуры фазовой автоподстройки частоты второго порядка имеют определенные преимущества с точки зрения производительности и являются основой анализа сопровождения фазы. Назовите два таких преимущества (см. раздел 10.2.1.1).

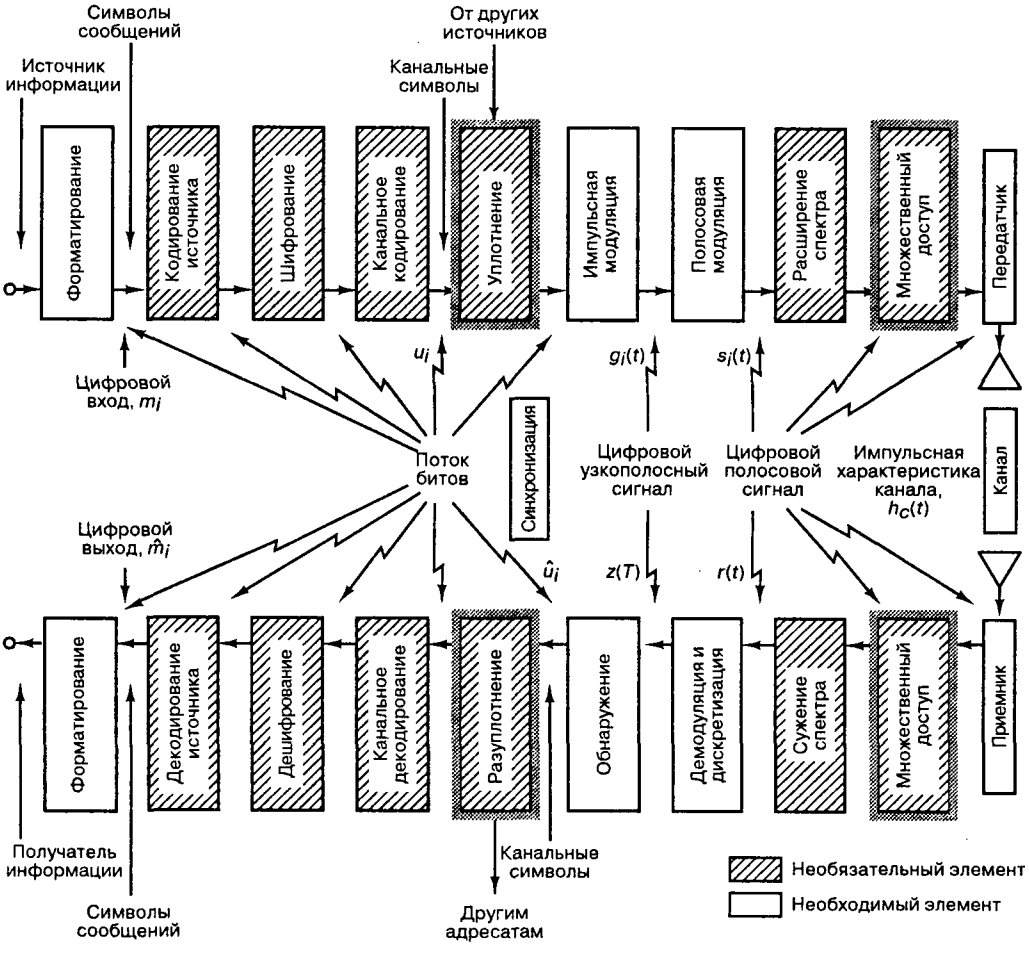
10.5. Почему схемы с *модуляцией без разрыва фазы* приобретают повышенное значение в современных системах связи и какие проблемы синхронизации возникают при их использовании (см. раздел 10.2.3.1)?

10.6. Назовите преимущества и недостатки *синхронизации с использованием данных* и *без использования данных* (см. раздел 10.2.3.2).

10.7. Опишите ситуацию, когда передатчик стоит синхронизировать для удовлетворения требований приемника (см. раздел 10.3).

ГЛАВА 11

Уплотнение и множественный доступ



Ресурс связи (communications resource — CR) представляет время и ширину полосы, доступные для передачи сигнала в определенной системе. Графически ресурс связи можно изобразить на двухмерном графике, где ось абсцисс представляет время, а ось ординат — частоту. Для создания эффективной системы связи необходимо спланировать распределение ресурса между пользователями системы, чтобы время/частота использовались максимально эффективно. Результатом такого планирования должен быть равноправный доступ пользователей к ресурсу.

С проблемой совместного использования ресурса связи связаны термины "уплотнение" и "множественный доступ". Разница между этими понятиями минимальна. При использовании термина *уплотнение* требования пользователя к совместному использованию ресурса связи постоянны либо (в большинстве случаев) изменяются незначительно. Распределение ресурса выполняется априорно, а совместное использование ресурса обычно привязывается к локальному устройству (к примеру, монтажной плате). Применение *множественного доступа*, как правило, требует *удаленного совместного использования* ресурса, как, например, в случае спутниковой связи. При динамической схеме множественного доступа контроллер системы должен учитывать потребности каждого пользователя ресурса связи. Время, необходимое для передачи соответствующей управляющей информации, устанавливает верхний предел эффективного использования ресурса связи.

11.1. Распределение ресурса связи

Существует три основных способа увеличения пропускной способности (общей скорости передачи данных) ресурса связи. Первый состоит в увеличении эффективной изотропно-излучаемой мощности (effective isotropic radiated power — EIRP) передатчика или в снижении потерь системы, что в любом случае приведет к увеличению отношения E_b/N_0. Второй способ — это увеличение ширины полосы канала. Третий способ заключается в повышении эффективности распределения ресурса связи. Одна из возможных реализаций этого способа — множественный доступ. Пример: спутниковый транспондер, который должен эффективно распределить ограниченный ресурс связи между большим количеством пользователей, обменивающихся цифровой информацией. При этом пользователи могут требовать различных скоростей передачи данных и иметь разные рабочие циклы. Основные способы распределения ресурса связи приводятся ниже (рис. 11.1, под заголовком уплотнение/множественный доступ).

1. *Частотное разделение* (frequency division — FD). Распределяются определенные поддиапазоны используемой полосы частоты.
2. *Временное разделение* (time division — TD). Пользователям выделяются периодические временные интервалы. В некоторых системах пользователям предоставляется ограниченное время для связи. В других случаях время доступа пользователей к ресурсу определяется динамически.
3. *Кодовое разделение* (code division — CD). Выделяются определенные элементы набора ортогонально (либо почти ортогонально) распределенных спектральных кодов, каждый из которых использует весь диапазон частот.
4. *Пространственное разделение* (space division — SD), или *многолучевое многократное использование частоты*. С помощью точечных лучевых антенн радиосигналы разделяются и направляются в разные стороны. Данный метод допускает многократное использование одного частотного диапазона.
5. *Поляризационное разделение* (polarization division — PD), или *двойное поляризационное многократное использование частоты*. Для разделения сигналов применяется ортогональная поляризация, что позволяет использовать один частотный диапазон.

Рис. 11.1. Основные преобразования цифровой связи

Ключевым моментом во *всех* схемах уплотнения и множественного доступа является то, что при использовании ресурса различными сигналами интерференция не дает неуправляемых взаимных помех, которые делают невозможным процесс обнаружения. Интерференция допустима до тех пор, пока сигналы одного канала незначительно увеличивают вероятность появления ошибок в другом канале. Избежать взаимных помех между разными пользователями позволяет использование в разных каналах ортогональных сигналов. Сигналы $x_i(t)$, где $i = 1, 2, ...$, являются ортогональными, если во временной области выполняется условие:

$$\int_{-\infty}^{\infty} x_i(t)x_j(t)dt = \begin{cases} K & \text{при } i = j \\ 0 & \text{при } i \neq j \end{cases}, \tag{11.1}$$

где K — ненулевая константа. Подобным образом сигналы ортогональны, если в частотной области выполняется условие:

$$\int_{-\infty}^{\infty} X_i(f)X_j(f)df = \begin{cases} K & \text{при } i = j \\ 0 & \text{при } i \neq j \end{cases}, \tag{11.2}$$

где функции $X_i(f)$ являются Фурье-образами сигналов $x_i(t)$. Распределение по каналам, характеризующееся ортогональными волнами, для которых выполняется условие (11.1), называют *уплотнением с временным разделением* (time-division multiplexing — TDM) или *множественным доступом с временным разделением* (time-division multiple access — TDMA). Распределение по каналам, характеризующееся ортогональными волнами, для которых выполняется условие (11.2), называют *уплотнением с частотным разделением* (frequency-division multiplexing — FDM) или *множественным доступом с частотным разделением* (frequency-division multiple access — FDMA).

11.1.1. Уплотнение/множественный доступ с частотным разделением

11.1.1.1. Использование уплотнения с частотным разделением в телефонной связи

На заре создания телефонной связи для каждой магистральной телефонной линии, соединяющей междугородные телефонные центры, было необходимо устанавливать два провода. Как видно из рис. 11.2, небо над крупными городами становилось все темнее по мере развития телефонной связи. Важное открытие в области телефонной связи в начале XX века — уплотнение с частотным разделением (frequency-division multiplexing — FDM) — позволило передавать несколько телефонных сигналов по одному проводу, а следовательно, изменить методы телефонной передачи.

Ресурс связи представлен на рис. 11.3 в виде частотно-временной зависимости. Спектральное распределение по каналам является примером технологии FDM или FDMA. Здесь распределение сигналов или пользователей по диапазону частот является *долгосрочным* или *постоянным*. Ресурс связи может одновременно содержать несколько сигналов, разнесенных в спектре. Первый частотный диапазон содержит сигналы, которые используют промежуток частот между f_0 и f_1, второй — между f_2 и f_3 и т.д. Области спектра, находящиеся между используемыми диапазонами, называют *защитными полосами частот*. Защитные полосы выполняют роль буфера, что позволяет снизить интерференцию между соседними (по частоте) каналами.

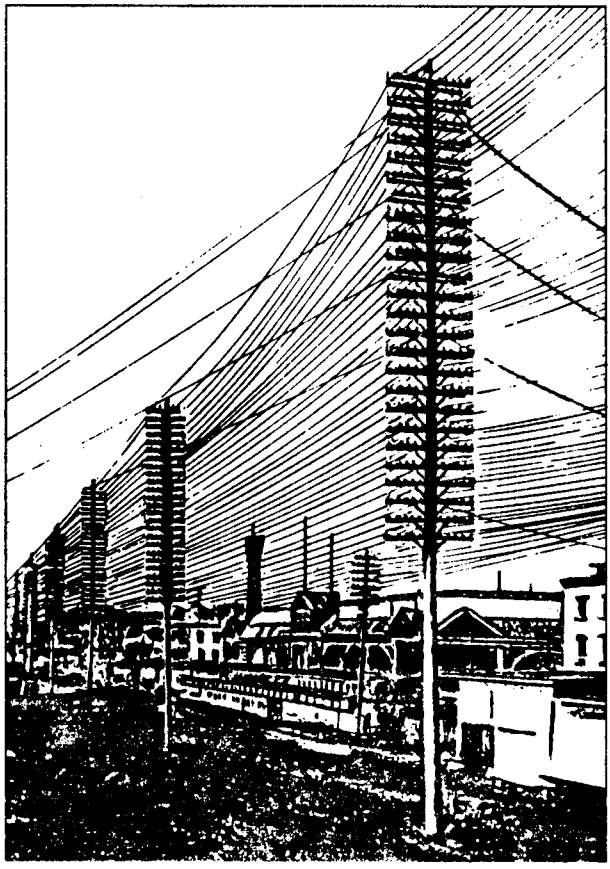


Рис. 11.2. На заре создания телефонной связи для каждой магистральной телефонной линии было необходимо устанавливать два провода

Может возникнуть вопрос: как преобразовать немодулированный сигнал так, чтобы он использовал более высокий диапазон частот? Ответ: при помощи *наложения* или *смешивания* (*модуляции*) информационного сигнала и синусоидального сигнала фиксированной частоты.

Если два модулируемых входящих сигнала описываются синусоидами с частотами f_A и f_B, их смешение или перемножение дает частоты f_{A+B} и f_{A-B}. Процесс модуляции описывается следующим тригонометрическим равенством.

$$\cos A \cos B = \frac{1}{2}\left[\cos(A+B) + \cos(A-B)\right] \tag{11.3}$$

На рис. 11.4, *а* показано модулирование типичного голосового телефонного сигнала $x(t)$ (частоты немодулированного сигнала принадлежат диапазону 300–3400 Гц) синусоидальным сигналом с частотой 20 кГц. Двусторонний спектр немодулированного сигнала, $|X(f)|$, показан на рис. 11.4, *а*. Может ли смеситель сигналов быть линейным устройством? Нет. Выходной сигнал линейного устройства будет иметь *те же* состав-

ляющие частоты, что и входной сигнал. Различие может быть лишь в амплитуде и/или фазе.

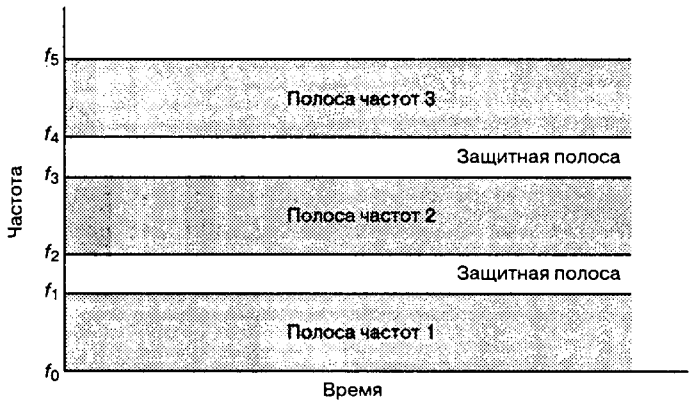


Рис. 11.3. Уплотнение с частотным разделением

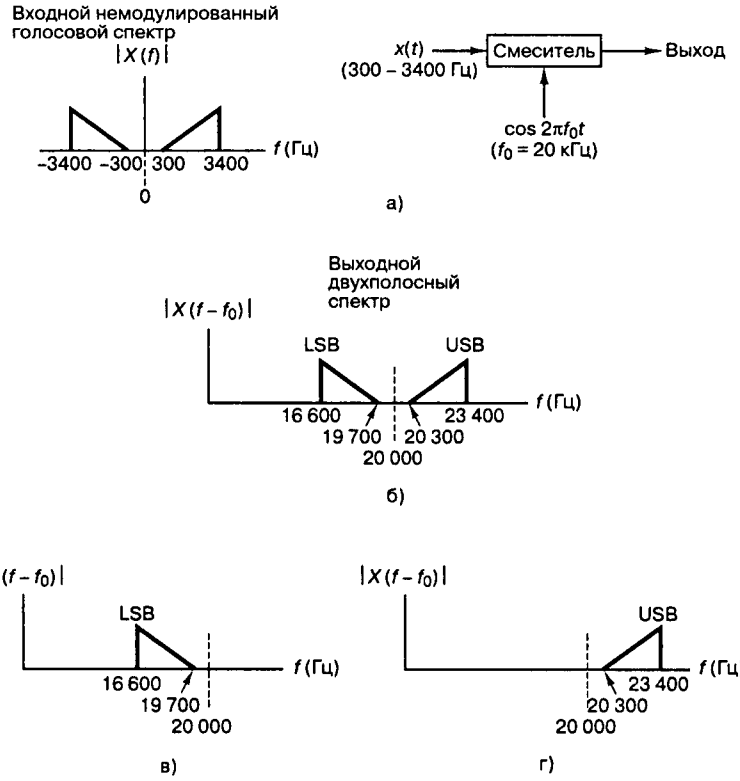


Рис. 11.4. Смешивание сигналов: а) процесс смешивания; б) выходной спектр смесителя; в) нижняя боковая полоса; г) верхняя боковая полоса

На рис. 11.4, *б* представлен односторонний спектр $|X(f-f_0)|$ на выходе смесителя. В результате смешивания, описанного в уравнении (11.3), спектр смещается в сторону

более высоких частот, по сравнению с немодулированным спектром, и центрирован теперь на частоте 20 кГц. Данный спектр называют двухполосным (double-sideband — DSB), поскольку информация находится в двух различных диапазонах частот. На рис. 11.4, *в* показана нижняя боковая полоса (lower sideband — LSB), которой принадлежат частоты 16 600–19 700 Гц. Иногда нижнюю боковую полосу называют *инвертированной боковой полосой*, поскольку частотные составляющие этой полосы расположены в обратном порядке, по сравнению с немодулированным сигналом. Подобным образом фильтрование может использоваться для выделения верхней боковой полосы (upper sideband — USB), которой, как показано на рис. 11.4, *г*, принадлежат частоты 20 300–23 400 Гц. Данную боковую полосу иногда называют *прямой*, поскольку частотные составляющие этой полосы расположены в том же порядке, что и в немодулированном сигнале. Обе боковые полосы спектра DSB содержат одну и ту же информацию. Таким образом, для восстановления исходных данных немодулированного сигнала необходима лишь одна боковая полоса — верхняя или нижняя.

На рис. 11.5 приведен простейший пример технологии FDM. В данном случае реализована схема с тремя каналами передачи речи. В канале 1 голосовой сигнал из диапазона 300–3 400 Гц модулируется сигналом с частотой 20 кГц. В каналах 2 и 3 аналогичный голосовой сигнал модулируется сигналами с частотами 16 и 12 кГц. В приведенном примере сохраняются лишь нижние боковые полосы. Результатом смешивания и фильтрации (для удаления верхних боковых полос) являются сдвинутые по частоте сигналы, показанные на рис. 11.5. Суммарный выходной сигнал есть суммой трех сигналов и принадлежит диапазону 8,6–19,7 кГц.

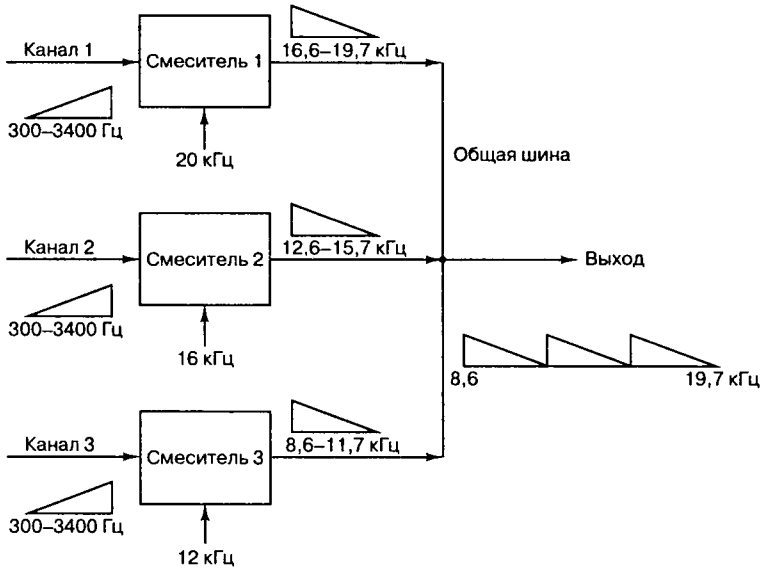


Рис. 11.5. Простейший пример FDM. Три сдвинутых по частоте канала передачи речи

На рис. 11.6 представлены два наиболее низких уровня иерархии уплотнения телефонных каналов с использованием FDM. Первый уровень состоит из *группы* 12 каналов, модулируемых поднесущими с частотами из диапазона 60–108 кГц. Второй уровень, состоящий из пяти групп (60 каналов), называют *супергруппой*. Супергруппа мо-

дулируется поднесущими с частотами из диапазона 312–552 кГц. Уплотненные каналы теперь рассматриваются как составной сигнал, который может передаваться по кабелю или модулироваться несущей с целью последующей радиопередачи.

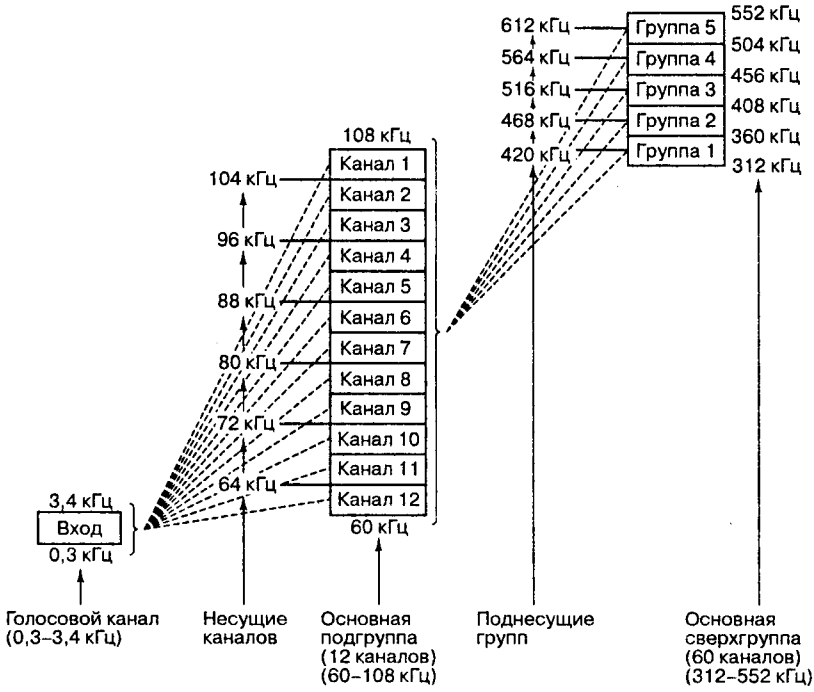


Рис. 11.6. Схема модулирования типичной системы уплотнения с частотным разделением

11.1.1.2. Множественный доступ с частотным разделением в спутниковых системах

Большинство спутников связи расположено на *геостационарной* или *геосинхронной* орбите. Это означает, что спутник находится на круговой орбите, лежащей в плоскости земного экватора. При этом спутник находится на такой высоте над уровнем моря (приблизительно 35 830 км), на которой период его обращения вокруг Земли равен периоду вращения самой Земли. Поскольку при наблюдении с Земли такие объекты кажутся неподвижными, три спутника, расположенных через 120° друг от друга, позволяют охватить территорию всего земного шара (за исключением, полярных областей). Большинство спутниковых систем связи используют нерегенеративные ретрансляторы или транспондеры. *Нерегенеративный* означает, что сигналы "земля-спутник" усиливаются, сдвигаются по частоте и ретранслируются на Землю без обработки сигнала, демодуляции или повторной модуляции. Наиболее широко используемым диапазоном в коммерческих системах спутниковой связи является так называемая *полоса С* (C-band). В данном диапазоне для передачи сигнала "земля-спутник" применяется несущая частота 6 ГГц и частота 4 ГГц передачи сигнала "спутник-земля". Согласно международным соглашениям, для систем передачи в полосе С разрешено использовать любой спутник, работающий в спектральном диапазоне шириной в 500 МГц. В большинстве случаев такой спутник имеет 12 транспондеров с шириной полосы 36 МГц каждый. Наиболее

распространенные транспондеры работают в режиме FDM/FM/FDMA (уплотнение с частотным разделением, частотная модуляция, множественный доступ с частотным разделением). Рассмотрим составляющие указанного режима.

1. *FDM.* Сигналы, подобные телефонным, имеющие одиночную боковую полосу шириной 4 кГц, обрабатываются с использованием FDM, в результате чего формируется составной многоканальный сигнал.
2. *FM.* Составной сигнал модулируется несущей и передается на спутник.
3. *FDMA.* Поддиапазоны полосы транспондера (36 МГц) могут распределяться между различными пользователями. Каждому пользователю выделяется определенная полоса, на которой он получает доступ к транспондеру.

Таким образом, составные каналы FDM модулируются (FM), после чего информация передается на спутник, будучи распределенной по различным полосам в соответствии с системой FDMA. Основным преимуществом технологии FDMA, в сравнении с TDMA, является простота. Каналы FDMA не требуют синхронизации или централизованного распределения времени. Каждый из каналов независим от остальных. Позднее будут рассмотрены преимущества TDMA в сравнении с FDMA.

11.1.2. Уплотнение/множественный доступ с временным разделением

На рис. 11.3 показано совместное использование ресурса связи, выполняемое посредством распределения частотных диапазонов. На рис. 11.7 тот же ресурс связи распределен путем предоставления каждому из *M* сигналов (или пользователей) всего спектра в течение небольшого отрезка времени, называемого *временным интервалом* (time slot). Промежутки времени, разделяющие используемые интервалы, называются *защитными интервалами* (guard time). Защитный интервал создает некоторую временную неопределенность между соседними сигналами и выступает в роли буфера, снижая тем самым интерференцию. На рис. 11.8 приведен пример использования технологии TDMA в спутниковой связи. Время разбито на интервалы, называемые *кадрами* (frame). Каждый *кадр* делится на *временные интервалы*, которые могут быть распределены между пользователями. Общая структура кадров периодически повторяется, так что передача данных по схеме TDMA — это один или более временных интервалов, которые периодически повторяются на протяжении каждого кадра. Каждая наземная передающая станция транслирует информацию в виде пакетов таким образом, чтобы они поступали на спутник в соответствии с установленным расписанием. После принятия транспондером такие пакеты ретранслируются на Землю вместе с информацией от других передающих станций. Принимающая станция обнаруживает и разуплотняет уплотненные данные соответствующего пакета, после чего информация поступает к соответствующим пользователям.

11.1.2.1. TDM/TDMA с фиксированным распределением временных интервалов

Простейшая схема TDM/TDMA именуется *TDM/TDMA с фиксированным распределением.* При использовании такой схемы *M* временных интервалов, составляющих кадр, заранее распределены между источниками сигнала на достаточно длительный промежуток времени. На рис. 11.9 в виде блок-схемы показана работа такой системы. Операция уплотнения состоит в предоставлении каждому источнику возможности использовать один или более интервалов. Разуплотнение — это распознавание интервалов с последующим распределением данных между соответствующими пользователями.

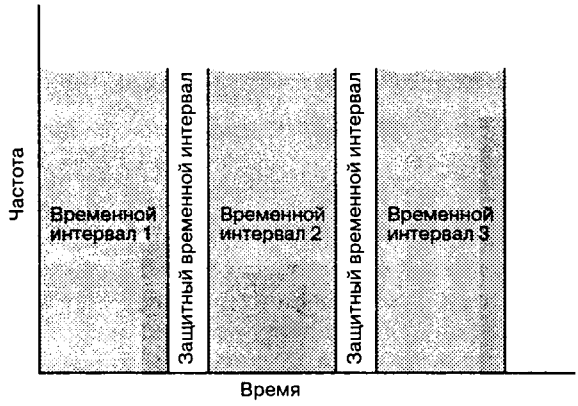


Рис. 11.7. Уплотнение с временным разделением

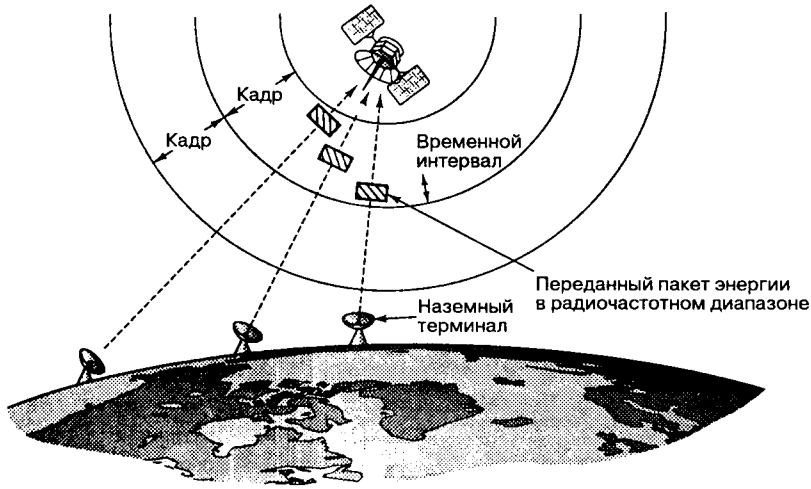


Рис. 11.8. Типичная конфигурация TDMA

Два коммутирующих ключа на рис. 11.9 должны быть синхронизированы таким образом, чтобы сообщение, соответствующее источнику 1, попадало на выход канала 1 и т.д. Само по себе сообщение в общем случае состоит из начальной комбинации битов (preamble) и собственно информационной части. Начальная комбинация обычно состоит из элементов, которые отвечают за синхронизацию, адресацию и защиту от ошибок.

Схема TDM/TDMA с фиксированным распределением является чрезвычайно эффективной, когда требования пользователя можно предвидеть, а поток данных значителен (т.е. временные интервалы практически всегда заполнены). В случае же пульсирующего или случайного потока данных указанный метод себя не оправдывает. Рассмотрим простой пример, представленный на рис. 11.10. Здесь кадр составляют четыре интервала, каждый из которых закреплен за пользователями *A*, *B*, *C* и *D*. На рис. 11.10, *а* изображены схемы активности четырех пользователей.

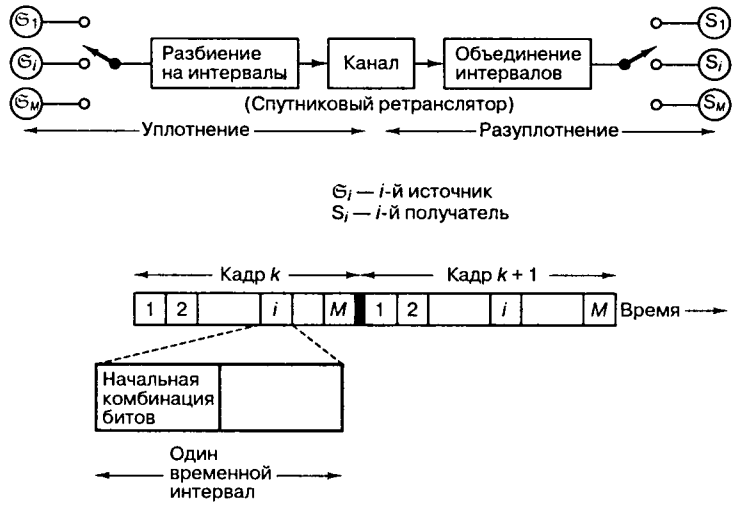


Рис. 11.9. TDM с фиксированным распределением

На протяжении первого интервала передачи кадра пользователь *C* не отправляет данных, пользователь *B* не передает данных в течение второго интервала, а *A* — в течение третьего. В случае использования TDMA с фиксированным распределением все интервалы кадра распределены заранее. Если "владелец" интервала не передает данных в течение указанного промежутка времени, данный интервал не используется. На рис. 11.10, *б* показан поток данных и неиспользованные интервалы. Если требования пользователей непредсказуемы, как в приведенном выше примере, то должны применяться более эффективные методы с использованием динамического распределения интервалов. Таких методов существует несколько — применение систем с коммутацией пакетов, статистических мультиплексоров или концентраторов. Данные системы позволяют достигнуть результата, изображенного на рис. 11.10, *в*, где пропускная способность системы остается постоянной благодаря использованию всех доступных временных интервалов.

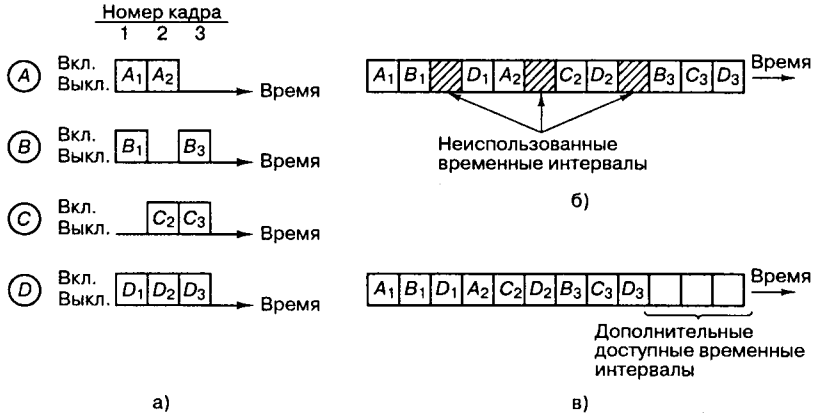


Рис. 11.10. TDM с фиксированным распределением и система с коммутацией пакетов: а) схема активности пользователей; б) TDM с фиксированным распределением; в) коммутация пакетов с временным разделением (концентрация)

11.1.3. Распределение ресурса связи по каналам

На рис. 11.3 приводилось распределение ресурса связи по спектральным диапазонам, а на рис. 11.7 был приведен пример его распределения по временным интервалам. На рис. 11.11 представлен более общий способ управления ресурсом связи, позволяющий распределять частотные диапазоны на заранее определенный период времени. Такую систему множественного доступа называют *комбинированной FDMA/TDMA*. Для изучения распределения частотных диапазонов рассмотрим случай равномерного пропорционального распределения полосы шириной W между M группами (или классами) пользователей. Подобным образом частотный диапазон будем считать разбитым на полосы шириной W/M Гц, которые будут постоянно доступны соответствующим группам. Аналогично для изучения распределения временных интервалов ось времени разбивается на интервалы продолжительностью T. В свою очередь, каждый из кадров разбивается на N интервалов продолжительностью T/N каждый. Предположим, что активность пользователей синхронизирована во времени и распределенные интервалы периодично расположены в кадрах. Каждый пользователь может передавать данные, когда начинается его интервал времени, а также на протяжении данного интервала пользователь может использовать выделенную полосу частот. Временной интервал однозначно задается как m-й интервал кадра n. Обратившись к рис. 11.11, можно описать интервал (n, m) следующим образом.

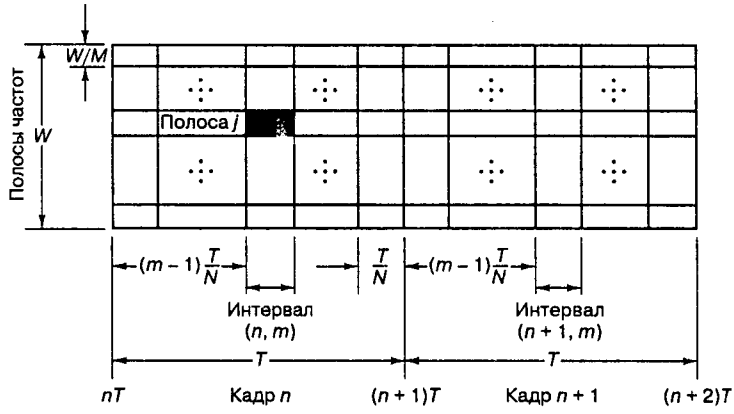


Рис. 11.11. Ресурс связи: временно-частотное распределение по каналам

$$\text{временной интервал } (n,m) = nT + \frac{(m-1)T}{N} \le t \le nT + \frac{mT}{N} \qquad (11.4)$$

$$n = 0, 1, \ldots; m = 1, 2, \ldots, N$$

Длительность кадра n, T, — это интервал $[nT, (n + 1)T]$. Как видно из рис. 11.11, область сигнала является пересечением временного интервала (n, m) и частотного диапазона (j). Предположим, что система модуляции/кодирования выбрана таким образом, что полная полоса W ресурса связи может поддерживать скорость передачи данных R бит/с. Для любого частотного диапазона, содержащего полосу W/M Гц, соответствующая скорость передачи данных будет составлять R/M бит/с. Технология FDMA позволяет использовать M диапазонов с шириной полосы $1/M$, а TDMA — полный

диапазон частот для каждого из N интервалов времени, при этом длительность каждого интервала составит $1/N$ длительности кадра.

11.1.4. Сравнение производительности FDMA и TDMA

11.1.4.1. Скорость передачи данных FDMA и TDMA

На рис. 11.12 представлены основные различия систем FDMA и TDMA для ресурса связи, поддерживающего скорость передачи данных R бит/с. На рис. 11.12, *а* полоса системы разделена на M ортогональных полос частот. Следовательно, все M источников $\mathfrak{S}_m (1 \le m \le M)$ могут одновременно производить передачу данных со скоростью R/M бит/с каждый. На рис. 11.12, *б* показан кадр, разбитый на M ортогональных временных интервалов. Таким образом, каждый из M источников передает данные со скоростью R бит/с, что в M раз больше скорости передачи от пользователя FDMA за время $(1/M)$. В обоих случаях источник $\mathfrak{S}_m$ передает информацию со средней скоростью R/M бит/с.

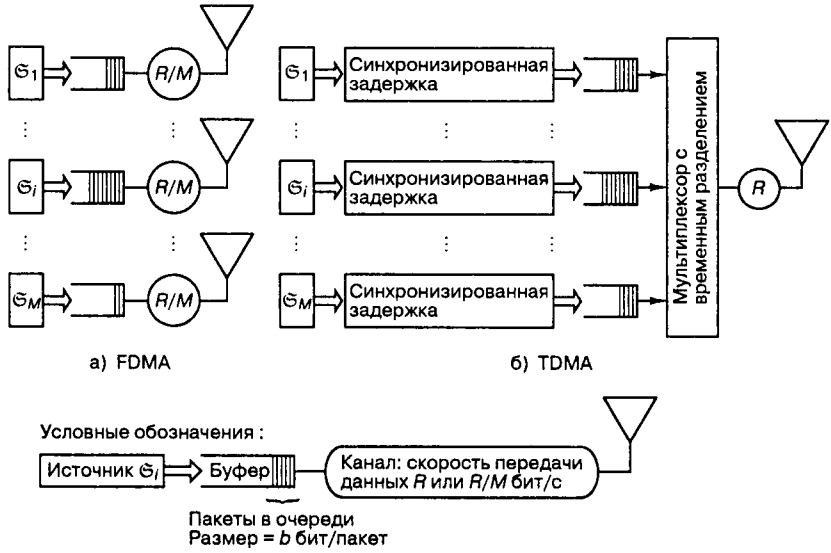


Рис. 11.12. Сравнительное представление технологий FDMA/TDMA: а) FDMA: частота делится на M ортогональных частотных диапазонов; б) TDMA: время разделено на M ортогональных временных интервалов (один пакет на интервал времени)

Пусть информация, передаваемая каждым источником на рис. 11.12, собирается в b-битовые группы или *пакеты*. В случае FDMA b-битовые пакеты передаются за T секунд по каждому из M непересекающихся каналов. Таким образом, полная скорость передачи данных может быть представлена в следующем виде.

$$R_{FD} = M\frac{b}{T} \text{ бит/с} \tag{11.5}$$

При использовании TDMA каждым источником за T/M секунд передается b бит. Следовательно, требуемая скорость передачи данных равна следующему.

$$R_{TD} = \frac{b}{T/M} \text{ бит/с} \tag{11.6}$$

Поскольку уравнения (11.5) и (11.6) идентичны, можно сделать следующий вывод.

$$R_{FD} = R_{TD} = R = \frac{Mb}{T} \text{ бит/с} \tag{11.7}$$

Следовательно, обе системы требуют одинаковой скорости передачи данных — R бит/с.

11.1.4.2. Задержка сообщений в системах FDMA и TDMA

Исходя из предыдущих разделов, можно сделать вывод, что, несмотря на некоторые различия, FDMA и TDMA не отличаются по производительности. Однако различие становится очевидным, если в качестве единицы измерения производительности используется средняя *задержка* пакета. Показано [1, 2], что TDMA значительно превосходит FDMA по данному параметру, т.е. среднее время задержки пакета при использовании первой схемы меньше, чем при использовании последней.

Как и ранее, предположим, что при FDMA диапазон частот системы разбит на M ортогональных полос; при использовании TDMA кадр разделен на M ортогональных временных интервалов. Для анализа времени задержки сообщения рассмотрим простейший случай детерминистических источников данных. Предположим, что ресурс связи используется на 100%. Тогда все частотные диапазоны при FDMA и все временные интервалы при TDMA будут заполнены пакетами данных. Для простоты будем считать, что *отсутствуют* дополнительные издержки, связанные с защитными полосами или интервалами. В таком случае время задержки сообщения можно выразить следующим образом.

$$D = w + \tau \tag{11.8}$$

Здесь w — среднее время ожидания пакета (до передачи), τ — время передачи пакета. При FDMA каждый пакет пересылается в течение T секунд; передача пакета для технологии FDMA будет следующей.

$$\tau_{FD} = T \tag{11.9}$$

При использовании TDMA каждый пакет пересылается в течение временного интервала T/M секунд. С помощью уравнения (11.7) время передачи пакета можно выразить следующим образом.

$$\tau_{TD} = \frac{T}{M} = \frac{b}{R} \tag{11.10}$$

Поскольку каналы FDMA доступны постоянно, а пакеты пересылаются непосредственно после создания, время ожидания w_{FD} составляет следующее.

$$w_{FD} = 0 \tag{11.11}$$

На рис. 11.13 сравниваются потоки данных для схем FDMA и TDMA. Как показано на рис. 11.13, *а*, при использовании TDMA временные интервалы пользователей начинаются в разных точках кадра протяженностью T секунд. Пакет S_{mk} начинает отправляться по

прошествии $(m-1)T/M$ секунд $(1 \le m \le M)$ после создания пакета. Таким образом, для TDMA среднее время ожидания пакета перед отправкой составит следующее.

$$w_{TD} = \frac{1}{M}\sum_{m=1}^{M}(m-1)\frac{T}{M} = \frac{T}{M^2}\sum_{n=0}^{M-1} n = \frac{T}{M^2}\frac{(M-1)M}{2} = \frac{T}{2}\left(1-\frac{1}{M}\right) \qquad (11.12)$$

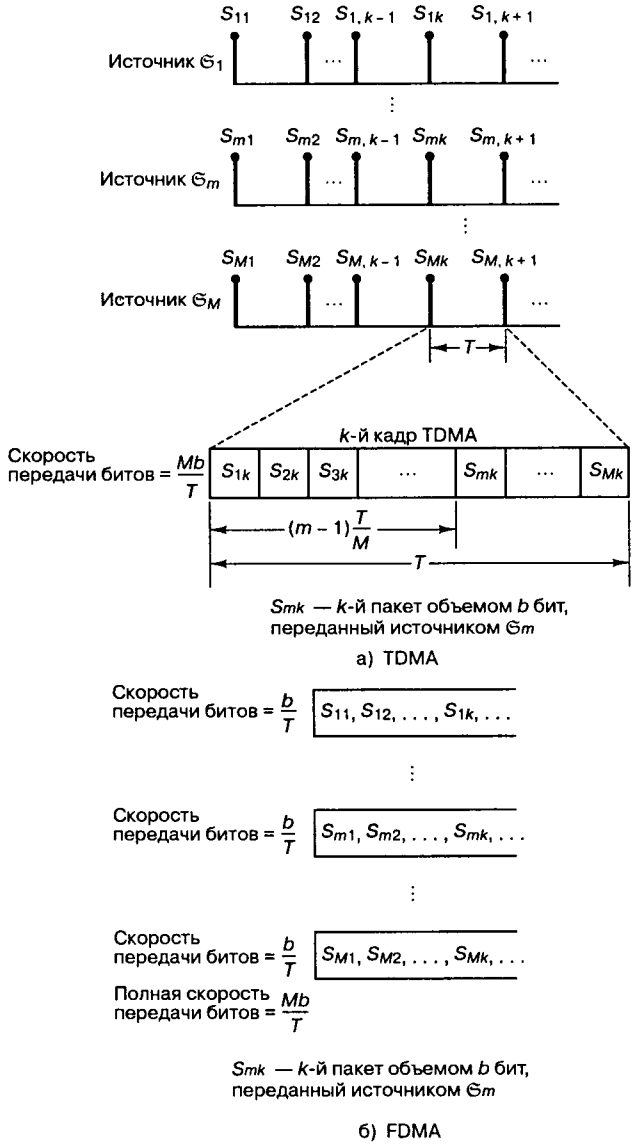


Рис. 11.13. Распределение по каналам: а) TDMA; б) FDMA

Максимальное время ожидания пакета перед отправкой составляет $(M-1)T/M$ секунд. В соответствии с уравнением (11.12), среднее время задержки пакета равно $1/2(M-1)(T/M) = (T/2)(1-1/M)$.

Для сравнения среднего времени задержки D_{FD} и D_{TD} при использовании FDMA и TDMA, соответственно, подставим уравнения (11.9) и (11.11) в (11.8) и уравнения (11.10) и (11.12) в (11.8). В результате получим следующее.

$$D_{FD} = T \tag{11.13}$$

$$D_{TD} = \frac{T}{2}\left(1-\frac{1}{M}\right)+\frac{T}{M} = D_{FD} - \frac{T}{2}\left(1-\frac{1}{M}\right) \tag{11.14}$$

С помощью уравнения (11.7) формулу (11.14) можно записать в следующем виде.

$$D_{TD} = D_{FD} - \frac{b}{2R}(M-1) \tag{11.15}$$

Результат свидетельствует о том, что FDMA значительно уступает TDMA по времени задержки сообщения. Несмотря на то что уравнение (11.5) строго справедливо для детерминистического источника данных, малые задержки передачи сообщений для TDMA сохраняются для любого независимого процесса получения данных [1, 2].

11.1.5. Множественный доступ с кодовым разделением

В случае FDMA (рис. 11.3) плоскость ресурса связи была разделена на горизонтальные отрезки, соответствующие частотным диапазонам. Та же плоскость на рис. 11.7 была разбита по вертикали на временные интервалы TDMA. Эти два подхода являются наиболее распространенными в приложениях множественного доступа. На рис. 11.14 приводится иллюстрация метода множественного доступа, являющегося результатом совмещения FDMA и TDMA. Этот метод называется *множественным доступом с кодовым разделением* (code-division multiple access — CDMA). CDMA является практическим приложением методов *расширения спектра* (spread-spectrum — SS), которые можно разделить на две основные категории: расширение спектра методом *прямой последовательности* (direct sequence — DS) и расширение спектра методом *скачкообразной перестройки частоты* (frequency hopping — FH). В данной главе будет рассмотрена схема CDMA с перестройкой частоты (FH-CDMA), описание схемы множественного доступа с кодовым разделением методом прямой последовательности приводится в главе 12.

Простейший пример CDMA с *перестройкой частоты*, кратковременное распределение частотного диапазона для различных источников сигнала, изображен на рис. 11.14. В каждом из коротких временных интервалов происходит перераспределение частотных диапазонов. Как показано на рисунке, в течение интервала 1 сигнал 1 использует диапазон 1, сигналы 2 и 3 — диапазоны 2 и 3. Во время интервала 2 сигнал 1 "перескакивает" в диапазон 3, сигнал 2 — в диапазон 1, сигнал 3 — в диапазон 2 и т.д. Таким образом, ресурс связи используется полностью, причем диапазоны пользователей перераспределяются в каждый последующий момент времени. Каждому пользователю присваивается псевдошумовой (pseudonoise — PN) код, который указывает последовательность перестройки частоты. Псевдошумовые коды ортогональны друг другу. Более подробно шумовые коды будут рассмотрены в разделе 12.2. На рис. 11.14 представлена существенно

упрощенная модель схемы CDMA с перестройкой частоты, поскольку в приведенном примере из требований симметрии вытекает, что каждый сигнал изменяет частоту синхронно со всеми остальными сигналами. Однако *в действительности этого не происходит*. Одним из преимуществ схемы CDMA в сравнении с TDMA является то, что группы пользователей не нуждаются в синхронизации (синхронизироваться должны только передатчики и приемники каждой группы).

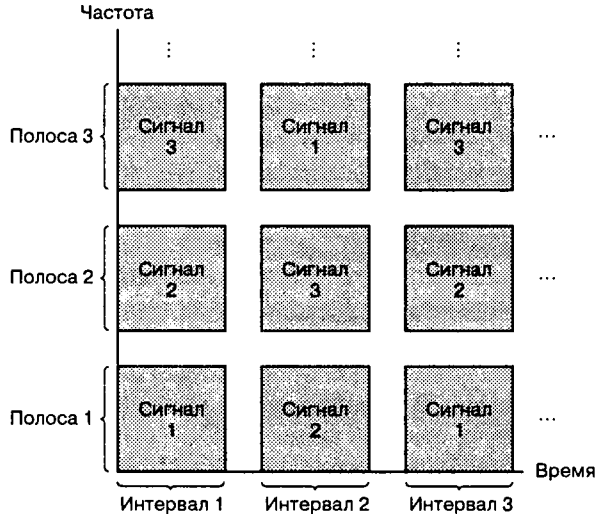


Рис. 11.14. Уплотнение с кодовым разделением

На блок-схеме, представленной на рис. 11.15, показан процесс модуляции с использованием перестройки частоты. Во время каждого изменения частоты генератор псевдошумовой последовательности направляет кодовую последовательность на *устройство скачкообразной перестройки частоты*. Данное устройство выдает одну из допустимых для скачка частот. Допустим, что используется M-арная частотная манипуляция (M-ary frequency shift keying — MFSK). При обычной системе MSFK данные модулируют несущую волну с *фиксированной* частотой. В случае MFSK с перестройкой частоты (FH-MFSK) частота несущей скачет по всему диапазону частот. FH-модуляцию на рис. 11.15 можно рассматривать как процесс, состоящий из двух этапов: модуляции данных и модуляции перестройки частоты. Указанные действия могут быть совмещены — в этом случае модулятор на основе псевдошумового кода и собственно данных генерирует тон передачи. Подробно системы с перестройкой частоты рассматриваются в разделе 12.4.

Может возникнуть вопрос: если схемы FDMA и TDMA достаточно эффективны при распределении ресурса связи, какой смысл в использовании смешанного метода? Ответом могут служить уникальные преимущества CDMA.

1. *Конфиденциальность.* Если код группы пользователей известен лишь разрешенным членам этой группы, CDMA обеспечивает конфиденциальность связи, поскольку несанкционированные лица, не имеющие кода, не могут получить доступ к передаваемой информации.
2. *Каналы с замираниями.* Если для определенной части используемого спектра характерно замирание, сигналы в данной части будут ослабленными. При исполь-

зовании схемы FDMA пользователь данной части спектра может испытывать постоянные затруднения со связью. При схеме FH-CDMA пользователь будет испытывать аналогичные проблемы только при изменении частоты в соответствующую часть спектра. Таким образом, возможные проблемы со связью равномерно распределяются между всеми пользователями.

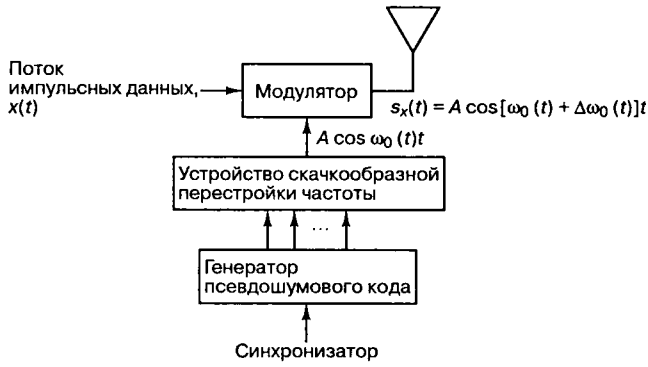


Рис. 11.15. Процесс модуляции схемы FH-CDMA

3. *Сопротивляемость подавлению.* В течение времени между изменениями частоты полоса сигнала идентична полосе обычной схемы MFSK, т.е. обычно равна минимальной ширине полосы, достаточной для передачи символа MFSK. В то же время в течение нескольких временных интервалов система совершает скачки в диапазоне частот, ширина которого намного превышает ширину полосы данных. Такое использование полосы называется расширением спектра. Расширение спектра и вытекающая из него сопротивляемость подавлению подробно описаны в главе 12.
4. *Гибкость.* Наиболее важным преимуществом CDMA, по сравнению с TDMA, является отсутствие необходимости синхронизации одновременно передающих устройств. Разные передачи не влияют на ортогональность процессов передачи с различными кодами. Данное утверждение станет понятнее при подробном описании в главе 12 автокорреляционных и взаимно корреляционных свойств кодов.

11.1.6. Множественный доступ с поляризационным и пространственным разделением

На рис. 11.16, *а* показано, как спутник INTELSAT IVA использует метод множественного доступа с пространственным разделением (space-division multiple access — SDMA), также называемый *многолучевым многократным использованием частоты.* INTELSAT IVA применяет двулучевую принимающую антенну, которая передает сигнал на два приемника. Это позволяет осуществлять одновременный доступ к спутнику из двух разных точек на Земле. Полосы частот, выделенные двум таким пользователям, одинаковы, поскольку сигналы этих пользователей разнесены в пространстве. Вообще, в таких случаях полосу называют *многократно используемой.*

На рис. 11.16, *б* показано применение спутником COMSTAR 1 множественного доступа с поляризационным разделением (polarization-division multiple access — PDMA), который также называют *двойным поляризационным многократным использованием частоты.*

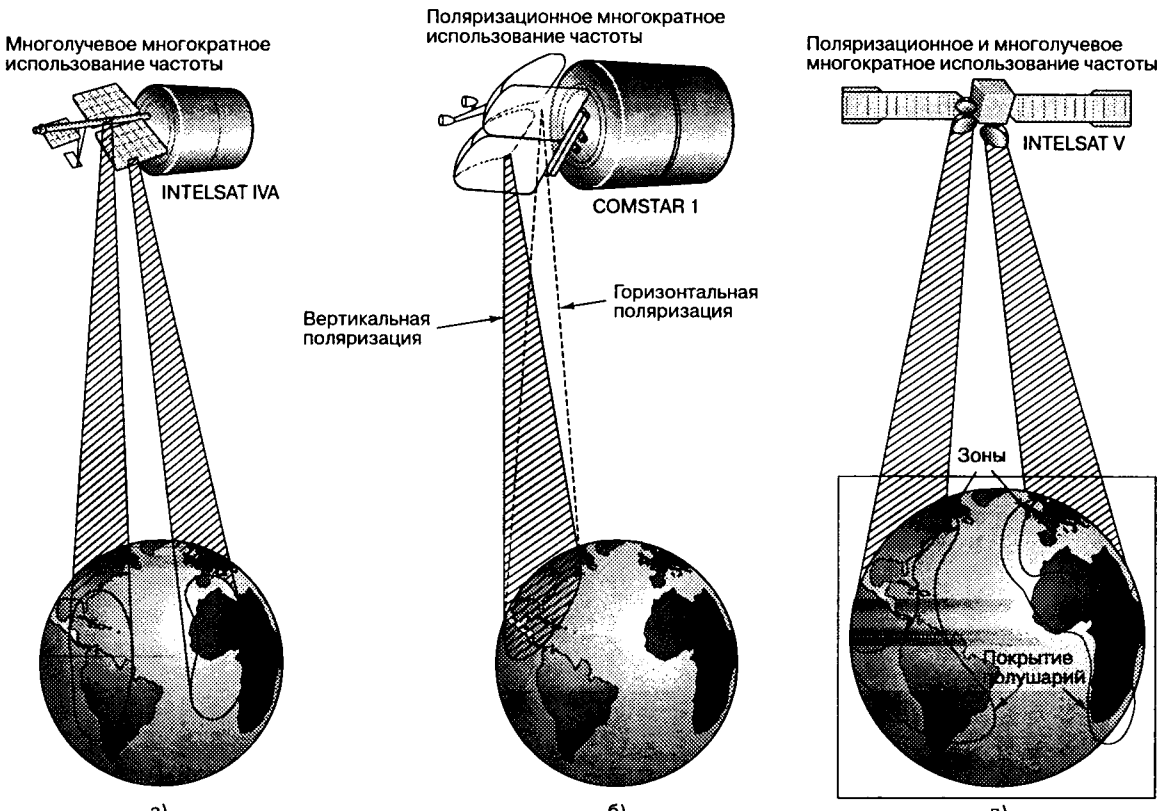


Рис. 11.16. SDMA и PDMA: а) INTELSAT IVA; б) COMSTAR 1; в) INTELSAT V (атлантическое покрытие)

В данном случае используются раздельные антенны с разными поляризациями, каждая из которых соотнесена со своим приемником. Это позволяет получить одновременный доступ к спутнику пользователям, находящимся на небольшом расстоянии друг от друга. Каждая из передающих антенн на Земле должна быть поляризована в соответствии с антенной спутника. (Обычно наземная станция оснащается антенной с двойной поляризацией). Полосы частот, используемые двумя антеннами, могут быть идентичными, поскольку они поляризованы ортогонально друг другу. Как и при SDMA, полосу частот PDMA называют многократно используемой. На рис. 11.16, *в* показано одновременное использование спутником INTELSAT V схем SDMA и PDMA. В данном случае покрытие спутника делится на две части: восточное и западное. Используется пара зональных лучей; каждый из которых частично пересекается с лучом полушария. Зональные лучи и лучи полушария взаимно ортогональны. Следовательно, в данном случае имеем четырехкратное использование спектра.

11.2. Системы связи множественного доступа и архитектура

Информация об использовании времени, частоты и кодовых функций, необходимая пользователям для сообщения между собой с помощью спутника, содержится в *протоколе* или *алгоритме множественного доступа* (multiple access algorithm — MAA). Система множественного доступа является объединением аппаратного и программного обеспечения, поддерживающим MAA. Основная задача такой системы — своевременное, упорядоченное и эффективное предоставление пользователю услуг связи.

На рис. 11.17 приводится несколько основных архитектур спутниковых систем связи множественного доступа. В условных обозначениях представлены символы, используемые для наземных станций, имеющих или не имеющих контроллер MAA. На рис. 11.17, *а* показана система, в которой одна из наземных станций определяется как основная (контроллер). На данной станции размещают компьютер, реагирующий на запросы на обслуживание, приходящие от всех остальных пользователей. Отметим, что пользовательский запрос влечет за собой передачу данных от контроллера к спутнику и обратно. Реакция контроллера приводит к другой передаче посредством спутника. Таким образом, каждая услуга требует двух сеансов передачи данных с Земли на спутник и обратно. Рис. 11.17, *б* соответствует случаю распределения управления MAA между всеми наземными станциями; выделенного контроллера не существует. Все наземные станции используют одинаковый алгоритм и располагают идентичными знаниями о запросах на доступ и распределении доступа. Следовательно, каждая услуга в этом случае требует одного цикла связи станция-спутник-станция. На рис. 11.17, *в* показан контроллер MAA, находящийся непосредственно на спутнике. Запрос пользователя поступает на спутник, который может немедленно послать ответный сигнал. Таким образом, в данной системе для предоставления услуги связи достаточно одного цикла связи.

11.2.1. Информационный поток в системах множественного доступа

На рис. 11.18 представлена блок-схема потока данных между алгоритмом множественного доступа (multiple access algorithm — MAA), или контроллером, и наземной станцией связи; нумерация пунктов в приведенном ниже списке соответствует нумерации на рисунке. Как указывалось в предыдущем разделе, за управление

может отвечать спутник или одна наземная станция; также управление может быть распределено между всеми наземными станциями. Передача данных происходит в следующем порядке.

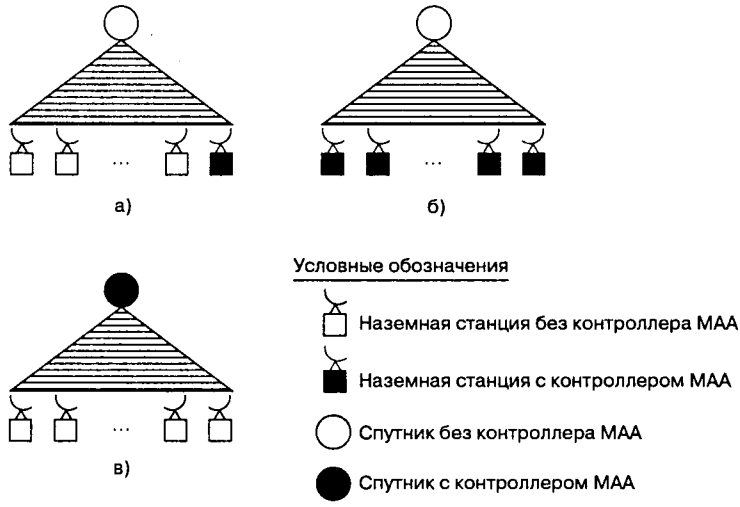


Рис. 11.17. Архитектура спутниковой системы множественного доступа: а) управление осуществляет одна наземная станция; б) управление распределено между всеми наземными станциями; в) управление осуществляет спутник

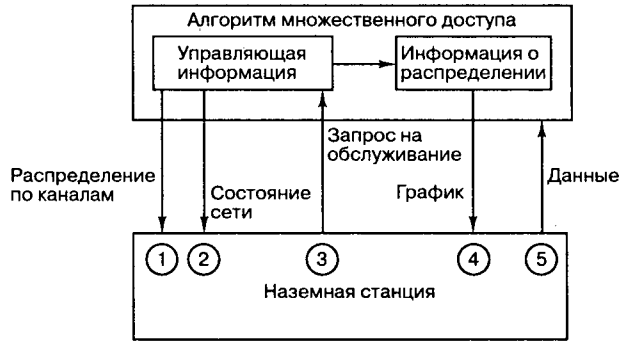


Рис. 11.18. Информационный поток в системах множественного доступа

1. *Распределение по каналам.* Данный термин относится к распределению информации (например, каналы 1–N могут быть предоставлены пользователю X, а каналы (N + 1)–M — пользователю Y). Данная информация изменяется редко и может распространяться между наземными станциями без использования системы связи, например, посредством информационного бюллетеня.
2. *Состояние сети* (network state — NS). Этот термин связан с состоянием ресурса связи. Наземная станция получает указания относительно доступности ресурса

связи, а также о том, как следует использовать время, частоту, кодовые позиции ресурса для передачи запроса на обслуживание.

3. *Запрос на обслуживание.* Станция передает запрос (запросы) на обслуживание (например, на выделение ресурса для передачи m сегментов сообщения).
4. По получении запроса (запросов) на обслуживание контроллер передает станции график, в соответствии с которым данные должны распределяться в ресурсе связи.
5. Станция передает данные в соответствии с указанным графиком.

11.2.2. Множественный доступ с предоставлением каналов по требованию

Системы множественного доступа, позволяющие передающей станции периодически получать доступ к каналу независимо от реальных потребностей, называются системами с *фиксированным распределением.* Существуют также системы с динамическим распределением, которые предоставляют доступ к каналу только при соответствующем запросе передающей станции. Их именуют системами *множественного доступа с предоставлением каналов по требованию* (demand-assignment multiple access — DAMA). Если передача данных станцией связи ведется нерегулярно или скачкообразно, схема DAMA может быть значительно эффективнее схемы фиксированного распределения. Полезность схемы DAMA объясняется тем, что фактическая потребность в ресурсах *редко* совпадает с максимальным спросом. Если пропускная способность системы равна общему максимальному спросу, а обмен данными производится нерегулярно, большую часть времени возможности системы будут использоваться не полностью. В то же время система с более низкой пропускной способностью, использующая буферизацию и схему DAMA, может успешно поддерживать скачкообразный процесс обмена данными, хотя в этом случае все же возможны некоторые задержки передачи данных. На рис. 11.19 обобщаются основные различия между системой с фиксированным распределением, пропускная способность которой равна сумме требований всех пользователей, и динамической системой, пропускная способность которой определяется средними требованиями пользователей.

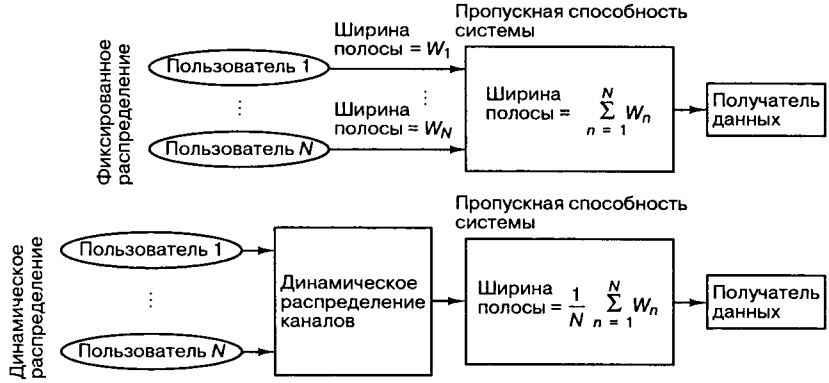


Рис. 11.19. Уменьшение ширины полосы для систем с динамическим распределением каналов

11.3. Алгоритмы доступа

11.3.1. ALOHA

В 1971 году Гавайский университет разработал и начал использовать систему ALOHA. В данном случае спутник применялся для связи нескольких университетских компьютеров посредством протокола произвольного доступа [3–7]. Принцип работы системы чрезвычайно прост и включает в себя следующие режимы.

1. *Режим передачи.* Пользователи передают данные в любой момент времени, кодируя свои сообщения с помощью кода обнаружения ошибок.
2. *Режим ожидания.* После передачи сообщения пользователь ожидает от приемника подтверждения (acknowledgment — ACK) приема данных. Иногда передачи различных пользователей перекрываются во времени, что приводит к возникновению ошибок в каждой передаче. В таком случае сообщения пользователей называют *конфликтующими.* Ошибки обнаруживаются, после чего пользователи получают отрицательное подтверждение приема (negative acknowledgment — NAK).
3. *Режим повторной передачи.* После получения сообщения NAK информация передается повторно. Естественно, если пользователи попытаются осуществить повторную передачу непосредственно после возникновения ошибки, конфликтная ситуация может повториться. Поэтому повторная передача производится после *случайной* задержки.
4. *Режим истечения времени ожидания.* Если после передачи пользователь в течение определенного времени не получил сообщения ACK или NAK, производится повторная передача.

11.3.1.1. Статистика получения сообщений

Предположим, что для работы некоторой системы необходима определенная средняя частота успешного поступления сообщений (пакетов) λ. Вследствие конфликтных ситуаций некоторые из сообщений не будут получены либо будут отклонены. Следовательно, общую частоту поступления сообщений λ_t можно определить как сумму частоты успешного поступления сообщений λ и частоты отклонения данных λ_r.

$$\lambda_t = \lambda + \lambda_r \tag{11.16}$$

Обозначим размер сообщения или пакета через b бит. Тогда средний объем успешно переданных данных, иначе говоря *пропускную способность* канала, ρ', можно представить следующим образом.

$$\rho' = b\lambda \text{ бит/с} \tag{11.17}$$

Также можно определить полный информационный обмен канала, G'.

$$G' = b\lambda_t \text{ бит/с} \tag{11.18}$$

Если считать максимальную скорость передачи битов (емкость канала) равной R бит/с, *нормированную пропускную способность* можно записать следующим образом.

$$\rho = \frac{b\lambda}{R} \tag{11.19}$$

Также можем записать *нормированный полный информационный обмен.*

$$G = \frac{b\lambda_t}{R} \tag{11.20}$$

Нормированная пропускная способность ρ выражает пропускную способность как часть ($0 \le \rho \le 1$) емкости канала. Нормированный полный информационный обмен G выражает полный информационный обмен как часть ($0 \le G \le \infty$) емкости канала. Следует отметить, что G может иметь значения, превышающие 1.

Время передачи пакета может быть выражено в следующем виде.

$$\tau = \frac{b}{R} \text{ секунд/пакет} \tag{11.21}$$

Подставляя уравнение (11.21) в (11.19) и (11.20), можем записать следующее.

$$\rho = \lambda\tau \tag{11.22}$$

и

$$G = \lambda_t\tau \tag{11.23}$$

Пользователь может успешно передавать данные, если ни один из пользователей не начал передачу в течение предыдущих τ секунд или не начнет ее в течение следующих τ секунд. В противном случае возникнет конфликт. Поэтому для успешной передачи каждого сообщения требуется 2τ секунд.

Статистика получения сообщений независимыми пользователями системы связи часто моделируется пуассоновским процессом. Вероятность поступления K новых сообщений в течение τ секунд описывается распределением Пуассона [8].

$$P(K) = \frac{(\lambda\tau)^K e^{-\lambda\tau}}{K!} \quad K \ge 0, \tag{11.24}$$

где λ — средняя частота поступления сообщений. Поскольку в системе ALOHA пользователи передают данные независимо друг от друга, приведенное выше выражение может быть использовано для вычисления вероятности события, когда в течение временного интервала 2τ будет получено точно $K = 0$ других сообщений. Таким образом, получаем P_s — вероятность успешной (бесконфликтной) передачи пользовательского сообщения. Для вычисления P_s предположим, что информационный обмен описывается распределением Пуассона, после чего подставим в уравнение (11.24) значения λ_t и 2τ.

$$P_s = P(K = 0) = \frac{(2\tau\lambda_t)^0 e^{-2\tau\lambda_t}}{0!} = e^{-2\tau\lambda_t} \tag{11.25}$$

В уравнении (11.16) общая частота поступления сообщений λ_t определялась как сумма частоты успешного поступления сообщений λ и частоты отклонения данных λ_r. Тогда, по определению, вероятность успешного получения пакета может быть выражена в следующем виде.

$$P_s = \frac{\lambda}{\lambda_t} \tag{11.26}$$

Объединяя уравнения (11.25) и (11.26), получаем следующее.

$$\lambda = \lambda_t e^{-2\tau\lambda_t} \tag{11.27}$$

Подставив в формулу (11.27) выражения (11.22) и (11.23), можно записать следующее.

$$\rho = Ge^{-2G} \tag{11.28}$$

Уравнение (11.28) связывает нормированную пропускную способность ρ и нормированный полный информационный обмен G при использовании канала системы ALOHA. График данной зависимости отмечен на рис. 11.20 как "чистый алгоритм ALOHA". По мере роста G увеличивается и ρ до тех пор, пока большое количество конфликтных ситуаций не приведет к снижению пропускной способности. Максимум ρ, равный $1/2e = 0{,}18$, достигается при $G = 0{,}5$. Таким образом, в канале с чистым алгоритмом ALOHA может быть использовано лишь 18% ресурса связи. Простота управления в данном алгоритме достигается за счет снижения емкости канала [7, 9].

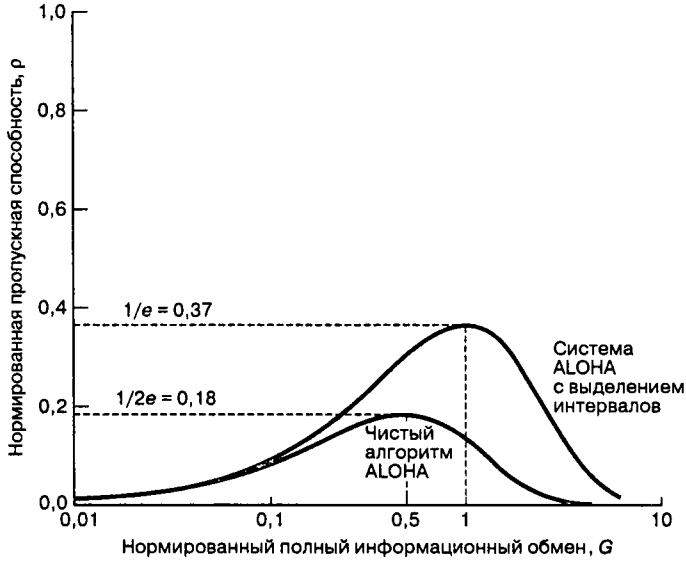


Рис. 11.20. Пропускная способность каналов ALOHA (зависимость доли успешных передач от их общего числа)

11.3.2. ALOHA с выделением временных интервалов

Чистый алгоритм ALOHA можно улучшить, если ввести небольшую координацию между станциями. Примером подобного алгоритма является система ALOHA с выделением временны*х* интервалов (slotted ALOHA — S-ALOHA). Всем станциям посредством метода широковещания передается последовательность синхронизирующих импульсов. Как и в случае чистой системы ALOHA, размер пакетов является постоянным. Сообщения могут передаваться только в течение временно*г*о интервала между синхронизирующими импульсами, а начало передачи пакета обязательно должно совпадать с *началом* интервала. Внесение таких незначительных дополнений в алгоритм ALOHA позволяет вдвое снизить число конфликтных ситуаций, поскольку теперь конфликтовать могут только сообщения,

передаваемые в течение одного временного интервала. Можно показать [9, 10], что при использовании алгоритма S-ALOHA сокращение *конфликтного промежутка* с 2τ до τ дает следующее соотношение между нормированной пропускной способностью ρ и нормированным полным информационным обменом G.

$$\rho = Ge^{-G} \tag{11.29}$$

График зависимости (11.29) приведен на рис. 11.20, где он отмечен как "система ALOHA с выделением временных интервалов". В данном случае максимальное значение ρ равно $1/e = 0{,}37$, что в два раза больше аналогичного показателя чистого алгоритма ALOHA.

Режим повторной передачи системы S-ALOHA отличается от соответствующего режима чистого алгоритма тем, что при получении пользователем отрицательного подтверждения (NAK) следующая попытка производится после *случайной* паузы, длительность которой кратна протяженности временного интервала. Работа алгоритма S-ALOHA представлена на рис. 11.21. После успешной передачи пакета данных пользователь k получает со спутника подтверждение о получении. Также показаны пользователи m и n, которые одновременно начинают передачу пакетов, что приводит к конфликту, и спутник передает сигнал NAK обоим пользователям. Для определения времени повторной передачи обе станции используют генератор случайных чисел. Далее на рисунке показано возможное продолжение: повторная передача пользователями m и n после случайно выбранной паузы. Разумеется, существует вероятность повторения конфликтной ситуации сразу же после конфликта. В этом случае после очередной случайной паузы будет предпринята еще одна попытка повторной передачи.

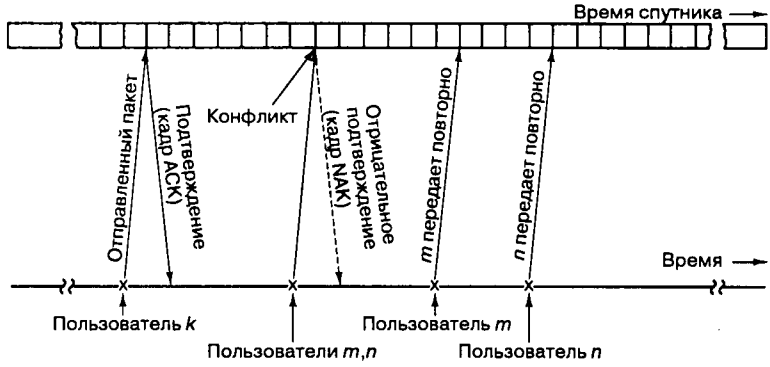


Рис. 11.21. Система произвольного доступа: работа алгоритма ALOHA с выделением временных интервалов

Пример 11.1. Процесс Пуассона

Пусть передачу и повторную передачу пакетов можно описать как пуассоновский процесс. Определите *вероятность* возникновения в процессе передачи пакета конфликта с *еще одним пользователем* (используется алгоритм S-ALOHA). Полная частота передачи пакетов равна $\lambda_t = 10$ пакетов в секунду; длительность пакета $\tau = 10$ мс.

Решение

$$P(K=1) = \left.\frac{(\tau\lambda_t)^K e^{-\tau\lambda_t}}{K!}\right|_{K=1} = (10\times 0{,}01)^1 e^{-0.1} = 0{,}1e^{-0.1} = 0{,}09$$

11.3.3. Алгоритм ALOHA с использованием резервирования

Работа систем ALOHA была значительно улучшена в результате введения резервирования (reservation-ALOHA — R-ALOHA) [11]. Системы R-ALOHA могут использоваться в двух основных режимах.

Режим без резервирования (состояние покоя)

1. Выделенный интервал времени разбивается на небольшие подынтервалы резервирования.
2. Эти подынтервалы используются для резервирования интервалов передачи сообщений.
3. После запроса резервирования пользователь ожидает подтверждения и распределения интервалов.

Режим с резервированием

1. Если не выполняется резервирование, временной интервал разбивается на $M+1$ интервалов.
2. Первые M интервалов используются для передачи сообщений.
3. Последний интервал разбивается на подынтервалы, которые используются для резервирования или передачи запросов.
4. Пользователи передают пакеты данных только в выделенных им элементах M интервалов.

Рассмотрим пример использования схемы R-ALOHA, представленный на рис. 11.22. В состоянии покоя время (с целью резервирования) разбивается на небольшие подынтервалы. После резервирования система конфигурируется так, что после $M=5$ интервалов передачи сообщений следуют $V=6$ подынтервалов резервирования; далее эта структура повторяется. На рисунке показан процесс отправления запроса и получения подтверждения. В данном примере передающей станции необходимо зарезервировать три интервала времени. В подтверждении спутника содержатся инструкции относительно размещения первого пакета данных. Управление распределено, поэтому все пользователи получают сигнал со спутника и, соответственно, информацию о резервировании и распределении времени. Поэтому в сигнале-подтверждении спутника находится *вся* необходимая информация, которая заключается в сообщении о выделении первого временного интервала. Как показано на рис. 11.22, в течение следующего интервала времени станция передает второй пакет. Далее пользователь знает, что следующий интервал состоит из шести подынтервалов, предназначенных для резервирования, поэтому передача информационных пакетов в течение этого времени *не производится*. Третий (последний) пакет отсылается в течение четвертого интервала. Если резервирование не производится, система возвращается в состояние покоя. Поскольку управление выполняется распределенно, все пользователи получают от спутника информацию об изменении состояния системы и соответствующие синхронизирующие импульсы. Другие интересные методы резервирования рассмотрены в [12, 13].

11.3.4. Сравнение производительности систем S-ALOHA и R-ALOHA

В главах 3 и 4 качество схемы цифровой модуляции определялось, в основном, зависимостью P_B от E_b/N_0. Это особенно полезно, поскольку E_b/N_0 является *нормированным отношением сигнал/шум*. Нормированные кривые позволяют сравнивать производи-

тельность различных схем модуляции. Для анализа систем множественного доступа используется подобный показатель — зависимость средней задержки от нормированной пропускной способности. На рис. 11.23 представлена *идеальная зависимость задержки от пропускной способности*. Для нормированных значений пропускной способности, $0 \le \rho < 1$, время задержки равно нулю, при $\rho = 1$ оно неограниченно возрастает. Помимо идеального случая, на рисунке изображена типичная зависимость, а также направление, соответствующее улучшению производительности.

Рис. 11.22. Пример алгоритма ALOHA с использованием резервирования. Передающая станция резервирует три интервала (M = 5 интервалов, V = 6 подынтервалов)

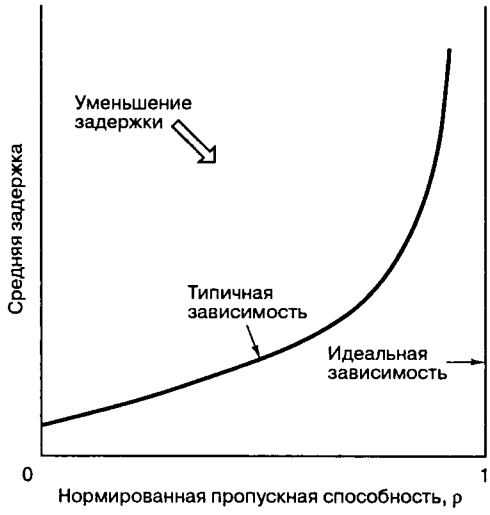


Рис. 11.23. Зависимость времени задержки от пропускной способности

На рис. 11.24 сравниваются зависимости времени задержки от пропускной способности для алгоритмов S-ALOHA и R-ALOHA (формат сообщений: два интервала передачи данных и шесть подынтервалов резервирования). Время задержки этих двух систем сравнивают с помощью *идеальной* кривой. Для пропускной способности $\rho < 0{,}2$

среднее время задержки для системы S-ALOHA меньше, чем для системы R-ALOHA. В то же время для ρ, принадлежащего диапазону 0,2–0,67, R-ALOHA превосходит S-ALOHA, поскольку у первой среднее время задержки существенно меньше. В чем причина превосходства схемы S-ALOHA при малоинтенсивном обмене данными? Данный алгоритм не требует служебных издержек для резервирования подынтервалов, как в случае R-ALOHA. Таким образом, при небольших значениях ρ производительность R-ALOHA ниже из-за более высоких расходов. При $\rho > 0{,}2$ конфликтные ситуации и повторная передача данных в системе S-ALOHA приводят к тому, что время задержки растет быстрее, чем в случае R-ALOHA (и неограниченно возрастает при $\rho = 0{,}37$). При более высоких значениях пропускной способности ($0{,}2 < \rho < 0{,}67$) служебные издержки схемы R-ALOHA полностью окупаются и обеспечивают менее резкое возрастание времени задержки при росте ρ. При использовании схемы R-ALOHA время задержки возрастает до бесконечности при $\rho = 0{,}67$.

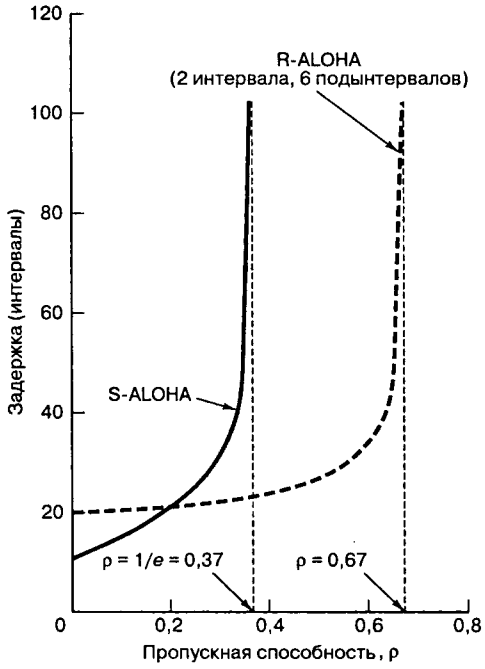


Рис. 11.24. Зависимость времени задержки от пропускной способности: спутниковый канал при использовании схем S-ALOHA и R-ALOHA

Пример 11.2. Использование канала связи

а) В качестве меры использования канала выбрана нормированная пропускная способность ρ. Ее можно найти как отношение успешно переданных данных к полному объему данных (включая отклоненные данные). Найдите нормированную пропускную способность канала связи с максимальной скоростью передачи данных $R = 50$ Кбит/с, который используется $M = 10$ станциями связи, каждая из которых передает данные со средней частотой $\lambda = 2$ пакета в секунду. Формат системы предусматривает пакеты по $b = 1350$ бит.

б) Применение какой из описанных систем ALOHA будет оптимальным в данном случае?

Решение

а) Обобщая уравнение (11.19) для информационного потока от нескольких станций, получаем следующее.

$$\rho = \frac{Mb\lambda}{R} = \frac{10(1350)(2)}{50\,000} = 0{,}54$$

б) В данной системе может использоваться только схема R-ALOHA, поскольку два других алгоритма не позволяют использовать 54% ресурса.

11.3.5. Методы опроса

Один из методов упорядочения работы системы произвольного доступа с множественными пользователями состоит во введении контроллера, выявляющего запросы на предоставление услуг путем периодического опроса всех пользователей. Если количество пользователей велико (например, тысячи терминалов), а процесс обмена данными происходит пульсирующим образом, время, выделяемое для опроса всех пользователей, может представлять существенные служебные издержки. Одним из методов быстрого опроса пользователей является *поиск по двоичному дереву* [4, 14]. На рис. 11.25 представлен пример использования данного метода для реализации "состязания" между пользователями спутниковой связи за обладание ресурсом. Пусть общее число пользователей равно восьми и каждому из них присвоен двоичный код от 000 до 111, как показано на рис. 11.25. Предположим, что терминалы 001, 100 и 110 соревнуются за один канал связи. При поиске по двоичному дереву группа пользователей периодически делится пополам, пока не останется лишь одна ветвь дерева. Терминал, соответствующий этой ветви, и получает право первым использовать канал. Затем операция повторяется, и доступ получает следующий "победитель". Алгоритм поиска состоит из следующих этапов (рис. 11.25).

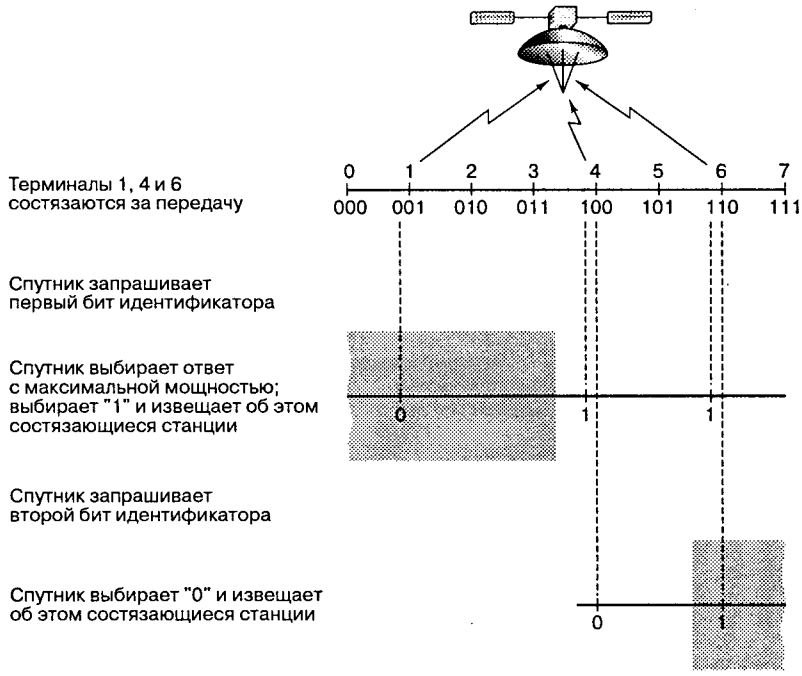


Рис. 11.25. Разрешение состязания между пользователями: поиск по двоичному дереву

1. Спутник запрашивает у состязающихся терминалов первую цифру их двоичных идентификаторов.
2. Терминал 001 передает "0", терминалы 100 и 110, соответственно, "1". Спутник, на основе мощности принятых сигналов, выбирает нуль или единицу. В данном примере была выбрана единица, и об этом были проинформированы пользователи. В настоящий момент половина пользователей прекращает состязание. В данном примере выбывает терминал 001.
3. Спутник запрашивает у оставшихся терминалов вторую цифру идентификационного номера.
4. Терминал 100 передает "0", терминал 110 — "1".
5. Предположим, что спутник выбрал нуль и уведомил об этом пользователей. Терминал 110 выбывает из состязания. Процесс продолжается до тех пор, пока терминал 100 не получит доступ к спутнику.
6. После того как канал связи освобождается, этапы 1–5 повторяются.

Пример 11.3. Сравнение поиска по двоичному дереву и непосредственного опроса

а) Поиск по двоичному дереву требует принятия $n = \log_2 Q$ решений при каждом опросе группы из Q терминалов. Экономия времени возможна в том случае, когда группа является достаточно большой, а среднее количество запросов на услугу невелико. Вычислите время, необходимое для непосредственного опроса группы из 4 096 терминалов, с целью предоставления канала связи 100 терминалам. Сравните результат со временем, необходимым для выполнения 100 операций поиска по двоичному дереву для той же группы пользователей. Время, необходимое для опроса одного терминала, и время принятия решения при поиске по двоичному дереву одинаковы и равны 1 с.

б) Выведите уравнение для максимального количества терминалов Q', при котором время непосредственного опроса равно (или меньше) времени поиска по двоичному дереву.

в) Вычислите Q' для п. а.

Решение

а) Время прямого опроса 4 096 терминалов равно следующему.

$$T = 4096 \times 1 \text{ с} = 4096 \text{ с}$$

Поиск по двоичному дереву для 100 терминалов требует 100 проходов по дереву.

$$T' = (100 \times \log_2 4096) \times 1 \text{ с} = 1200 \text{ с}$$

б) Q' является максимальным числом терминалов, при котором в условиях п. а $T' \leq T$. Это происходит в следующем случае.

$$Q'' \log_2 Q \times 1 \text{ с/решение} = Q \times 1 \text{ с/опрос}$$

$$Q' = \lfloor Q'' \rfloor = \left\lfloor \frac{Q}{\log_2 Q} \right\rfloor \tag{11.30}$$

Здесь $\lfloor x \rfloor$ — наибольшее целое число, не превышающее x.

в) Q' для п. а равно следующему.

$$Q' = \left\lfloor \frac{4096}{\log_2 4096} \right\rfloor = 341 \text{ терминал}$$

Поиск по двоичному дереву для 341 терминала требует 4 092 с.

11.4. Методы множественного доступа, используемые INTELSAT

В 1965 году запуск первого коммерческого геостационарного спутника связи (INTELSAT I или Early Bird) ознаменовал начало новой эпохи телекоммуникаций. 240 каналов передачи речи предоставляли больше возможностей, чем все подводные кабели, проложенные между США и Европой за последние 10 лет [15].

Early Bird представлял собой жестко ограниченный по мощности нелинейный транспондер со схемой FDMA. Результатом одновременного использования нелинейного устройства несколькими сигналами с разными несущими частотами являются сигналы, частоты которых равны всем возможным суммам и разностям исходных частот [16–18]. Потеря энергии сигнала вследствие такой взаимной модуляции — это потеря полезной энергии сигнала. Кроме того, если такие комбинированные сигналы появляются в полосе, принадлежащей другим сигналам, результат аналогичен добавлению к этим сигналам *шума*.

Нелинейный транспондер Early Bird позволяет одновременный доступ к спутнику только двум наземным станциям (одной в Европе, другой — в США). На рис. 11.26 показана передача данных спутником. Три передающие станции в Европе соединены наземной сетью. Каждый месяц одна из них получает прямой доступ к спутнику и управляет процессом обмена данными двух других станций.

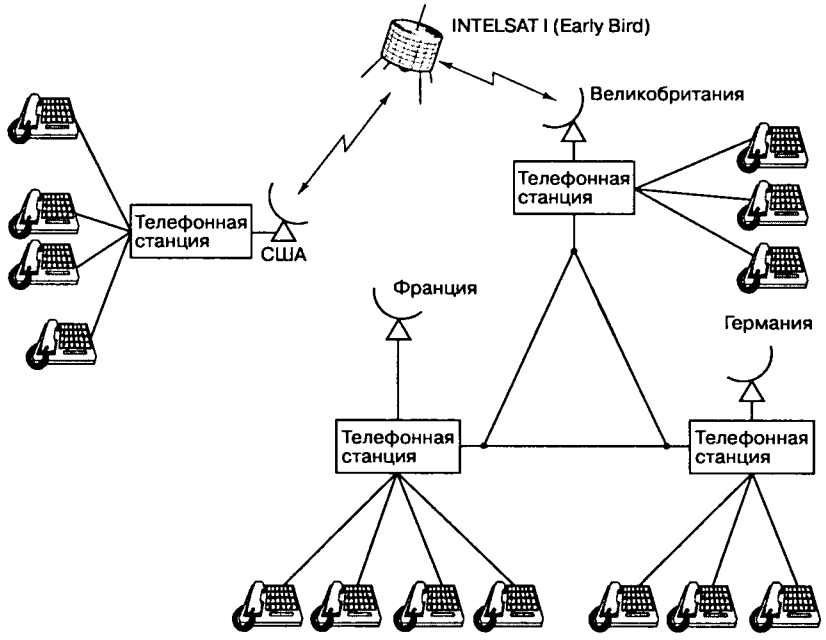


Рис. 11.26. Схема работы спутника INTELSAT I

11.4.1. Режимы работы FDM/FM/FDMA и MCPC

Возможности множественного доступа спутников INTELSAT II и III были значительно улучшены благодаря использованию усилителей на лампах бегущей волны (traveling-wave tube amplifiers — TWTA, ЛБВ), работающих в линейном режиме. Данный метод позволяет удерживать взаимную модуляцию на допустимом уровне и предоставляет одновременный

доступ более чем двум пользователям. (Ценой стало снижение эффективности усилителей мощности). Таким образом, множество частотно-модулированных несущих от различных наземных станций может одновременно получать доступ к спутнику. Такой режим работы называют либо FDM/FM/FDMA с предварительным распределением (или просто FDM/FM), либо многоканальным использованием несущей (multichannel per carrier — MCPC). Данный режим изображен на рис. 11.27. Международные звонки из страны *A* поступают в телефонную сеть и уплотняются в супергруппу (5 групп по 12 каналов передачи речи). Каждая группа супергруппы предварительно выделена наземной станции страны *A* для телефонной информации, адресованной в страны *B–F*. Все эти страны получают сигнал на частоте f_A. В стране-адресате полученный сигнал демодулируется и разуплотняется, причем каждая страна отбирает только те 12 каналов, которые соответствуют этой стране.

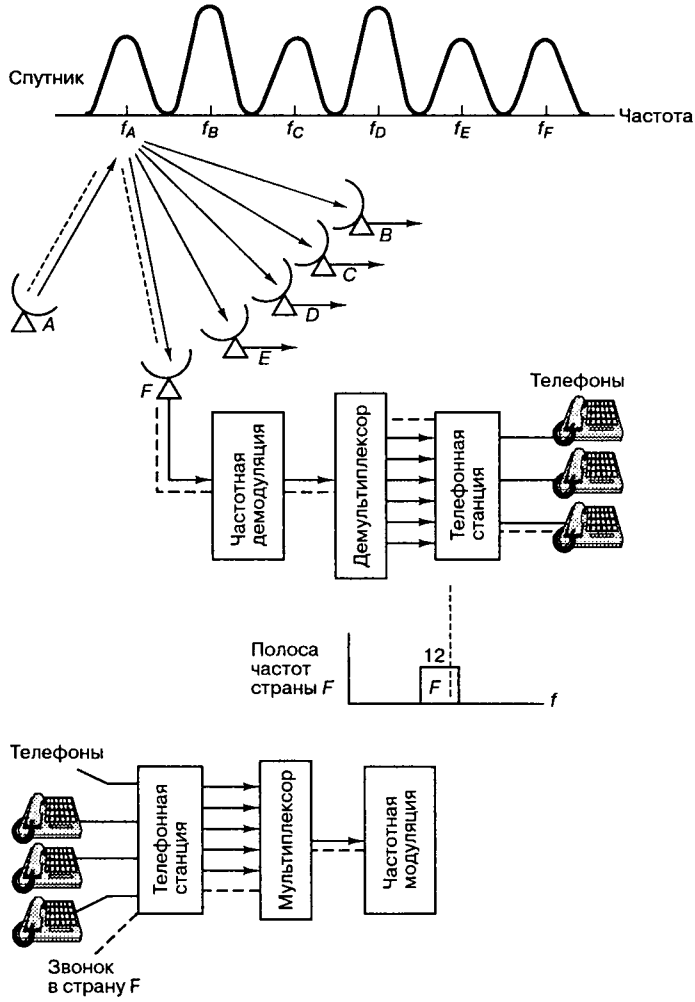


Рис. 11.27. FDM/FM с предварительным распределением. (Перепечатано с разрешения авторов из Puente J. G. and Werth A. M. "Demand-Assigned Service for the INTELSAT Global Network". IEEE Spectrum, January, 1971. © 1971, IEEE.)

11.4.2. MCPC-режимы доступа к спутнику INTELSAT

В настоящее время спутники INTELSAT используют стандартизированные методы совместного использования транспондеров с шириной полосы 36 МГц: множеству пользователей выделяется занимаемая полоса радиочастот и определенное количество каналов шириной 4 кГц. Некоторые стандартные каналы представлены в табл. 11.1. Следует отметить, что пропускная способность транспондера (последний столбец табл. 11.1) снижается по мере увеличения числа несущих. Это можно объяснить следующим образом.

Таблица 11.1. Стандартные режимы доступа INTELSAT MCPC

Число несущих на транспондер	Ширина полосы несущей	Число каналов шириной 4 кГц на несущую	Число каналов шириной 4 кГц на транспондер
1	36 МГц	900	900
4	3 полосы по10 МГц	132	456
	5 МГц	60	
7	5 МГц	60	420
14	2,5 МГц	24	336

1. Между несущими волнами необходимы защитные интервалы. Чем больше несущих волн, тем больше требуется защитных интервалов, что и приводит к снижению пропускной способности.
2. Для нелинейных усилителей на ЛБВ использование большого количества несущих волн приводит к возникновению взаимной модуляции. Если для снижения интерференции усилитель перевести в линейный режим работы, его общая мощность снизится. Канал становится ограниченным по мощности и может обслуживать меньшее число несущих.

Из табл. 11.1 видно, что возможности транспондера будут наиболее эффективны при наличии одной несущей. Почему же тогда INTELSAT не всегда использует транспондеры в таком режиме? Причина в том, что далеко не все наземные передающие станции могут обмениваться данными в таком объеме, чтобы полностью использовать возможности транспондера с шириной полосы 36 МГц. Поэтому применение других режимов позволяет нескольким станциям с небольшими запросами получить одновременный доступ к транспондеру.

11.4.2.1. Ограничения по ширине полосы и мощности

В предыдущем разделе утверждалось, что число поддерживаемых каналов для транспондера с небольшой загрузкой меньше, чем для транспондера, работающего в режиме насыщения. Полезно рассмотреть два условия работы спутникового транспондера: режимы с ограничениями по ширине полосы и мощности. На рис. 11.28 представлен транспондер с шириной полосы 36 МГц и максимальной выходной мощностью 20 Вт. На рис. 11.28, *а* изображено совместное использование четырьмя несущими волнами полосы шириной 36 МГц в режиме MCPC. Предположим, каждая несущая требует 4 Вт выходной мощности. Тогда полная выходная мощность равна 16 Вт (меньше максимальной мощности усилителя); следовательно, возможности транспондера используются не полностью. В то же

время, помимо существующих пользователей, доступ к полосе 36 МГц не может получить никто. Данный пример — это случай ограничения по ширине полосы.

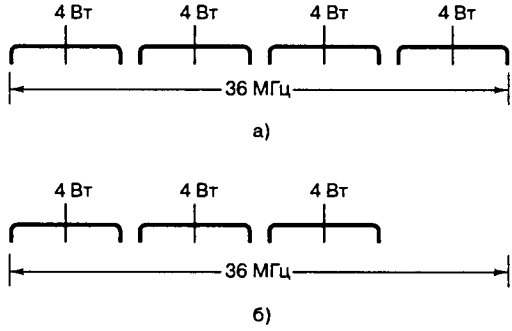


Рис. 11.28. Различные условия работы: а) ограниченная ширина полосы; б) ограниченная мощность

Предположим, что в предыдущем примере возникла существенная взаимная модуляция и необходимо перевести транспондер в линейный режим путем снижения максимальной выходной мощности до 12 Вт. При этом транспондер уже не может поддерживать связь с четырьмя пользователями, каждому из которых требуется 4 Вт мощности. Один из пользователей должен быть "отключен", что показано на рис. 11.28, *б*. В данном примере ширина полосы позволяет доступ еще одного пользователя, но для этого недостаточно выходной мощности. Другими словами, имеем случай ограниченной выходной мощности.

11.4.3. Работа алгоритма SPADE

Схема множественного доступа MCPC с предварительным распределением эффективна при достаточно интенсивном обмене данными, когда каналы используются практически полностью. В то же время, если в группе из 12 каналов используется только один, остальные 11 выключить нельзя. Передача данных по схеме FDM/FM осуществляется вместе с телефонными сигналами или без них. Следовательно, долгосрочное распределение несущих для систем с недостаточно интенсивным обменом данными нерационально. Поэтому для систем с большим числом подобных слабо нагруженных каналов был необходим гибкий механизм обслуживания. Также требовался метод управления перегрузками в процессе обмена данными для линий средней мощности. При такой постановке задачи решением стал усовершенствованный алгоритм DAMA, получивший название SPADE. Впервые схема SPADE использовалась в системе INTELSAT IV. Перевод с английского аббревиатуры SPADE звучит как "оборудование импульсно-кодовой модуляции с множественным доступом с распределением запросов по требованию и одноканальным использованием несущей" (single-channel-per-carrier PCM multiple access demand assignment equipment). Ниже перечислены основные характерные особенности схемы SPADE [15].

1. Отдельный канал передачи речи со скоростью 64 Кбит/с преобразовывается из аналоговой формы в цифровую.
2. Полученный узкополосный цифровой сигнал модулирует несущую с использованием квадратурной фазовой манипуляции (quadrature phase shift keying — QPSK).

В отличие от метода MCPC, для каждой несущей волны существует *только один* речевой канал.

3. Расстояние между каналами равно 45 кГц. На транспондере доступно 800 несущих каналов. Шесть из них резервируются системой; таким образом, для использования доступны 794 канала.
4. Несущие распределяются динамически *по требованию.*
5. Динамическое распределение осуществляется с помощью канала общего доступа (common signaling channel — CSC) с шириной полосы 160 кГц. Скорость передачи данных в канале CSC равна 128 Кбит/с, в качестве модуляции используется двоичная фазовая манипуляция (binary phase shift keying — BPSK).

На рис. 11.29 изображено распределение частот канала CSC, а также 800 несущих системы SPADE. Рассмотрим использование алгоритма SPADE, изображенного на рис. 11.30. Канал CSC работает в широковещательном режиме TDMA с фиксированным распределением. Все наземные станции наблюдают за каналом CSC и получают информацию о текущем распределении каналов. Каждой станции в канале CSC выделяется временной интервал 1 мс (один раз в каждые 50 мс) для отправки запроса на выделение канала или сообщения об освобождении канала. Когда наземной станции требуется канал, она "захватывает" произвольный свободный канал (пару частот) и сообщает о своем выборе через канал CSC. Произвольный выбор позволяет снизить вероятность одновременного запроса одного канала двумя станциями. Вероятность такого события возрастает, если количество незанятых каналов мало. После того как наземная станция получает доступ к каналу, остальные станции исключают его из списка доступных каналов. Изменения в список вносятся через канал CSC. Таким образом, управление доступом в схеме SPADE *распределено* между всеми наземными станциями.

По окончании сеанса связи станция освобождает канал, отправляя во время выделенного интервала времени соответствующий сигнал через канал CSC. Этот сигнал получают все станции, после чего в соответствующем списке помечают освободившийся канал как доступный. Если две станции пытаются одновременно получить доступ к одному каналу — обе получают сигнал, что канал занят. После этого станции повторяют запрос, выбирая произвольным образом один из доступных каналов.

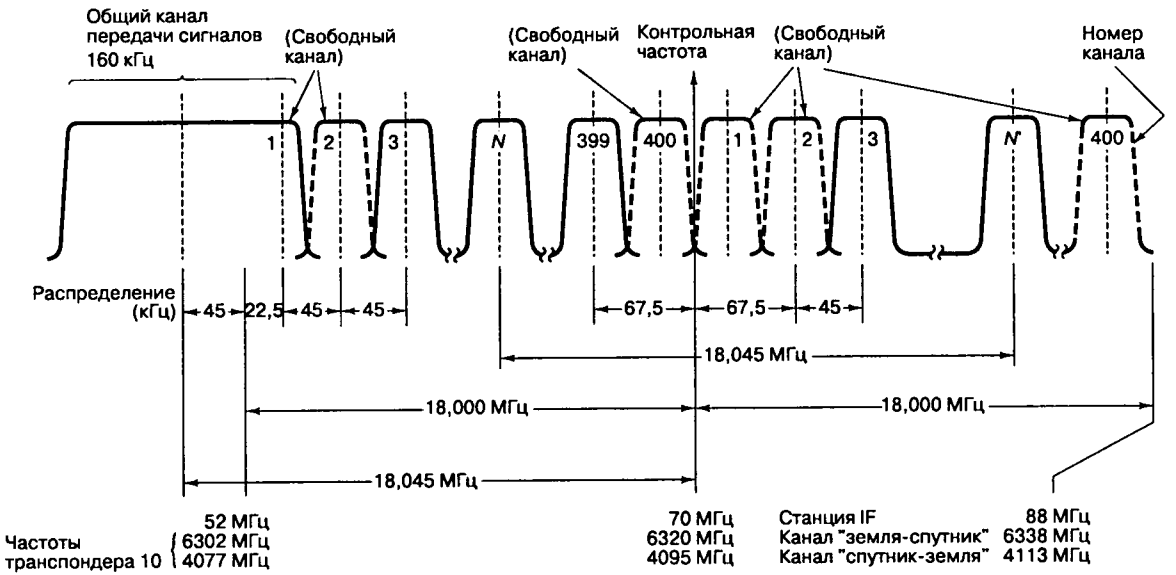


Рис. 11.29. Распределение частот при использовании алгоритма SPADE. (Перепечатано с разрешения авторов из Puente J. G. and Werth A. M. "Demand-Assigned Service for the INTELSAT Global Network". IEEE Spectrum, *January, 1971. © IEEE.)

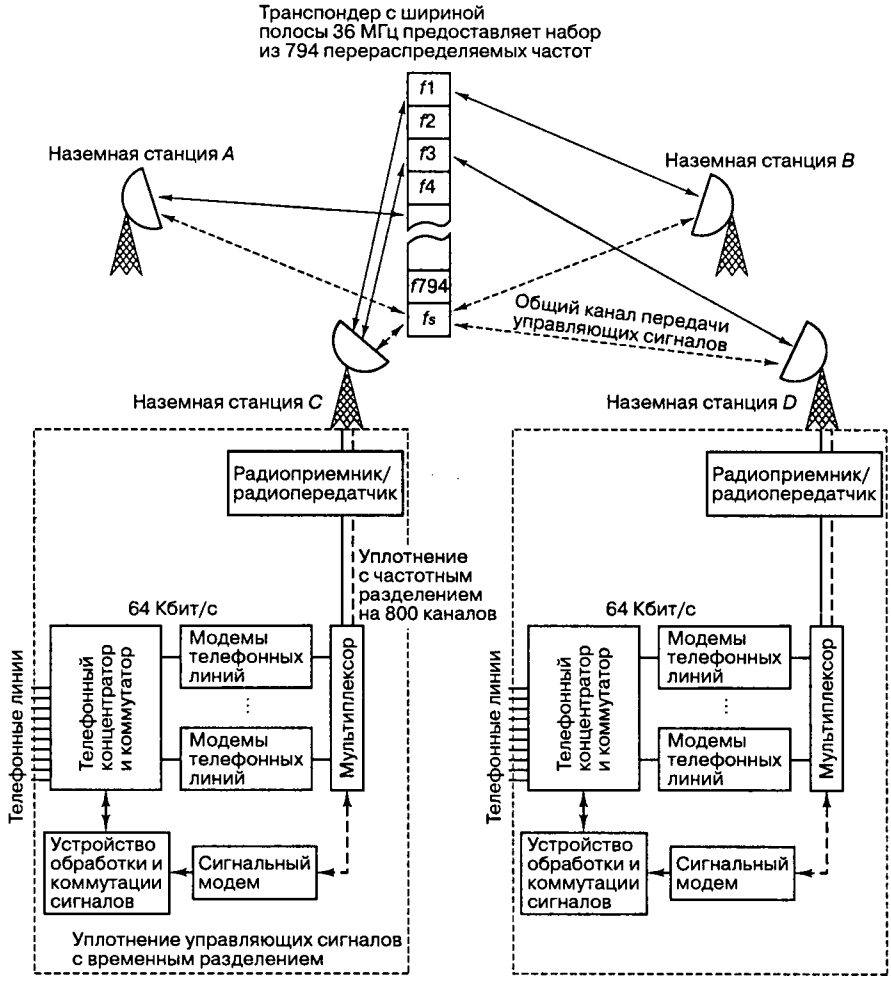


Рис. 11.30. Работа системы SPADE. (Перепечатано с разрешения издательства Prentice-Hall, Englewood Cliffs, N. J. из James Martin, Communications Satellite Systems, *Fig. 15.2, p. 236. © 1978.)*

11.4.3.1. Использование пропускной способности транспондера при выборе схемы SPADE

Из табл. 11.2, которая является продолжением табл. 11.1, видим, что использование полосы транспондера при выборе алгоритма SPADE дает общую пропускную способность 800 каналов передачи речи на транспондер. Сравним данные, приведенные в табл. 11.1 и 11.2. В первом случае по мере роста числа несущих от 1 до 14 полное число каналов уменьшается с 900 до 336. Почему же тогда система SPADE не дает меньшую пропускную способность, чем система с 336 каналами, связанными с 14 несущими? Причина в следующем — когда на каждую несущую приходится только один канал передачи речи, несущая может быть отключена, если голосовой сигнал отсутствует. Даже если работают все каналы, они могут отключаться приблизительно 60%

всего времени. Поскольку мощность транспондера ограничена, ее экономия позволяет использовать для передачи больше каналов. Кроме того, SPADE применяет цифровую передачу речи (схема QPSK). Эффективность использования полосы системы соответствует получаемой при использовании схемы FDM/FM с одной несущей.

Таблица 11.2. Режимы доступа SPADE

Количество несущих на транспондер	Ширина полосы несущей	Число каналов шириной 4 кГц на несущую	Число каналов шириной 4 кГц на транспондер
800	45 МГц	1	800

11.4.3.2. Эффективность схемы SPADE

При использовании схемы MCPC пропускная способность системы распределяется заранее, и неиспользуемые каналы не могут перераспределяться. Система SPADE является модификацией системы DAMA, где все каналы используются совместно. Каналы выделяются пользователю, когда в них действительно возникает необходимость. Важной мерой качества телефонной системы, называемой вероятностью блокировки, является вероятность недоступности запрошенного канала. Для получения 1% вероятности блокировки системы MCPC необходимо в четыре раза больше каналов, чем для SPADE. По этому параметру транспондер SPADE с 800 каналами эквивалентен транспондеру MCPC с 3200 каналами [15].

11.4.3.3. Сеть наземных станций разной мощности с использованием SPADE

Стандартная наземная станция INTELSAT характеризуется чувствительностью приемника $G/T° = 40{,}7$ дБ/К, тогда как станции меньшего размера имеют $G/T° = 35$ дБ/К. Если 125 каналов SPADE выделены для использования малыми станциями, общая пропускная способность транспондера снижается до 525 каналов. В данном случае половина доступных ресурсов транспондера применяется для обслуживания стандартных станций. Связь пропускной способности транспондера и числа каналов, используемых малыми станциями, показана на рис. 11.31. Лучшим пояснением для этого рисунка может служить рис. 11.32. На рис. 11.32, *а* представлен случай, когда вся мощность усилителя на ЛБВ используется для обслуживания крупных станций, транспондер с шириной полосы 36 МГц поддерживает приблизительно 800 несущих, каждая из которых имеет мощность x дБВт (в данном случае имеем дело с ограниченной шириной полосы). На рис. 11.32, *б* показана другая ситуация: для обслуживания малых станций требуется половина мощности, для использования стандартными станциями резервируется половина исходных несущих (400) с уровнем мощности x дБВт каждая. Рассмотрим оставшиеся 400 несущих. В главе 5 показывалось, что вероятность ошибок, возникающих в канале связи, прямо связана с произведением EIRP и $G/T°$. Для любого канала можно достичь приемлемого компромисса между этими параметрами, поддерживая таким образом фиксированный уровень вероятности ошибки. Поскольку отношение $G/T°$ для малой станции на 5,7 дБ меньше, чем для стандартной станции, малой станции необходимо обеспечить на 5,7 дБ большую мощность EIRP, чтобы уравновесить производительность станций. *Увеличение* мощности несущей для малой передающей станции приводит к соответствующему *снижению* количества несущих. В результате, вместо 400 несущих для обслуживания малых станций используется 125 (снижение на 5,1 дБ); транспондер становится ограниченным по мощности.

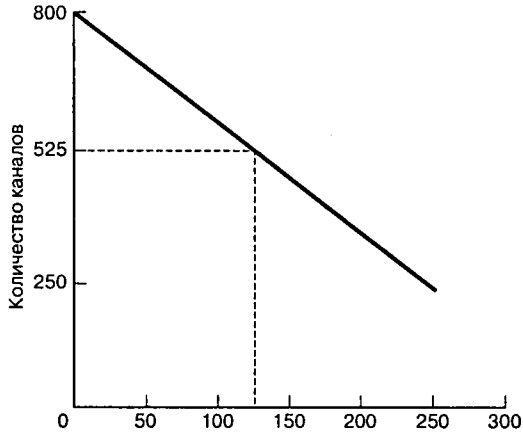


Рис. 11.31. Пропускная способность транспондера SPADE в сети наземных станций различной мощности

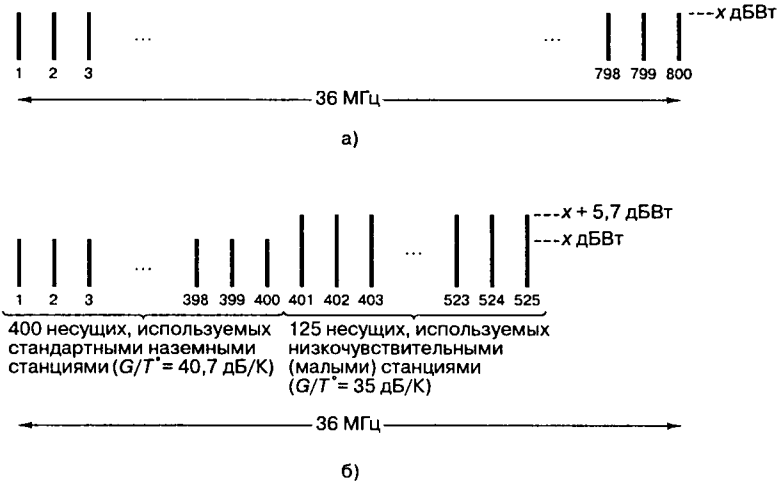


Рис. 11.32. Сеть наземных станций различной мощности: а) полная мощность усилителя на ЛБВ используется для обслуживания крупных станций; ограничение по ширине полосы (800 каналов); б) половина мощности усилителя на ЛБВ применяется для обслуживания малых станций; ограничение по мощности (525 каналов)

В момент выделения канала по запросу передающая станция получает информацию о размере станции-адресата. Напомним, что данные спутники являются нерегенеративными, поэтому пропорциональное разделение мощности EIRP канала связи "спутник-земля" выполняется передающей станцией (см. раздел 5.7.1). Передающая станция устанавливает свой уровень мощности в зависимости от потребностей станции-адресата.

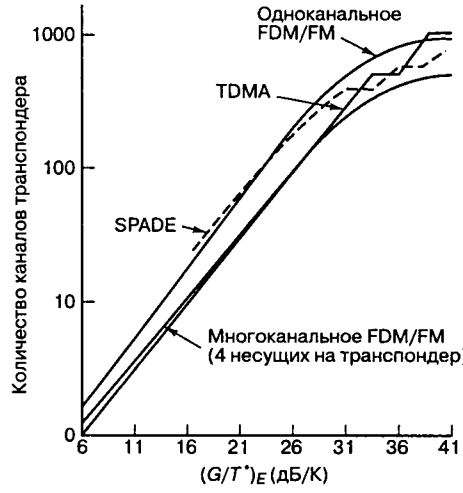
11.4.4. Использование TDMA в системах INTELSAT

В первом поколении систем связи множественного доступа преобладали системы с использованием FDMA. В настоящее время, благодаря наличию точных схем синхро-

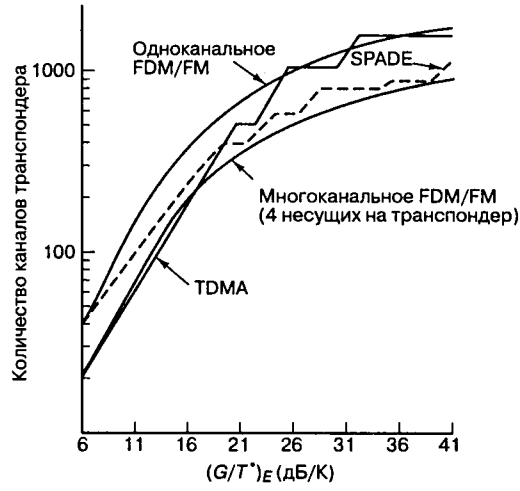
низации и высокоскоростных коммутирующих элементов, предпочтение отдается технологии TDMA [19–24]. В INTELSAT IV для управления сетью SPADE применялся канал CSC со скоростью 128 Кбит/с, в котором использовалась схема TDMA. Для многолучевой международной системы цифровой связи в спутник INTELSAT V была введена схема TDMA со скоростью передачи данных 120 Мбит/с. Одним из недостатков реализации схемы TDMA является необходимость точной *синхронизации* всех наземных станций и спутника. Системы FDMA, не имеющие такого требования, значительно проще с точки зрения работы с сетью. Ниже приводятся основные преимущества и недостатки схем TDMA и FDMA.

1. Применение FDMA может привести к возникновению взаимной модуляции. Во избежание этого усилитель на ЛБВ должен работать в линейной области, снижая тем самым номинальную мощность.
2. При использовании TDMA на усилителе может находиться только одна несущая. Поэтому возникновение взаимной модуляции невозможно.
3. Оборудование наземной станции TDMA сложнее и потому дороже оборудования для станции FDMA. В то же время для наземных станций FDMA, использующих множественные двухточечные каналы, требуется выполнение особых этапов обработки сигналов — преобразование с переносом частоты в область радиочастот и обратное преобразование. Следовательно, при применении схемы FDMA растет число единиц оборудования и требуемых соединений между ними. При использовании схемы TDMA этого не происходит, поскольку выбор канала осуществляется по времени, а не по частоте. Таким образом, для наземных станций с большим количеством соединений более рентабельна схема TDMA, а не FDMA.
4. В многолучевых системах может возникать необходимость установления связи одного луча со всеми остальными. TDMA предоставляет возможности создания удобного последовательного соединения, такого как TDMA со спутниковой коммутацией (satellite-switched TDMA — SS/TDMA). Использование SS/TDMA на спутнике INTELSAT VI описывается в разделе 11.4.5.

На рис. 11.33 в виде графика зависимости пропускной способности канала от отношения G/T° наземной станции приведена сравнительная производительность схем TDMA, FDM/FM и SPADE для транспондера INTELSAT IV. Рис. 11.33, *а* соответствует антенне обзора земной поверхности, а рис. 11.33, *б* — сфокусированной антенне. При одинаковом расположении ширина луча половинной мощности составляет, соответственно, 17° и 4,5°. Из графиков видно, что схема FDM/FM с одной несущей так же эффективна, как и схема TDMA, если система работает со стандартными наземными станциями ($G/T^\circ = 40{,}7$ дБ/К). Для меньших станций ($G/T^\circ \leq 31$ дБ/К), использующих транспондеры обзора земной поверхности, метод SPADE эффективнее TDMA и FDM/FM со множественными несущими (MCPC) (на рисунке изображен график для четырех несущих). Для обычных наземных станций (G/T° лежит в диапазоне 19–40,7 дБ/К), использующих сфокусированные транспондеры, схема TDMA значительно выгоднее схем SPADE и MCPC. Для меньших станций (G/T° от 6 до 19 дБ/К), использующих сфокусированный транспондер, схема SPADE значительно лучше схем TDMA и MCPC. Вообще, при работе со *стандартными* наземными станциями наиболее эффективным методом множественного доступа к спутнику системы INTELSAT IV является применение схемы TDMA [19].



а)



б)

Рис. 11.33. Зависимость пропускной способности от отношения G/T° наземной станции для схем FDMA, TDMA и SPADE: а) пропускная способность канала транспондера обзора земной поверхности как функция $(G/T^\circ)_E$, где $(G/T^\circ)_E$ означает (G/T°) наземной станции; б) пропускная способность канала сфокусированного транспондера как функция $(G/T^\circ)_E$. (Из работы Chakraborty D. "INTELSAT IV Satellite System (Voice) Channel Capacity versus Earth Station Performance". IEEE Trans. Commun. Tech., *vol. COM19, n. 3, June, 1971, pp. 355–362. © 1971, IEEE.)*

11.4.4.1. Структуры кадров уплотнения PCM

В настоящее время используется два цифровых стандарта телефонной связи для структуры кадра PCM. Североамериканский стандарт называется *T-Carrier*; в его основе лежит 193-битовый кадр, показанный на рис. 11.34, *а*. Всего существует 24 канала, каждый из которых содержит восьмибитовую выборку речи. Кроме того, для цикличной синхронизации используется один бит кадра, значение которого чередуется от кадра к кадру (1 0 1 0 ...). Поскольку телефонный канал передачи речи имеет ширину 4 кГц (включая защитные полосы), частота дискретизации Найквиста для восстановления аналоговой информации в диапазоне 4 кГц равна $f_S = 2W = 8000$ выборок/с. Следовательно, основной кадр PCM, называемый *кадром Найквиста* (Nyquist frame), содержит 24 выборки речи из 24 различных источников информации и передается со скоростью 8000 кадров/с (1 кадр за 125 мкс). Таким образом, скорость передачи битов при использовании стандарта T-Carrier равна 193 бит/кадр × 8000 кадров/с = 1,544 Мбит/с.

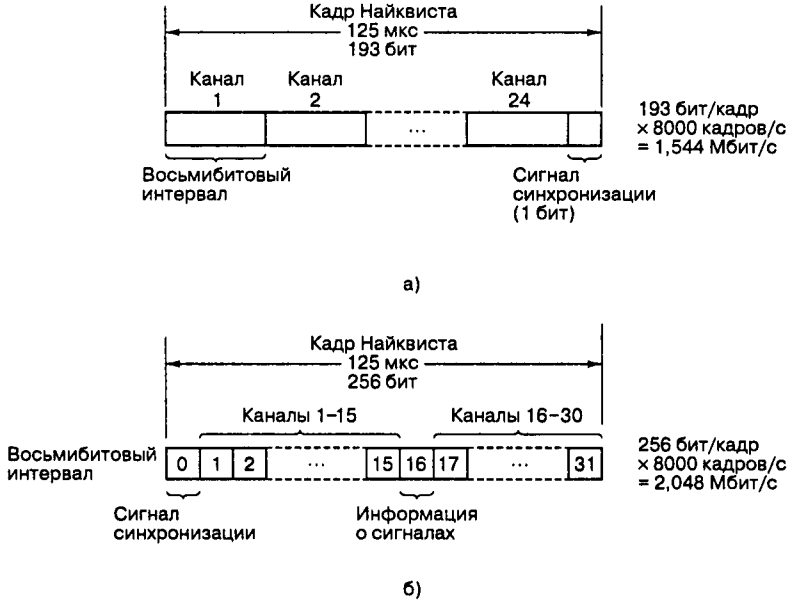


Рис. 11.34. Структура кадров уплотнения PCM: а) стандарт T-Carrier (Северная Америка); б) европейский стандарт

Европейский стандарт создан на основе 256-битового кадра, показанного на рис. 11.34, *б*. Существует 30 каналов передачи сообщений, каждый из которых содержит восьмибитовую выборку речи. Кроме того, для цикличной синхронизации используется один 8-битовый интервал, а другой 8-битовый интервал применяется для передачи информации по адресу. Скорость передачи кадров для обоих описанных стандартов одинакова. Следовательно, скорость передачи для европейского стандарта равна 256 бит/кадр × 8000 кадров/с = 2,048 Мбит/с.

11.4.4.2. Высокоскоростной кадр TDMA европейского стандарта

На рис. 11.35, *а* показано 16 кадров Найквиста европейского формата уплотнения сигналов PCM. Каждый кадр содержит 8-битовую выборку от каждого из 30 наземных каналов связи, а также 8 бит служебной информации и 8 бит данных о сигнале. Длительность такого кадра TDMA равна следующему.

16 кадров Найквиста × 125 мкс/кадр Найквиста = 2 мс

В течение этих 2 мс передается

16 кадров Найквиста × 256 бит/кадр Найквиста = 4096 бит.

Одной из основ схемы TDMA является возможность совместного доступа к ресурсу связи пользователей, передающих низкоскоростные потоки данных, путем пакетной передачи с более высокой скоростью, чем могут давать отдельные пользователи. На рис. 11.35, *б* представлен высокоскоростной кадр TDMA длительностью 2 мс. Кадр начинается с опорного пакета (RB1), передаваемого опорной станцией. В пакете содержится информация, которая позволяет другим станциям правильно разместить свои данные в кадре. Кроме того, для повышения надежности может быть использован второй опорный пакет, RB2, за которым следует последовательность слотов данных. Эта последовательность может упорядочиваться заранее или же распределяться согласно протоколу DAMA [20].

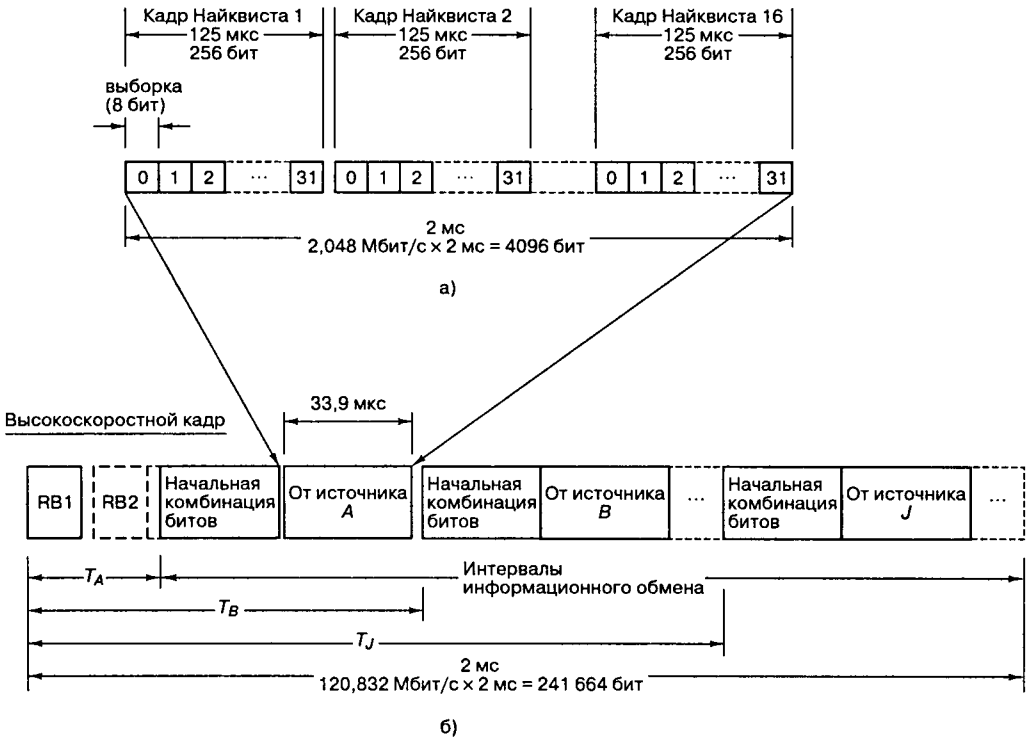


Рис. 11.35. Европейские стандарты цифровой передачи для спутника INTELSAT: а) наземное уплотнение сигналов PCM; б) высокоскоростной кадр

Уплотненный сигнал PCM со скоростью передачи $R_0 = 2{,}048$ Мбит/с и длительностью кадра $T = 2$ мс сжимается (в 59 раз), после чего передается с использованием модуляции QPSK со скоростью $R_T = 120{,}832$ Мбит/с (или 60,416 миллионов символов в секунду). Длительность поля данных T_{tr} для высокоскоростного кадра TDMA вычисляется следующим образом.

$$T_{tr} = \frac{R_0 T}{R_T} = \frac{2{,}048 \times 10^6 \times 2 \times 10^{-3}}{120{,}832 \times 10^6} = 33{,}9 \text{ мкс} \tag{11.31}$$

Для расчета полной продолжительности пакета данных необходимо учесть время, затраченное на передачу начальной комбинации данных. Если начальная комбинация состоит из S_P символов, то, предполагая модуляцию QPSK, общая длина пакета символов, выраженная в символах, равна следующему.

$$S_T = \frac{R_0 T}{2} + S_P \tag{11.32}$$

Длительность пакета равна следующей величине.

$$T_T = \frac{2S_T}{R_T} \tag{11.33}$$

Если начальная комбинация содержит 300 символов, тогда получаем следующее.

$$S_T = \frac{2{,}048 \times 10^6 \times 2 \times 10^{-3}}{2} + 300 = 2348 \text{ символов}$$

Подставляя это число в уравнение (11.33), получим следующее.

$$T_T = \frac{2 \times 2348}{120{,}832 \times 10^6} = 38{,}9 \text{ мкс}$$

11.4.4.3. Высокоскоростной кадр TDMA североамериканского стандарта

Скорость передачи данных (пакетов TDMA) $R_T = 120{,}832$ Мбит/с в системе INTELSAT соответствует европейскому и североамериканскому стандартам. Рис. 11.36 подобен рис. 11.35, за исключением того, что уплотненный сигнал PCM разбит на 24 канала (стандарт T-Carrier), а не на 30 (европейский стандарт). Перечислим важные отличительные особенности стандарта T-Carrier.

1. Каждый кадр Найквиста состоит из 24 каналов (или выборок) × 8 бит + 1 бит цикличной синхронизации = 193 бит.
2. 16 кадров Найквиста содержат 16 × 193 = 3088 бит.
3. Скорость передачи данных T-Carrier равна 1,544 Мбит/с.
4. Длительность информационного поля кадра в высокоскоростном кадре TDMA вычисляется из уравнения (11.31).

$$T_{tr} = \frac{1{,}544 \times 10^6 \times 2 \times 10^{-3}}{120{,}832 \times 10^6} = 25{,}6 \text{ мкс}$$

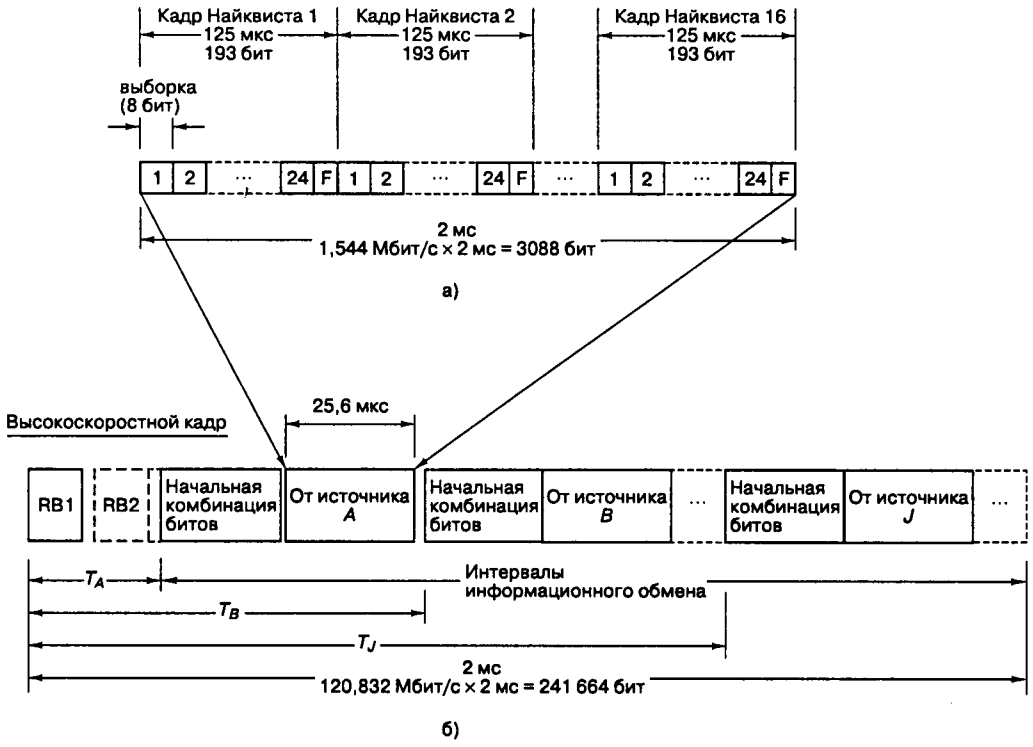


Рис. 11.36. Стандарты цифровой передачи T-Carrier для спутника INTELSAT: а) наземное уплотнение сигналов PCM; б) высокоскоростной кадр

11.4.4.4. Работа спутника INTELSAT с использованием схемы TDMA

На передающей наземной станции непрерывный низкоскоростной поток данных поступает на один из пары буферов, изображенных на рис. 11.37, *а*. В то время как первый буфер заполняется данными с низкой скоростью (1,544 или 2,048 Мбит/с), второй очищается с высокой скоростью (120,832 Мбит/с). В каждом кадре функции буферов чередуются. Благодаря использованию быстродействующего счетчика, пакеты передаются в надлежащие интервалы времени и прибывают на спутник в выделенный им момент времени (согласно схеме TDMA).

В принимающей станции поток кадров направляется к одному из пары буферов расширения (рис. 11.37, *б*), функции которых обратны по отношению к функциям буфера сжатия (рис. 11.37, *а*). Пока один буфер на высокой скорости заполняется данными, другой освобождается с желаемой выходной скоростью.

Основной проблемой в работе TDMA является необходимость точной синхронизации для достижения ортогональности временных интервалов [20]. На рис. 11.38 приведена иллюстрация общего принципа, используемого в большинстве коммерческих схем синхронизации спутников. Одна из наземных станций назначается главной (или управляющей). Эта станция передает периодические пакеты импульсов эталонного времени. Пользовательские станции также передают собственные тактовые импульсы, обозначенные на рис. 11.38 как "подчиненные". По каналу "спутник-земля" станция, в дополнение к собственным тактовым импульсам, получает эталонные импульсы

управляющей станции. Разность во времени между этими импульсами соответствует ошибке синхронизации. Для ее снижения наземные станции должны регулировать собственные схемы синхронизации.

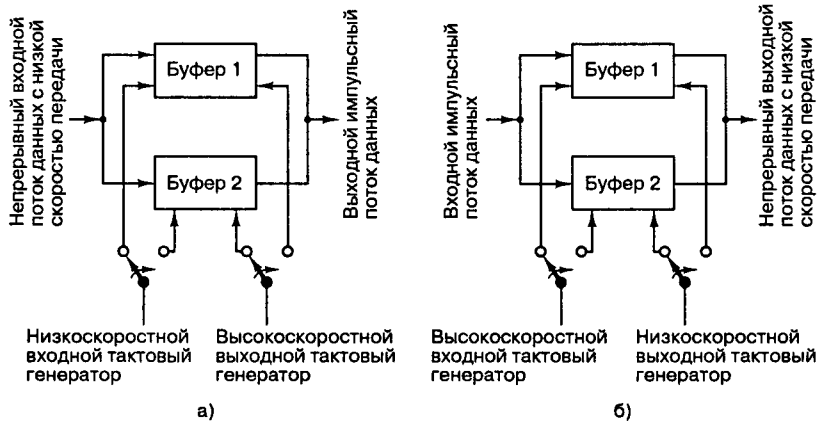


Рис. 11.37. Буферы сжатия и расширения пакетов: а) буферы сжатия в передатчике; б) буферы расширения в приемнике

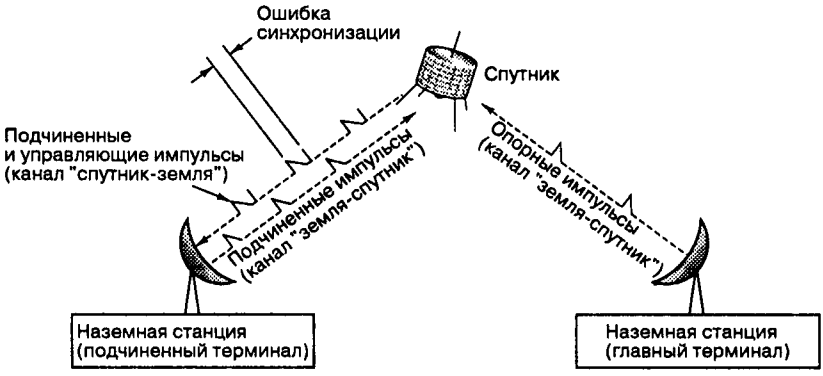


Рис. 11.38. Принцип синхронизации TDMA

11.4.5. Использование схемы TDMA со спутниковой коммутацией на спутнике INTELSAT

Современные спутники связи обычно используют несколько лучей, обеспечивающих покрытие в определенном регионе. К примеру, если спутник находится над Атлантическим океаном, отдельные лучи могут быть направлены в Северную Америку, Европу, Южную Америку и Африку. Для взаимосвязи станций различных регионов используются коммутаторы. Основной целью схемы TDMA со спутниковой коммутацией (satellite-switched TDMA — SS/TDMA) является обеспечение эффективной циклической взаимосвязи данных TDMA из областей охвата различных спутников.

Основой системы служит расположенная на спутнике микроволновая матрица коммутации, программируемая посредством наземного управления на циклическое изменение состояний. Таким образом, в каждый момент коммутации связываются раздельные лучи каналов "земля-спутник". Наземная станция может связаться со станциями, использую-

щими другой луч, посылая пакеты TDMA во время соответствующих выделенных интервалов времени. Схема коммутации состояний выбирается так, чтобы максимально увеличить пропускную способность системы с учетом существующих ограничений по обмену данными [21]. Для достижения полной взаимосвязанности N лучей, требуется $N!$ различных состояний или *режимов* спутника. В табл. 11.3 показаны шесть режимов, необходимых для полной взаимосвязанности трехлучевой системы.

Таблица 11.3. Режимы коммутации трехлучевого спутника

	Выход					
Вход	**Режим 1**	**Режим 2**	**Режим 3**	**Режим 4**	**Режим 5**	**Режим 6**
A	*A*	*A*	*B*	*B*	*C*	*C*
B	*B*	*C*	*A*	*C*	*A*	*B*
C	*C*	*B*	*C*	*A*	*B*	*A*

В режиме 1 приемники спутника на лучах A, B и C соединены с передатчиками для лучей A, B и C. Наземная станция, использующая один из этих лучей, может связаться с другой станцией, использующей тот же луч. Такой луч называют *самоориентированным.*

На рис. 11.39 представлен пример трехлучевой (лучи A, B и C) системы SS/TDMA. Микроволновая матрица коммутации для данного спутника является *координатной.* Такая матрица может быть представлена как набор продольных и поперечных линий. При активизации линий, одной продольной и одной поперечной, возникает контакт на пересечении. Координатный коммутатор позволяет одновременно устанавливать связь только между двумя компонентами матрицы, одним продольным и одним поперечным. Если канал станции A_U связан с каналом станции B_D, *ни один из этих каналов* не может быть одновременно связан с каким-либо другим каналом.

На рис. 11.39 показаны три схемы процедуры обмена данными в течение интервалов времени T_1, T_2 и T_3 при существовании трех состояний коммутации S_1, S_2 и S_3. В течение интервала T_1 имеем режим S_1: лучи самоориентированы. В течение интервала T_2 режим S_2 позволяет передать сигналы со станций A_U, B_U и C_U на станции B_D, C_D и A_D. На протяжении интервала T_3 (режим S_3) каналы передачи подобным образом связываются с каналами приема, что позволяет обеспечить доставку данных требуемому адресату.

Схемы процедуры обмена данными, а также их длительность выбираются с целью оптимизации пропускной способности спутника и максимально эффективного обслуживания пользователей. Для учета изменений в информационном потоке циклическая схема в случае необходимости может изменяться наземной станцией.

11.4.5.1. Матрица информационного обмена

На рис. 11.40 представлена матрица, характеризующая обмен данными между N областями, обслуживаемыми сфокусированным лучом. На данном рисунке t_{ij} — объем информационного потока от луча i к j. Промежуточная сумма

$$S_i = \sum_{j=1}^{N} t_{ij} \tag{11.34}$$

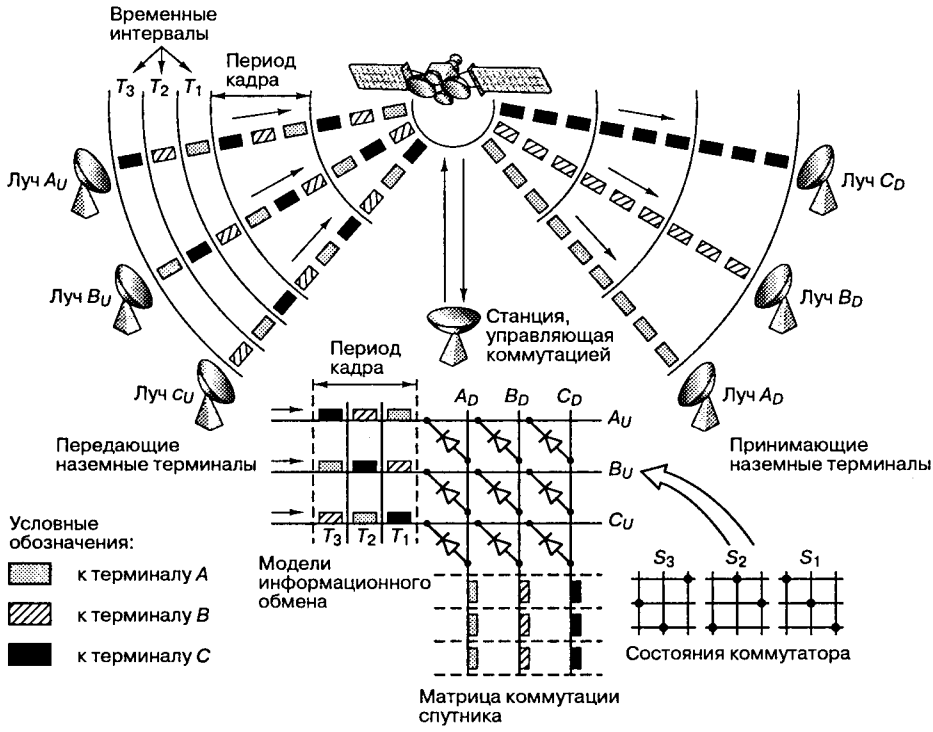


Рис. 11.39. TDMA со спутниковой коммутацией (satellite-switched TDMA — SS/TDMA)

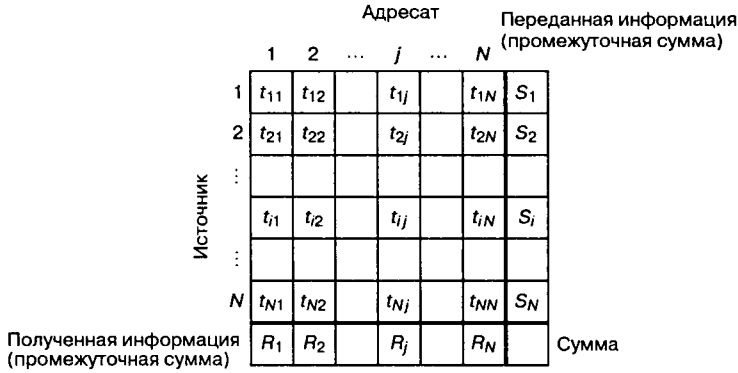


Рис. 11.40. Матрица информационного обмена

является полным информационным потоком от i-го луча наземной станции, а

$$R_j = \sum_{i=1}^{N} t_{ij} \tag{11.35}$$

полным информационным потоком к j-му лучу наземной станции. Если обмен данными системы SS/TDMA управляется неблокирующим коммутатором (позволяющим передачу *всех* сообщений без выдачи какого-либо аналога сигнала "занято"), каждому

каналу в кадре TDMA назначается временной интервал длительностью k секунд. Для эффективного использования ресурса связи полный информационный обмен на рис. 11.40 должен быть выполнен в течение времени кадра T, которое должно быть как можно меньше. Минимальное время передачи кадра для обеспечения подобной неблокирующей связности можно выразить следующим образом [22].

$$T_{\min} = k \max(\{S_i\},\{R_j\}) \tag{11.36}$$

Здесь $\max(\{S_i\},\{R_j\})$ — максимальное значение, выбранное из всех возможных $\{S_i\}$ и $\{R_j\}$. Выражение (11.36) описывает минимальное время, необходимое для передачи *всех* данных *всем* адресатам (и то, и другое указано в матрице информационного обмена), если все каналы имеют полосы равной ширины.

11.5. Методы множественного доступа в локальных сетях

Локальные сети (local area network — LAN) могут использоваться для связи компьютеров, терминалов, принтеров и других устройств, расположенных недалеко друг от друга (например, в одном здании). Если из экономических соображений в глобальных сетях применяются телефонные сети общего пользования, то для создания локальных сетей обычно устанавливаются собственные кабели высокой пропускной способности. Следовательно, в последнем случае ширина полосы не является столь "дефицитным" ресурсом, как при глобальных сетях. Поскольку в оптимизации использования полос нет необходимости, в системах локальных сетей могут применяться простые алгоритмы доступа [6, 25–27].

11.5.1. Сети CSMA/CD

Схема Ethernet, представляющая собой метод доступа для локальных сетей, была разработана корпорацией Xerox. Данный метод основывается на предположении, что каждое локальное устройство может узнать состояние общего широковещательного канала связи перед попыткой его использования. Такой метод называется *множественным доступом к среде с обнаружением конфликтов и детектированием несущей* (carrier-sense multiple access with collision detection — CSMA/CD). В данном случае "несущая" означает *любую* электрическую активность в кабеле. На рис. 11.41, *а* изображен формат битового поля данных для спецификации Ethernet. Пояснения к рисунку приводятся ниже.

1. Максимальный размер пакета равен 1526 байт, где байт включает 8 бит. Структура пакета является следующей: начальная комбинация битов (8 байт) + заголовок (14 байт) + данные (1500 байт) + биты четности (4 байт).
2. Минимальный размер пакета равен 72 байт. Пакет включает начальную комбинацию битов (8 байт) + заголовок (14 байт) + данные (46 байт) + биты четности (4 байт).
3. Минимальная пауза между пакетами равна 9,6 мкс.
4. Начальная комбинация битов содержит 64-битовый шаблон синхронизации, состоящий из чередующихся единиц и нулей, причем два последних символа — единицы: (1 0 1 0 1 0 ...1 0 1 0 1 1).
5. Принимающая станция изучает поле адреса в заголовке пакета, после чего решает, принимать ли ей этот пакет. Первый бит указывает тип адреса (0 — индиви-

дуальный адрес, 1 — групповой). Поле, состоящее из одних единиц, обозначает широковещание на все станции.

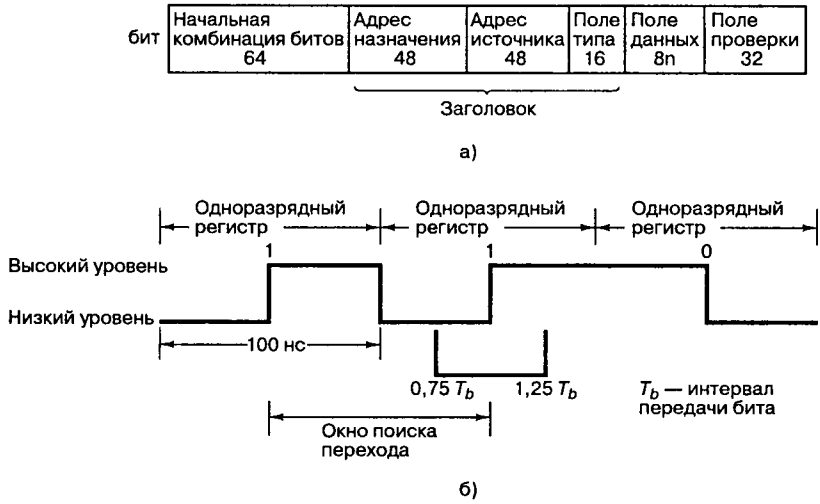


Рис. 11.41. Поле данных и формат PCM схемы Ethernet: а) спецификация Ethernet; б) формат манчестерской модуляции PCM

6. Адрес источника — это уникальный адрес передающей машины.
7. Тип поля определяет, как необходимо интерпретировать поле данных. Например, биты поля могут использоваться для описания кодировки данных, шифрования, приоритета сообщения и т.д.
8. Поле данных состоит из целого числа байт (минимум — 46, максимум — 1500 байт).
9. Поле проверки четности содержит биты четности, генерируемые с помощью следующего полинома (см. раздел 6.7).

$$X^{32} + X^{26} + X^{23} + X^{22} + X^{16} + X^{12} + X^{11} + X^{10} + X^8 + X^7 + X^5 + X^4 + X^2 + X + 1$$

В алгоритме множественного доступа Ethernet определены следующие действия или отклики пользователя.

1. *Отложить.* Пользователь не должен передавать данные при наличии несущей или в течение минимального времени, разделяющего пакеты.
2. *Передать.* Если не используется предыдущее действие, пользователь может передавать данные до окончания времени передачи пакета или до возникновения конфликта.
3. *Прервать.* При возникновении конфликта пользователь должен прекратить передачу данных и оповестить других пользователей, участвующих в конфликте.
4. *Передать повторно.* Пользователь должен предпринять попытку повторной передачи после паузы случайной протяженности (аналогично схеме ALOHA).
5. *Откат.* Пауза перед n-й попыткой повторной передачи — это равномерно распределенное случайное число от 0 до $2^n - 1$, где $(0 < n \leq 10)$. При $n > 10$ интервал остается в пределах от 0 до 1023. Единицей измерения времени для интервала задержки перед повторной передачей является 512 бит (51,2 мкс).

На рис. 11.41, *б* показан поток данных со скоростью 10 Мбит/с при использовании манчестерской схемы PCM из спецификации Ethernet. Отметим, что при таком форматировании каждый однобитовый элемент или позиция двоичного разряда содержит переход. Двоичная единица описывается переходом с низкого уровня на высокий, двоичный нуль — переходом с высокого уровня на низкий. Следовательно, наличие переходов служит показателем наличия несущей. Если в течение определенного промежутка времени (от 0,75 до 1,25 периода передачи бита) переход не наблюдается — несущая потеряна, что свидетельствует об окончании пакета.

11.5.2. Сети Token Ring

Сеть с детектированием несущей состоит из кабеля, к которому пассивно подключаются все станции. *Кольцевая сеть* включает в себя несколько двухточечных кабелей, последовательно соединяющих станции. Сопряжение между кольцом и станциями является уже не пассивным, а активным. На рис. 11.42, *а* представлено стандартное однонаправленное кольцо с подключением через интерфейсы к нескольким станциям. На рис. 11.42, *б* показано состояние интерфейса для режима ожидания и режима передачи. В *режиме ожидания* входные биты копируются на выход с задержкой, равной времени прохождения одного бита. В *режиме передачи* соединение разрывается так, что станция может вводить в кольцо собственные данные. Маркер (token) — это специальная последовательность бит (например, 1 1 1 1 1 1 1 1), которая циркулирует по кольцу, когда все станции находятся в "холостом" состоянии. Как система может гарантировать, что последовательность бит, составляющая маркер, не встретится как часть передаваемых данных? Для этого используется метод *заполнения битами* (bit stuffing). Для приведенного примера 8-битового маркера, после каждой информационной последовательности из семи единиц система будет помещать нуль. При извлечении данных приемник использует подобный алгоритм для удаления введенного бита, перед которым идут семь единиц. Кольцевая сеть с маркерным доступом (сеть Token Ring) работает следующим образом.

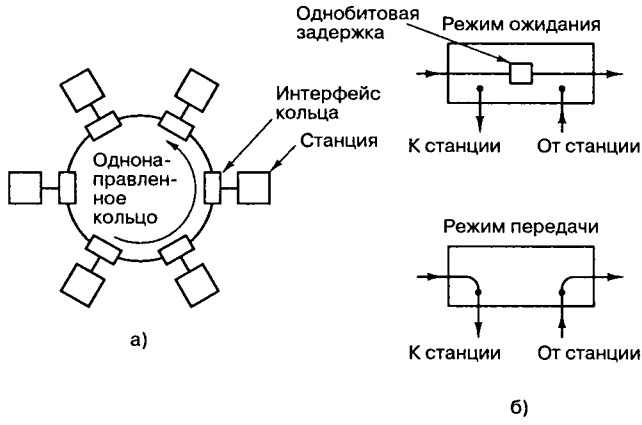


Рис. 11.42. Кольцевая сеть с маркерным доступом: а) сеть; б) режимы ожидания и передачи

1. Станция, желающая передавать, отслеживает появление маркера на интерфейсе. При прохождении маркера станция инвертирует последний бит (например, 1 1 1

1 1 1 1 0). Затем она прерывает интерфейсное соединение и вводит в кольцо собственные данные.

2. После прохождения по кольцу биты удаляются отправителем. Размер пакетов не ограничен, поскольку никакой пакет не появится в сети мгновенно.
3. После передачи последнего бита сообщения станция должна восстановить маркер. После прохождения по кольцу последний бит данных удаляется, а интерфейс переключается в режим ожидания.
4. В системе с маркерным доступом возникновение конфликтных ситуаций невозможно. При весьма активном обмене данными маркер сразу после восстановления захватывается следующей станцией кольца. Таким образом, разрешение на передачу данных последовательно передается по кольцу. Поскольку используется только один маркер, конфликтные ситуации не возникают.

Кольцевая система должна делать такую паузу, чтобы позволить передачу маркера по кольцу, когда все станции находятся в холостом состоянии. Важным моментом при проектировании кольцевых сетей является расстояние распространения или "длина" бита. Если скорость передачи данных равна R Мбит/с, бит выпускается за каждые ($1/R$) мкс. Поскольку скорость распространения по типичному коаксиальному кабелю равна 200 м/мкс, бит занимает $200/R$ метров кольца.

Пример 11.4. Минимальный размер кольца

Пусть скорость передачи данных в кольцевой сети с маркерным доступом равна 5 Мбит/с, а размер маркера — 8 бит. Определите минимальное *расстояние распространения* d_p, необходимое для охвата кольца. Скорость распространения v_p равна 200 м/мкс.

Решение

$$R = 5 \text{ Мбит/с}$$

Время, необходимое для передачи одного бита, t_b, равно следующему.

$$t_b = \frac{1}{5 \times 10^6} \text{с}$$

Время передачи восьмибитового маркера, t_t

$$t_t = \frac{8}{5 \times 10^6} \text{с}$$

Расстояние распространения восьмибитового маркера.

$$d_p = t_t \times v_p = \frac{8}{5} \text{мкс} \times 200 \text{ м/мкс} = 320 \text{ м}$$

11.5.3. Сравнение производительности сетей CSMA/CD и Token Ring

На рис. 11.43 сравнивается зависимость задержки от пропускной способности для сети CSMA/CD и кольцевой сети с маркерным доступом. В каждом случае используется кабель протяженностью 2 км, сеть включает 50 станций, средняя длина пакета равна 1 000 бит, размер заголовка сообщения равен 24 бит. На рис. 11.43, *а*, где скорость передачи данных равна 1 Мбит/с, графики практически совпадают. На рис. 11.43, *б*, по сравнению с предыдущим, был изменен один параметр — скорость передачи данных увеличена до 10 Мбит/с. Видим, что в данном случае разница между двумя системами является значительной. При нормированной пропускной способности $\rho < 0{,}22$,

CSMA/CD превосходит по производительности систему с маркерным доступом. Однако при $\rho > 0{,}22$ характеристики системы с маркерным доступом значительно лучше, чем системы CSMA/CD. Чтобы понять причину низкой производительности CSMA/CD (рис. 11.43, *б*), напомним определение ρ из уравнений (11.17) и (11.19).

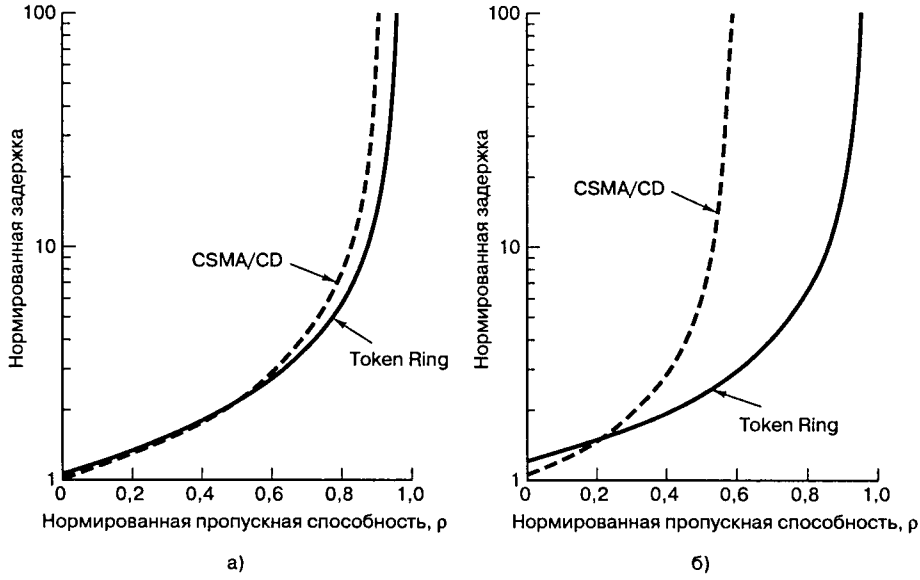


Рис. 11.43. Зависимость задержки от нормированной пропускной способности для сетей с маркерным доступом и CSMA/CD: а) скорость передачи данных 1 Мбит/с; б) скорость передачи данных 10 Мбит/с. (Перепечатано с разрешения автора из Bux W. "Local-Area Subnetworks: A Performance Comparison". IEEE Trans. Commun., vol. COM29, n. 10, October, 1981, pp. 1465–1473. © 1981, IEEE.)

$$\rho = \frac{b\lambda}{R} = \frac{\rho'}{R}$$

Здесь $\rho' = b\lambda$ — пропускная способность канала в бит/с, а R — емкость канала (максимальная скорость передачи битов). По мере роста R пропускная способность канала должна возрастать в соответствии с заданным значением ρ. При высокой пропускной способности большинство попыток передачи в системе CSMA/CD приводит к конфликтам [26].

11.6. Резюме

В этой главе рассмотрены концепции совместного использования ресурсов и подробно описаны классические подходы: схемы FDM/FDMA и TDM/TDMA. Приведено также описание смешанного метода множественного доступа — CDMA. Кроме того, дано введение в некоторые спутниковые методы множественного доступа, получившие широкое распространение в 70–80-х годах: многолучевое многократное использование частоты и двойное поляризационное многократное использование частоты.

В контексте нескольких модификаций алгоритма ALOHA рассмотрены методы множественного доступа с выделением ресурса по требованию (DAMA). Также приведено описание нескольких методов множественного доступа, используемых системами

INTELSAT, в частности FDM/FM, SPADE, TDMA и SS/TDMA. В заключение выполнено сравнение двух распространенных алгоритмов, используемых в локальных сетях, — множественного доступа к среде с обнаружением конфликтов и детектированием несущей (CSMA/CD) и маркерного доступа (Token Ring). Основная задача данной главы — общее представление информации о методах множественного доступа.

Литература

1. Rubin I. *Message Delays in FDMA and TDMA Communication Channels.* IEEE Trans. Commun., vol. COM27, n. 5, May, 1979, pp. 769–777.
2. Nirenberg L. M. and Rubin I. *Multiple Access Systems Engineering — A Tutorial.* IEEE WESCON/78 *Professional Program*, Modern Communications Techniques and Applications, session 21, Los Angeles, September, 13, 1978.
3. Abramson N. *The ALOHA System — Another Alternative for Computer Communications.* Proc. Fall Joint Comput. Conf. AFIPS, vol. 37, 1970, pp. 281–285.
4. Hayes J. F. *Local Distribution in Computer Communications.* IEEE Commun. Mag., March, 1981, pp. 6–14.
5. Schwartz M. *Computer — Communication Network Design and Analysis.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1977.
6. Tanenbaum A. S. *Computer Networks.* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1981.
7. Abramson N. *The ALOHA System;* in N. Abramson and F. F. Kuo, eds., *Computer Communication Networks*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1973.
8. Kleinrock L. *Queueing Systems*, vol. 1. Theory, John Wiley & Sons, Inc., New York, 1975.
9. Abramson N. *Packet Switching with Satellites.* AFIPS Conf. Proc., vol. 42, June, 1973, pp. 695–702.
10. Rosner R. D. *Packet Switching.* Lifelong Learning Publications, Wadsworth Publishing Company, Inc., Belmont, Calif., 1982.
11. Crowther W., Rettberg R., Walden D., Ornstein S. and Heart F. *A System for Broadcast Communication: Reservation ALOHA.* Proc. Sixth Hawaii Int. Conf. Syst. Sci., January, 1973, pp. 371–374.
12. Roberts L. *Dynamic Allocation of Satellite Capacity through Packet Reservation.* AFIPS Conf. Proc., vol. 42, June, 1973, p. 711.
13. Binder R. *A Dynamic Packet-Switching System for Satellite Broadcast Channels.* Proc. Int. Conf. Commun., June, 1975, pp. 41-1–41-5.
14. Capetanakis J. *Tree Algorithms for Packet Broadcast Channels.* IEEE Trans. Inf. Theory, vol. IT25, September, 1979, pp. 505–515.
15. Puente J. G. and Werth A. M. *Demand-Assigned Service for the INTELSAT Global Network.* IEEE Spectrum, January, 1971, pp. 59–69.
16. Jones J. J. *Hard Limiting of Two Signals in Random Noise.* IEEE Trans. Inf. Theory, vol. IT9, January, 1963, pp. 34–42.
17. Bond F. E. and Meyer H. F. *Intermodulation Effects in Limited Amplifier Repeaters.* IEEE Trans. Commun. Technol., vol. COM18, n. 2, April, 1970, pp. 127–135.
18. Shimbo O. *Effects of Intermodulation, AM-PM Conversion, and Additive Noise in Multicarrier TWT Systems.* Proc. IEEE, vol. 59, February, 1971, pp. 230–238.
19. Chakraborty D. *INTELSAT IV Satellite System (Voice) Channel Capacity versus Earth-Station Performance.* IEEE Trans. Commun. Technol., vol. COM19, n. 3, June, 1971, 355–362.
20. Campanella S. and Schaefer D. *Time Division Multiple Access Systems (TDMA);* in K. Feher, *Digital Communications, Satellite/Earth Station Engineering*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1983.
21. Scarcella T. and Abbott R. V. *Orbital Efficiency Through Satellite Digital Switching.* IEEE Commun. Mag., May, 1983, pp. 38–46.
22. Muratani T. *Satellite-Switched Time-Domain Multiple Access.* Proc. IEEE Electron. and Aerosp. Conf. (EASCON), 1974, pp. 189–196.
23. Dill G. D. *TDMA, The State-of the-Art.* Rec. IEEE Electron. Aerosp. Syst. Conv. (EASCON), September, 26–28, 1977, pp. 31-5A–31-5I.

24. Jarett K. *Operational Aspects of Intelsat VI Satellite - Switched TDMA Communication System.* AIAA Tenth Commun. Satell. Syst. Conf. March, 1984, pp. 107–111.
25. Stallings W. *Local Network Performance.* IEEE Commun. Mag., vol. 22, n. 2, February, 1984, pp. 27–36.
26. Bux W. *Local-Area Subnetworks: A Performance Comparison.* IEEE Trans. Commun., vol. COM29, n. 10, October, 1981, pp. 1465–1473.
27. Dixon R. C., Strole N. C. and Markov J. D. *A Token-Ring Network for Local Data Communications.* IBM Syst. J., vol. 22, n. 1-2, 1983, pp. 47–62.

Задачи

11.1. Разработайте набор сигналов FDM, состоящий из 5 каналов передачи речи, каждый в диапазоне 300–3400 Гц. Уплотненный набор сигналов должен состоять из инвертированных боковых полос и занимать спектральную область от 30 до 50 кГц.

а) Изобразите составной спектр, указав отдельные спектры и положение защитных полос.

б) Изобразите блок-схему, показывающую процессы смешивания частот и фильтрования, а также необходимые параметры местного гетеродина приемника.

11.2. Приемник настроен на прием нижней боковой полосы (lower sideband — LSB) радиочастотной несущей с частотой $f_c = 8$ МГц. Ширина полосы сигнала LSB равна 100 кГц. Для переноса принятого сигнала на нижнюю промежуточную частоту используется местный гетеродин приемника с частотой f_{LO}. Пусть $f_{LO} > f_c$, а усилитель промежуточной частоты центрирован на частоте 2 МГц. Изобразите блок-схему гетеродинного преобразования, на которой будут указаны радиочастотный фильтр, местный гетеродин и фильтр промежуточной частоты. Укажите частоту центрирования каждого фильтра и типичные спектры сигналов в разных точках диаграммы.

11.3. Из уравнений (11.13) и (11.15) следует, что средняя величина задержки сообщения в схеме TDMA меньше, чем в схеме FDMA. Какими будут практические результаты уменьшения времени задержки в схеме TDMA (как функции времени передачи кадра) для спутникового канала с односторонним радиусом действия 36 000 км? Для каких значений времени передачи кадра схема TDMA будет иметь значительное преимущество перед FDMA?

11.4. Группа станций совместно использует канал с чистой схемой ALOHA, поддерживающий скорость 56 Кбит/с. В среднем каждые 10 с любая станция передает пакет данных, даже если на данный момент предыдущий пакет еще не отправлен (т.е. станция заносит пакеты в буфер). Размер каждого пакета равен 3 000 бит. Найдите максимальное число станций, которые могут одновременно использовать данный канал. Процесс прибытия пакетов считать пуассоновским.

11.5. Группа из трех станций совместно использует канал с чистой схемой ALOHA, поддерживающий скорость 56 Кбит/с. Средняя скорость передачи данных станциями равна следующему: $R_1 = 7{,}5$ Кбит/с, $R_2 = 10$ Кбит/с, $R_3 = 20$ Кбит/с. Размер каждого пакета составляет 100 бит. Вычислите нормированный объем информации, которой обмениваются через канал, нормированную пропускную способность, вероятность успешной передачи и скорость поступления успешно переданных пакетов. Процесс поступления пакетов считать пуассоновским.

11.6. Докажите, что при использовании чистой схемы ALOHA нормированная пропускная способность не превышает $1/2e$, а максимум наблюдается при нормированном объеме переданной информации, равном 0,5.

11.7. а) Докажите, что уравнение (11.24) является действительной функцией плотности вероятности дискретной случайной переменной.

б) Найдите среднее значение дискретной случайной переменной, функция плотности вероятности которой описывается уравнением (11.24).

в) Докажите, что результат, полученный в п. б, не противоречит утверждению, что λ — средняя скорость поступления пакетов.

11.8. Рассмотрим процесс получения данных в чистом алгоритме ALOHA, показанный на рис. 311.1. Вертикальная стрелка указывает момент поступления пакета. N_n — число пакетов, полученных в промежутке времени $(T_{n-1}, T_n]$, где $(t_x, t_y]$ обозначает интервал $t_x < t \le t_y$. N_{n+1} — число пакетов, полученных в промежутке $(T_n, T_{n+1}]$; τ — продолжительность пакета в секундах. Средняя скорость поступления пакетов равна λ_t. Предполагать, что пакеты поступают независимо друг от друга.

а) Найдите функцию совместной плотности вероятности N_n и N_{n+1}.

б) Пусть T_n — время получения пакета пользователя A; выразите через совместную вероятность N_n и N_{n+1} вероятность того, что передача пользователя A будет успешной.

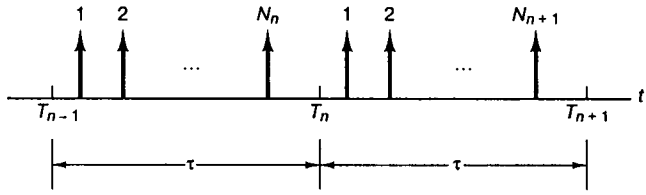


Рис. 311.1

11.9. Пусть $N = N_n + N_{n+1}$, где N_n и N_{n+1} определены в задаче 11.8. Найдите функцию плотности вероятности для N и объясните значение N.

11.10. 6 000 станций состязаются за доступ к каналу системы S-ALOHA. Средняя станция делает 30 запросов в час, причем каждый раз запрашивается интервал 500 мкс. Рассчитайте нормированный объем информации, переданной по каналу.

11.11. Рассмотрим сценарий, изображенный на рис. 311.1; указанные времена поступления пакетов допустимы для чистого алгоритма ALOHA, но не для алгоритма S-ALOHA, где поступление пакетов возможно только в заданные моменты времени T_i, $i = 0, 1, \ldots$. Пусть среднее время поступления пакетов равно λ_t.

а) Как изменится рис. 311.1 для схемы S-ALOHA? Как при этом изменятся функции плотности вероятности N_n и N_{n+1}?

б) Какова вероятность успешной передачи данных, если пакет пользователя A поступил в момент времени T_n?

11.12. Группа станций, использующих алгоритм S-ALOHA, генерирует в общем 120 запросов в течение секунды, включая исходные и повторные передачи. Каждый раз запрашивается интервал 12,5 мс.

а) Рассчитайте нормированный объем информации, переданной по каналу.

б) Определите вероятность успешной передачи данных при первой попытке.

в) Какова вероятность возникновения ровно двух конфликтов непосредственно перед успешной передачей?

11.13. Статистика использования канала S-ALOHA показывает, что 20% интервалов не используется.

а) Определите нормированный объем информации, переданной по каналу.

б) Определите нормированную пропускную способность канала.

в) Является ли канал перегруженным или его мощность используется не полностью?

11.14. Покажите, что сумма двух пуассоновских процессов со скоростями λ_1 и λ_2 также является пуассоновским процессом со скоростью $\lambda_t = \lambda_1 + \lambda_2$. Обобщите результат на сумму n пуассоновских процессов.

11.15. Транспондер с шириной полосы 10 МГц использует 200 идентичных несущих, половина которых обслуживает станции с $G/T = 40$ дБ/К, остальные — станции с $G/T = 37$ дБ/К.

Вероятность возникновения битовой ошибки для каждой станции не должна превышать 10^{-5}. Транспондер ограничен по мощности.

а) Определите максимальную ширину полосы для каждой несущей.

б) Пусть ширина полосы каждой несущей равна 40 кГц, а транспондер должен обслуживать только группу более мощных ($G/T = 40$ дБ/К) станций. Сколько станций сможет обслуживать транспондер? Будет ли транспондер ограничен по мощности или по ширине полосы?

в) Рассмотрите п. б при условии, что транспондер должен обслуживать только малые станции ($G/T = 37$ дБ/К).

11.16. Система TDMA работает со скоростью 100 Мбит/с, длительность кадра равна 2 мс. Пусть все временные интервалы равны (по длительности), а защитная полоса между ними 1 мкс.

а) Рассчитайте эффективность использования ресурса связи для кадров, содержащих 1, 2, 5, 10, 20, 50 и 100 интервалов.

б) Решите п. а, считая, что в начале каждого интервала требуется начальная комбинация из 100 бит. Рассчитайте эффективность использования ресурса связи в зависимости от объема переданной информации.

в) Изобразите графически результаты пп. а и б.

11.17. С помощью уравнения (11.36) выполните следующее.

а) Проанализируйте эффективность использования ресурса связи, если все S_i и R_j равны.

б) Проанализируйте, что произойдет, если отдельные S_i и R_j будут значительно больше остальных. Как можно улучшить эффективность использования ресурса связи?

в) Укажите, в каком случае распределения S_i и R_j будут подобны между собой. В каком случае они будут различны?

11.18. а) Кольцевая сеть с маркерным доступом работает со скоростью передачи данных 10 Мбит/с по кабелю со скоростью распространения 200 м/мкс. Какая протяженность кабеля приведет к задержке в 1 бит на каждом интерфейсе кольца?

б) Пусть длина маркера равна 10 бит, а все станции сети, кроме трех, не работают в вечернее время. Какова минимальная длина кабеля необходима для создания кольца?

Вопросы для самопроверки

11.1. Что обычно подразумевается под *ресурсом связи* (см. вступление)?

11.2. В чем сходство и различие *уплотнения* и *множественного доступа* (см. вступление)?

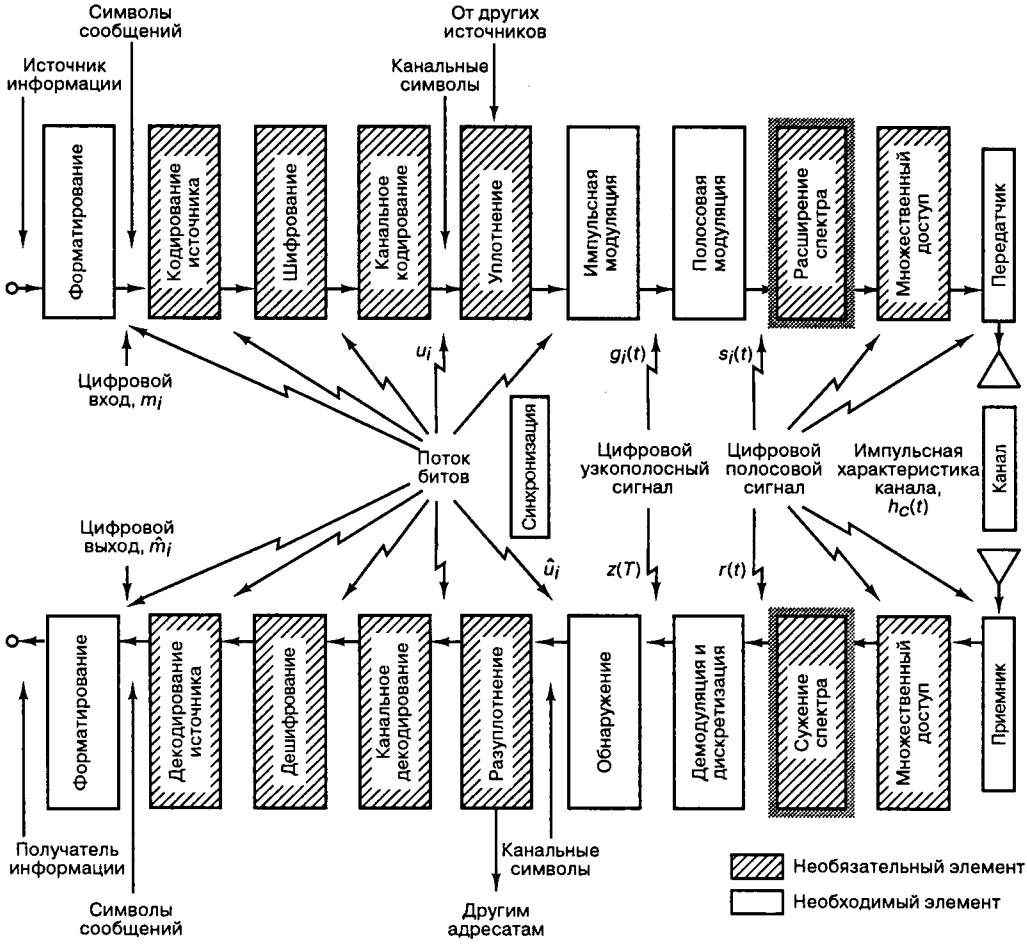
11.3. Почему *линейное устройство* невозможно использовать в качестве смесителя частот (см. раздел 11.1.1.1 и приложение А)?

11.4. Существует ли теоретическое преимущество по пропускной способности при предоставлении услуг FDMA и TDMA (см. раздел 11.1.4.1)?

11.5. Укажите преимущества схемы CDMA перед схемами FDMA и TDMA (см. раздел 11.1.5).

ГЛАВА 12

Методы расширенного спектра



12.1. Расширенный спектр

Изначально методы расширенного спектра (spread-spectrum — SS) применялись при разработке военных систем управления и связи. К концу второй мировой войны в радиолокации расширение спектра применялось для борьбы с преднамеренными помехами [1], а в последующие годы развитие данной технологии объяснялось желанием создать помехоустойчивые системы связи. В процессе исследований расширенному спектру нашлось и другое применение — снижение плотности энергии, высокоточная локация и использование при множественном доступе. Все эти практические приложения расширенного спектра будут рассмотрены в данной главе. Методы *расширенного спектра* получили свое название благодаря тому, что полоса, используемая для передачи сигнала, намного шире минимальной, необходимой для передачи данных. Система связи называется системой с расширенным спектром в следующих случаях.

1. Используемая полоса значительно шире минимальной, необходимой для передачи данных.
2. Расширение спектра производится с помощью так называемого *расширяющего (или кодового) сигнала*, который не зависит от передаваемой информации. Подробное описание таких сигналов приводится в последующих разделах главы.
3. Восстановление исходных данных приемником ("сужение спектра") осуществляется путем сопоставления полученного сигнала и синхронизированной копии расширяющего сигнала.

Следует отметить, что расширение спектра сигнала также происходит при использовании некоторых стандартных схем модуляции, таких как частотная и импульсно-кодовая модуляция. Однако эти схемы не относятся к методам расширенного спектра, поскольку не удовлетворяют всем приведенным выше условиям.

12.1.1. Преимущества систем связи расширенного спектра

12.1.1.1. Подавление помех

По определению белый гауссов шум — это математическая модель шума бесконечно большой мощности, равномерно распределенного по всему спектру частот. Наличие такого шума не обязательно означает отсутствие эффективной связи, поскольку интерферировать с сигналом могут лишь шумовые составляющие ограниченной мощности, находящиеся в сигнальном пространстве (другими словами, имеющие *те же координаты*, что и компоненты сигнала). Прочие составляющие эффективно отсеиваются детектором (см. раздел 3.1.3). Для типичного узкополосного сигнала это означает, что характеристики связи ухудшают только шумы, находящиеся в диапазоне сигнала. Поскольку изначально методы расширенного спектра разрабатывались для военных систем связи, работающих при повышенном уровне помех, создаваемых противником, вначале будет рассмотрена помехоустойчивость данных методов (коммерческое использование данных систем рассматривается в разделах 12.7 и 12.8).

Рассмотрим основополагающий принцип применения расширенного спектра для создания помехоустойчивых систем связи. Предположим, что для передачи сигнала можно использовать множества ортогональных координат (или измерений), причем в каждый момент времени используется только малая их часть. Допустим также, что станция-постановщик помех не способна определить подмножество координат, ис-

пользуемое в данный момент. Количество координат для сигнала с шириной полосы W и длительностью T будет приблизительно равно $2WT$ [2]. При определенном построении системы вероятность ошибки в ней будет функцией только E_b/N_0. При наличии белого гауссова шума *бесконечно большой* мощности использование расширения (т.е. больших значений $2WT$) не улучшает качества связи. В то же время, если шум происходит от постановщика помех с *постоянной конечной мощностью* и нельзя точно установить координаты сигнала в пространстве сигналов, то для подавления сигнала можно использовать только следующие методы.

1. Создание помех *равной* мощности во *всем* сигнальном пространстве. В таком случае мощность помех на каждой координате будет небольшой.
2. Создание помех *большей* мощности для *небольшого* количества координат диапазона (более общий случай — создание помех различной мощности для всех координат диапазона).

На рис. 12.1 приводится сравнение систем с расширенным спектром при наличии белого шума и при постановке преднамеренных помех. Спектральная плотность мощности сигнала обозначается $G(f)$ до расширения и $G_{ss}(f)$ после расширения. Для простоты на рисунке рассматривается только частотный диапазон. Как показано на рис. 12.1, *а*, односторонняя спектральная плотность мощности белого шума N_0 не изменяется при расширении полосы сигнала с W до W_{ss}. Средняя мощность белого шума (площадь под кривой спектральной плоскости) является бесконечной. Следовательно, расширение не улучшает качества связи. На рис. 12.1, *б* (верхняя диаграмма) представлено создание намеренных помех ограниченной мощности J. Спектральная плотность мощности в данном примере равна $J'_0 = J/W$, где W — ширина нерасширенной полосы, подвергающейся воздействию помех. После расширения диапазона сигнала станция намеренных помех может использовать один из двух изложенных выше методов. Для метода 1 это означает рассеивание спектральной плотности шумов J'_0 по всему диапазону сигнала (на единицу ширины полосы теперь приходится в (W/W_{ss}) раз меньшая мощность помех). Получаемую спектральную плотность шумов $J_0 = J/W_{ss}$ называют *спектральной плотностью шума широкополосного постановщика помех*. При использовании метода 2 уменьшается количество точек диапазона, в которых создаются помехи. В то же время постановщик помех может увеличить спектральную плотность шумов с J_0 до J_0/ρ $(0 < \rho \leq 1)$, где ρ — часть полосы расширенного спектра, в которой создаются помехи. При неудачном выборе координат постановки помех средняя их эффективность будет ниже, чем при удачном. Чем больше набор координат для передачи сигнала, тем сложнее задача по его подавлению, и соответственно, связь будет более защищенной от преднамеренных помех. Сравнение систем связи с расширенным спектром и нерасширенным должно производиться в предположении о равной полной средней мощности обеих систем. Поскольку площадь под кривыми спектральной плотности мощности (power spectral density — PSD) представляет собой полную среднюю мощность, площадь под кривыми PSD для расширенного и нерасширенного спектров должна быть неизменной. Таким образом, должно быть очевидно, что графики $G_{ss}(f)$ на рис. 12.1, *а* и *б* имеют разный масштаб.

Возникновение помех не всегда является результатом преднамеренных действий. В некоторых случаях помехи могут быть следствием природных явлений. Кроме того, так называемый *многолучевой эффект* способен вызвать самоинтерференцию, т.е. основной сигнал и его отражения, имеющие различные направления распространения, интерферируют между собой.

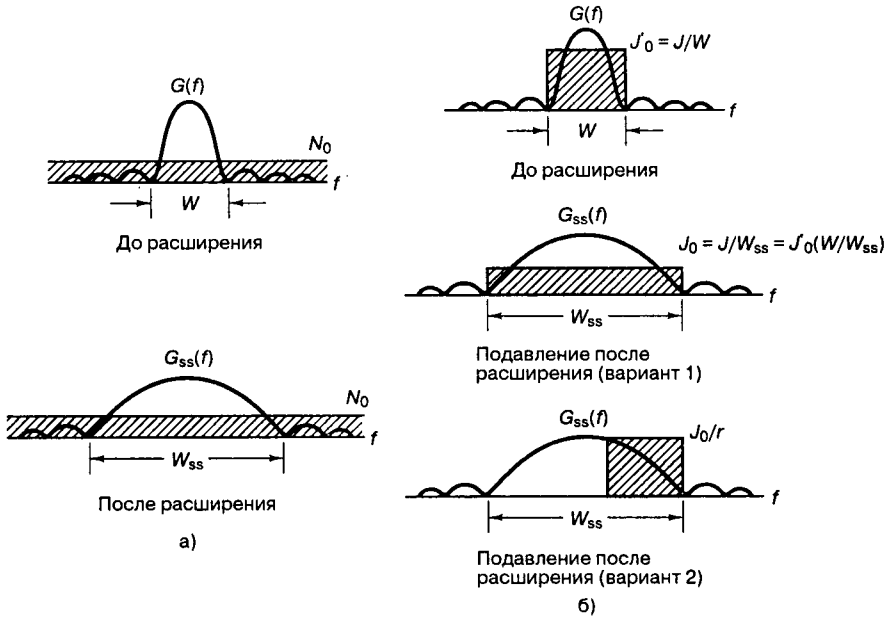


Рис. 12.1. Расширение спектра: а) при наличии белого шума; б) при постановке намеренных помех

12.1.1.2. Снижение плотности энергии

Представим себе ситуацию, когда сигнал в процессе связи не должен быть получен никем, кроме определенного приемника. Устройства, используемые в таких случаях, называют системами связи с *низкой вероятностью обнаружения* (LPD — low probability of detection) или же *системами с низкой вероятностью перехвата* (LPI — low probability of intercept). Основной особенностью этих систем является минимальная вероятность обнаружения сеанса связи кем-либо, кроме определенного приемника, при использовании минимальной мощности сигнала и оптимальной схемы передачи. Следовательно, в контексте систем связи расширенного спектра распределение по множеству координат приводит к тому, что сигнал более равномерно и менее плотно (по сравнению с традиционными схемами модуляции) распределяется в заданной области спектра. Таким образом, не только повышается помехоустойчивость сигнала, но и снижается вероятность его перехвата. Для того, кто не располагает синхронизированной копией расширенного сигнала, данный сигнал будет теряться в шуме.

Для обнаружения расширенного сигнала в заданном диапазоне W может быть использован *радиометр*. Как видно из рис. 12.2, радиометр состоит из полосового фильтра (bandpass filter — BPF) с полосой W, схемы возведения в квадрат, которая обеспечивает положительную выходную мощность (поскольку обнаруживается *энергия* сигнала), а также интегрирующей схемы. В момент времени $t = T$ выход интегратора сравнивается с порогом. Если выход больше порога, считается, что сигнал присутствует, в противном случае считается, что сигнала нет. Подробное описание возможности обнаружения сигналов расширенного спектра с помощью радиометра и более сложных устройств, использующих особенности сигналов, приводится в работах [3, 4].

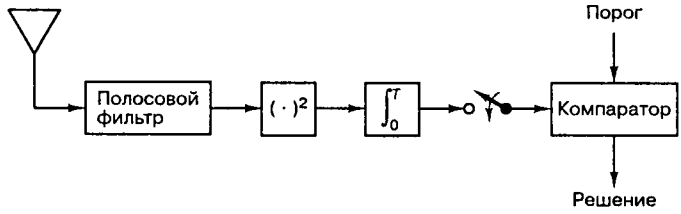


Рис. 12.2. Радиометр

При создании систем LPI может проявляться эффект *снижения вероятности определения местоположения* (LPPF — low probability of position fix), т.е. даже при обнаружении наличия сигнала затруднительно определить местоположение передатчика. В некоторых системах связи расширенного спектра применяется метод *снижения вероятности использования сигнала* (LPSE — low probability of signal exploitation), что усложняет идентификацию передатчика.

Метод расширенного спектра может применяться для уменьшения плотности энергии сигнала, что иногда требуется для согласования систем связи с государственными стандартами. Сигналы, передаваемые спутниками, должны соответствовать международным стандартам относительно спектральной плотности вблизи поверхности Земли. Путем распределения энергии сигнала спутника по расширенному диапазону можно увеличить полную энергию переданного сигнала, что позволяет улучшить производительность системы, а также удовлетворить требования стандартов относительно плотности энергии.

12.1.1.3. Высокая временная разрешающая способность

Сигналы расширенного спектра могут использоваться для определения местоположения. Расстояние можно определить с помощью измерения задержки распространения импульсного сигнала. Как следует из рис. 12.3, погрешность такого измерения, Δt, прямо пропорциональна времени нарастания сигнала, которое, в свою очередь, обратно пропорционально ширине полосы сигнала.

$$\Delta t \approx \frac{1}{W} \tag{12.1}$$

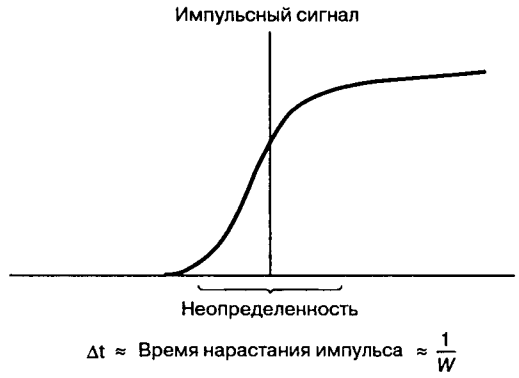


Рис. 12.3. Измерение времени задержки распространения

Точность измерения расстояния может быть повышена за счет увеличения ширины полосы сигнала. При использовании гауссова канала результат, полученный путем одноразового измерения единичного импульсного сигнала, не будет достаточно точным. Метод расширенного спектра предполагает применение кодированного сигнала, состоящего из длинной последовательности изменений полярности (например, сигнал с модуляцией PSK). В приемнике полученная последовательность сопоставляется с локальной копией, и результаты такого сопоставления позволяют произвести точное измерение расстояния.

12.1.1.4. Множественный доступ

Методы расширенного спектра применяются в системах связи множественного доступа для управления совместным использованием ресурса связи большим числом пользователей. Данный метод называется *множественным доступом с кодовым разделением* (code-division multiple access — CDMA); его краткое описание приведено в главе 11. Одной из особенностей CDMA является сохранение конфиденциальности связи между пользователями, имеющими разные сигналы расширенного спектра. Отслеживание сеанса связи будет непростой задачей для пользователя, не имеющего доступа к определенному сигналу. Более подробно данный вопрос будет рассмотрен позже.

12.1.2. Методы расширения спектра

На рис. 12.4 отмечены распространенные методы расширения информационного сигнала на большее число координат диапазона. Для сигнала с длительностью T и шириной полосы W размерность пространства сигналов приблизительно равна $2WT$. Размерность диапазона можно повысить за счет увеличения W (расширение спектра) или T (расширение временного диапазона или переключение временных интервалов). При расширении спектра сигнал расширяется в частотной области. При переключении временных интервалов сообщению, передаваемому со скоростью R, выделяется более длительное время, чем необходимо для передачи данных с помощью обычного метода модуляции. В течение этого времени данные передаются отдельными пакетами согласно требованиям кода. Можно сказать, что при переключении временных интервалов сигнал расширяется во временной области. В обоих случаях создание преднамеренных помех будет осложнено тем, что область, используемая сигналом в каждый момент времени, будет неопределенной.

Первые два метода, указанные в разделе "расширение спектра" на рис. 12.4, — *метод прямой последовательности* (direct sequencing — DS) и *метод скачкообразной перестройки частоты* (frequency hopping — FH) — являются наиболее распространенными. Третий метод, *переключение временных интервалов* (time hopping — TH), используется при наличии преднамеренных помех, поскольку он позволяет скрывать координаты сигнала от потенциального противника. Кроме того, существуют смешанные методы, такие как DS/FH, FH/TH или DS/FH/TH. Поскольку эти методы — просто развитие основных, детально они рассматриваться не будут. В данной главе основное внимание обращается на два основных метода расширения спектра: прямой последовательности и скачкообразной перестройки частоты.

Рис. 12.4. Основные преобразования цифровой связи

12.1.3. Моделирование подавления интерференции с помощью расширения спектра методом прямой последовательности

На рис. 12.5 представлена модель подавления интерференции с использованием расширения спектра методом прямой последовательности (direct-sequence spread-spectrum — DS/SS). Сигнал $x(t)$, характеризующийся скоростью передачи данных R бит/с, модулируется путем умножения на расширяющий кодовый сигнал $g(t)$, скорость передачи которого равна R_{ch} элементарных сигналов/с. Предположим, что полосы передачи для $x(t)$ и $g(t)$ равны R и R_{ch} Гц. Умножение данных двух функций во временной области соответствует их свертке в частотной области.

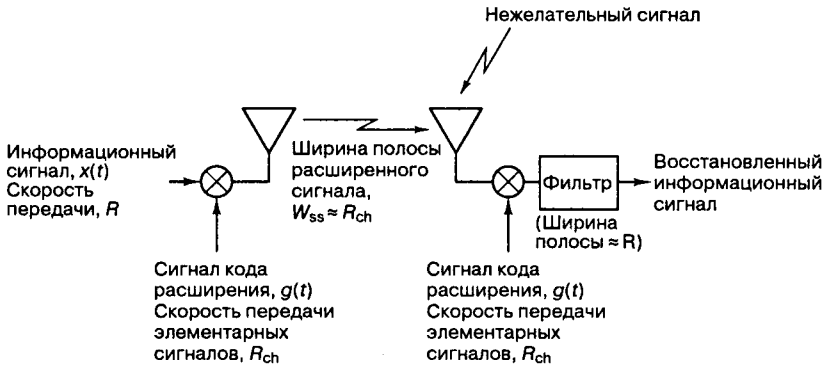


Рис. 12.5. Основа метода расширенного спектра

$$x(t)g(t) \leftrightarrow X(\omega) * G(\omega) \tag{12.2}$$

Следовательно, если информационный сигнал является узкополосным (по сравнению с расширяющим сигналом), произведение $x(t)g(t)$ будет приблизительно равно ширине полосы расширяющего сигнала (см. раздел А.5).

В демодуляторе полученный сигнал умножается на синхронизированную копию расширяющего сигнала $g(t)$, в результате чего получается суженный сигнал. Для отсеивания побочных высокочастотных компонентов используется фильтр с шириной полосы R. Следует отметить, что любой нежелательный сигнал, полученный приемником, будет расширен путем умножения на $g(t)$, точно так же как передатчик расширяет исходный сигнал. Рассмотрим, как это скажется на станции-постановщике помех, которая пытается создать узкополосную помеху в диапазоне передачи информации. Первая операция на входе приемника — умножение на расширяющий сигнал расширения. Таким образом, помехи будут расширены по всему диапазону этого сигнала.

Наиболее важные особенности помехоустойчивой системы связи расширенного спектра можно сформулировать следующим образом.

1. *Однократное* умножение на $g(t)$ приводит к расширению диапазона сигнала.
2. *Повторное* умножение и последующее фильтрование восстанавливают исходный сигнал.
3. Исходный сигнал умножается *дважды*, тогда как сигнал-помеха умножается только *один раз*.

12.1.4. Историческая справка

12.1.4.1. Передача или хранение опорного сигнала

В течение первых нескольких лет исследования систем расширенного спектра синхронизация работы приемника и передатчика производилась с помощью *истинно случайного* расширяющего сигнала (например, широкополосного шума). Такие устройства получили название *систем связи с передачей опорного сигнала* (transmitted reference — TR). В системах TR передатчик отправляет две версии непредсказуемых широкополосных несущих, одна из которых модулируется данными, а другая остается немодулированной. Указанные два сигнала передаются по разным каналам. Приемник использует немодулированную несущую для сужения несущей, модулированной данными. Основное преимущество систем TR — отсутствие серьезных проблем синхронизации в приемнике, поскольку оба сигнала передаются одновременно. Существенные недостатки TR заключаются в следующем: (1) расширяющий код отправляется незашифрованным, потому доступен для прослушивания; (2) в систему легко внедрить чужеродную информацию, если послать пару сигналов, приемлемых с точки зрения приемника; (3) наличие шумов в обоих сигналах приводит к росту вероятности ошибки при низкой мощности сигнала; (4) для передачи опорного сигнала требуется удвоить ширину полосы и мощность сигнала.

Все современные системы расширенного спектра построены с использованием метода *хранения опорного сигнала* (stored reference — SR). В этом случае опорный сигнал независимо генерируется приемником и передатчиком. Основным преимуществом систем SR является то, что при правильном выборе кода сигнал не может быть определен путем прослушивания. Нужно отметить, что кодовый сигнал системы SR, сходный по характеристикам с белым шумом, не может быть истинно случайным, как в случае системы TR. Поскольку один и тот же код должен быть независимо сгенерирован двумя или более пользователями, последовательность кода должна быть детерминированной (хотя для "неуполномоченных слушателей" она может казаться случайной). Такая последовательность детерминированных сигналов называется псевдошумовой (pseudonoise — PN), или же псевдослучайной (pseudorandom) последовательностью. Более подробно генерирование псевдослучайных последовательностей будет рассмотрено позже.

12.1.4.2. Шумовые колеса

В конце 40-х–начале 50-х годов Мортимер Рогофф (Mortimer Rogoff), сотрудник ITT (International Telephone and Telegraph Corporation — Международная телефонная и телеграфная корпорация, США), провел новаторский эксперимент с использованием систем расширенного спектра [5]. Используя фотографию, Рогофф построил "шумовое колесо", содержащее информацию о псевдослучайном сигнале. Из телефонного справочника Манхеттена были выбраны 1440 номеров, не заканчивающихся на "00". Две средние из четырех последних цифр каждого номера были радиально расположены с интервалом 1/4°, после чего график был перенесен на пленку в виде колеса (рис. 12.6). При вращении колеса свет, излучаемый из прорези, образует модулированный интенсивностью луча пучок света, фактически представляющий собой псевдослучайный расширяющий сигнал, который может быть зафиксирован фотоэлементом.

Рогофф установил два идентичных шумовых колеса на ось, вращаемую синхронным двигателем с частотой 900 об/мин. Расширяющий сигнал одного из колес модулировался данными (и помехами), после чего поступал на один из входов принимающего коррелятора. На другой вход коррелятора поступал немодулированный сигнал второго колеса. Экс-

перименты проводились с узкополосными сигналами на скорости 1 бит/с. В результате была доказана возможность передачи информации в виде сигналов, подобных шуму [4].

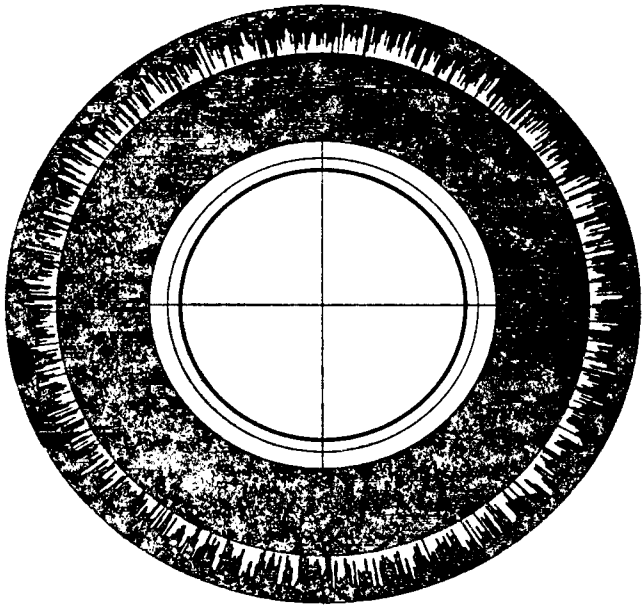


Рис. 12.6. Шумовое колесо Рогоффа. (Перепечатано с разрешения ITT из Section I (Communications) of "Application of Statistical Methods to Secrecy Communication Systems," Proposal 946, Fed. Telecomm. Lab., August, 28, 1950, Fig. 6.)

12.2. Псевдослучайные последовательности

Системы связи расширенного спектра с передачей опорного сигнала (transmitted reference — TR) могут использовать *истинно* случайный кодовый сигнал для расширения и сужения, поскольку кодовый сигнал и модулированный данными кодовый сигнал одновременно передаются в разных областях спектра. Метод хранения опорного сигнала (stored reference — SR) *не позволяет* использовать истинно случайные кодовые сигналы, поскольку код должен храниться или генерироваться приемником. В системах SR должен применяться *псевдошумовой* (pseudonoise) или *псевдослучайный* (pseudorandom) кодовый сигнал.

В чем отличие псевдослучайного кода от истинно случайного? Случайная последовательность *непредсказуема* и может быть описана только в статистическом смысле. Псевдослучайный код на самом деле не является случайным — это детерминированный периодический сигнал, известный передатчику и приемнику. Так почему же он называется "псевдослучайным"? Причина в том, что он имеет все статистические свойства дискретного белого шума. Для "неуполномоченного" пользователя такой сигнал будет казаться абсолютно случайным.

12.2.1. Свойства случайной последовательности

Каким должен быть псевдослучайный код, чтобы казаться истинно случайным? Существует три основных свойства любой периодической двоичной последовательности, которые могут быть использованы в качестве проверки на случайность.

1. *Сбалансированность.* Для каждого интервала последовательности количество двоичных единиц должно отличаться от числа двоичных нулей не больше чем на на один элемент.
2. *Цикличность. Циклом* называют непрерывную последовательность одинаковых двоичных чисел. Появление иной двоичной цифры автоматически начинает новый цикл. Длина цикла равна количеству цифр в нем. Желательно, чтобы в каждом фрагменте последовательности приблизительно половину составляли циклы обоих типов длиной 1, приблизительно одну четверть — длиной 2, приблизительно одну восьмую — длиной 3 и т. д.
3. *Корреляция.* Если часть последовательности и ее циклично сдвинутая копия поэлементно сравниваются, желательно, чтобы число совпадений отличалось от числа несовпадений не более чем на единицу.

В следующем разделе для проверки данных свойств будет сгенерирована псевдослучайная последовательность.

12.2.2. Последовательности, генерируемые регистром сдвига

Рассмотрим линейный регистр сдвига с обратной связью (рис. 12.7), который состоит из четырехразрядного регистра для хранения и сдвига, сумматора по модулю 2 (операция суммирования по модулю 2 была определена в разделе 2.9.3), а также контура обратной связи с входом регистра. Работа регистра сдвига управляется последовательностью синхронизирующих импульсов (не показанных на рисунке). С каждым импульсом содержимое регистров сдвигается на одну позицию вправо, а содержимое регистров X_3 и X_4 суммируется по модулю 2 (линейная операция). Результат суммирования по обратной связи подается на разряд X_1. Последовательность, генерируемая регистром сдвига, — это, по определению, выход последнего регистра (в данном случае X_4).

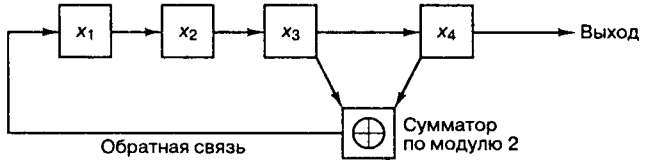


Рис. 12.7. Пример линейного регистра сдвига с обратной связью

Предположим, что разряд X_1 содержит единицу, а все остальные разряды — нули, т.е. начальным состоянием регистра является 1 0 0 0. В соответствии с рис. 12.7, последующие состояния регистра будут следующими.

```
1000 0100 0010 1001 1100 0110 1011 0101
1010 1101 1110 1111 0111 0011 0001 1000
```

Поскольку последнее состояние, 1 0 0 0, идентично начальному, видим, что приведенная последовательность повторяется регистром через каждые 15 тактов. Выходная последовательность определяется содержимым разряда X_4 на каждом такте. Эта последовательность имеет следующий вид.

```
000100110101111
```

Здесь крайний левый бит является самым ранним. Проверим полученную последовательность на предмет соответствия критериям, приведенным в предыдущем разделе. По-

следовательность содержит семь нулей и восемь единиц, что соответствует условию сбалансированности. Рассмотрим циклы нулей — всего их четыре, причем половина их имеет длину 1, а одна четвертая — длину 2. То же получаем для циклов единиц. Последовательность слишком коротка, чтобы продолжать проверку, но видно, что условие цикличности выполняется. Условие корреляции будет проверено в разделе 12.2.3.

Последовательность, сгенерированная регистром сдвига, зависит от количества разрядов, места подсоединения отводов обратной связи и начальных условий. Последовательности на выходе генератора могут классифицироваться как имеющие *максимальную* или *немаксимальную* длину. Период повторения (в тактах) последовательности максимальной длины, генерируемой n-каскадным линейным регистром сдвига с обратной связью, равен следующему.

$$p = 2^n - 1 \tag{12.3}$$

Очевидно, что последовательность, сгенерированная регистром сдвига на рис. 12.7, является примером последовательности с максимальной длиной. Если длина последовательности меньше $(2^n - 1)$, говорят, что последовательность имеет немаксимальную длину.

12.2.3. Автокорреляционная функция псевдослучайного сигнала

Автокорреляционная функция $R_x(\tau)$ периодического сигнала $x(t)$ с периодом T_0 была представлена в уравнении (1.23) и приводится ниже в нормированной форме.

$$R_x(\tau) = \frac{1}{K}\left(\frac{1}{T_0}\right)\int_{-T_0/2}^{T_0/2} x(t)x(t+\tau)dt \quad \text{при } -\infty < \tau < \infty, \tag{12.4}$$

где

$$K = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} x^2(t)\,dt. \tag{12.5}$$

Если $x(t)$ является периодическим импульсным сигналом, представляющим псевдослучайный код, каждый из элементарных импульсов такого сигнала называют *кодовым символом* (code symbol) или *элементарным сигналом* (chip). Нормированная автокорреляционная функция псевдослучайного сигнала с единичной длительностью элементарного сигнала и периодом p элементарных сигналов может быть записана следующим образом.

$$R_x(\tau) = \frac{1}{p}\cdot\left(\begin{array}{l}\text{разница между числом соответствий и несоответствий}\\ \text{при сравнении одного полного периода последовательности}\\ \text{с ее модификацией, полученной путем циклического}\\ \text{сдвига на } \tau \text{ позиций}\end{array}\right) \tag{12.6}$$

График нормированной автокорреляционной функции последовательности максимальной длины $R_x(\tau)$ показан на рис. 12.8. Очевидно, что для $\tau = 0$, т.е. когда сигнал $x(t)$ и его копия идеально совпадают, $R(\tau) = 1$. В то же время для любого циклического сдвига между $x(t)$ и $x(t + \tau)$ при $(1 \le \tau < p)$ автокорреляционная функция равна $-1/p$ (для больших значений p последовательности практически декоррелируют между собой при сдвиге на один элементарный сигнал).

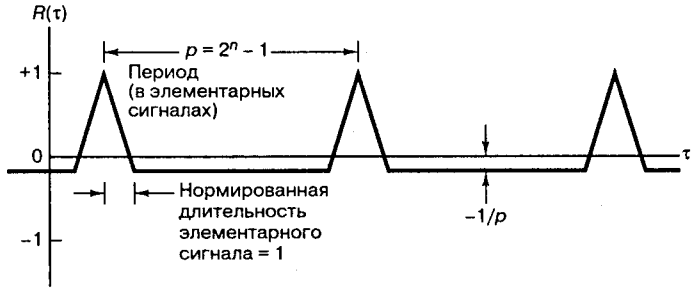


Рис. 12.8. Автокорреляционная функция псевдослучайной последовательности

Теперь легко можно провести проверку свойства корреляции для псевдослучайной последовательности, сгенерированной регистром сдвига на рис. 12.7. Запишем выходную последовательность и ее модификацию со сдвигом на один регистр вправо.

0	0	0	1	0	0	1	1	0	1	0	1	1	1	1
1	0	0	0	1	0	0	1	1	0	1	0	1	1	1
d	*a*	*a*	*d*	*d*	*a*	*d*	*a*	*d*	*d*	*d*	*d*	*a*	*a*	*a*

Совпадение цифр отмечено символом *a*, а несовпадение — *d*. Согласно уравнению (12.6) автокорреляционная функция при подобном сдвиге на один элементарный сигнал равна следующему.

$$R(\tau = 1) = \frac{1}{15}(7 - 8) = -\frac{1}{15}$$

Любой циклический сдвиг, который приводит к отклонению от идеальной синхронизации, дает значение автокорреляционной функции $-1/p$. Следовательно, третье свойство псевдослучайной последовательности в данном случае выполняется.

12.3. Системы расширения спектра методом прямой последовательности

На блок-схеме, приведенной на рис. 12.9, *а*, изображен модулятор схемы *прямой последовательности* (direct-sequence — DS). "Прямая последовательность" — это модуляция несущей информационным сигналом $x(t)$ с последующей модуляцией высокоскоростным (широкополосным) расширяющим сигналом $g(t)$. Рассмотрим модулированную данными несущую с постоянной огибающей, которая имеет мощность P, угловую частоту ω_0, информационную модуляцию фазы $\theta_x(t)$.

$$s_x(t) = \sqrt{2P}\cos[\omega_0 t + \theta_x(t)] \tag{12.7}$$

После модуляции расширяющим сигналом $g(t)$ с постоянной огибающей переданный сигнал можно представить в следующем виде:

$$s(t) = \sqrt{2P}\cos[\omega_0 t + \theta_x(t) + \theta_g(t)], \tag{12.8}$$

причем фаза несущей теперь состоит из двух компонентов: $\theta_x(t)$, который соответствует данным, и $\theta_g(t)$, возникший из-за применения расширяющего сигнала.

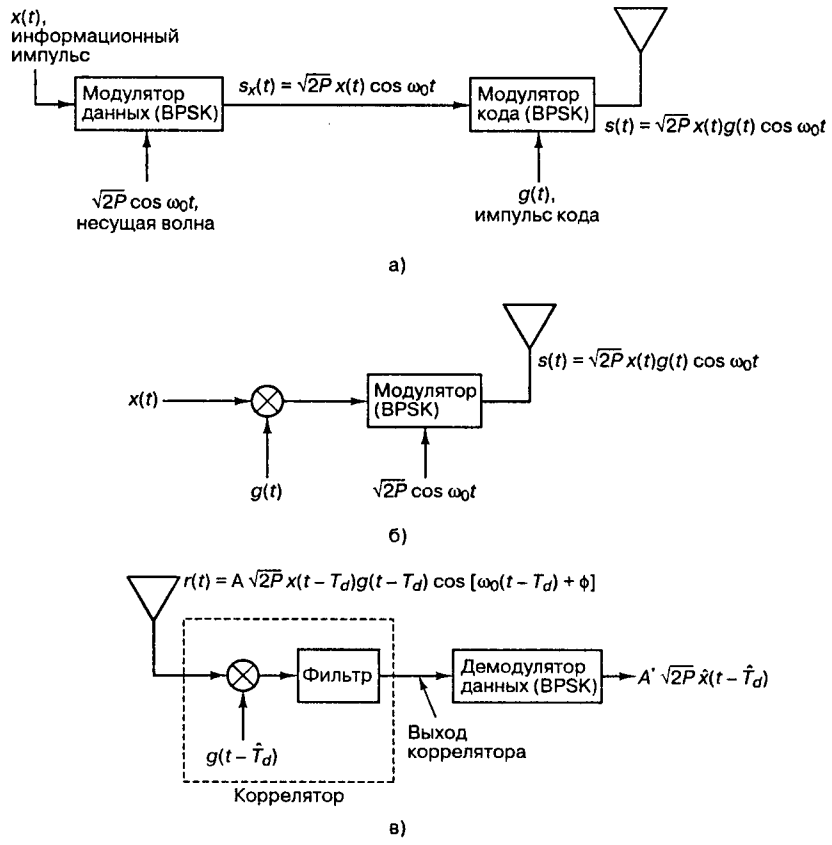


Рис. 12.9. Система расширения спектра методом прямой последовательности: а) передатчик BPSK; б) упрощенный передатчик BPSK; в) приемник BPSK

В главе 4 было показано, что двоичная фазовая манипуляция (binary phase shift keying — BPSK) с подавлением несущей приводит к мгновенным изменениям фазы несущей на π радиан согласно передаваемой информации. Формулу (12.7) также можно записать как произведение несущей и $x(t)$, потока антиподных импульсов со значениями импульсов +1 либо –1.

$$s_x(t) = \sqrt{2P}\,x(t)\cos\omega_0 t \tag{12.9}$$

Если модуляция расширяющей последовательности — это также BPSK, а $g(t)$ — антиподный поток импульсов со значениями импульсов +1 либо –1, уравнение (12.8) может быть представлено в следующем виде.

$$s(t) = \sqrt{2P}\,x(t)g(t)\cos\omega_0 t \tag{12.10}$$

Модулятор, построенный согласно формуле (12.10), изображен на рис. 12.9, *б*. Вначале производится перемножение потока импульсных данных и расширяющего сигнала, после чего несущая модулируется полученным сигналом $x(t)$. Если присвоение значений импульсов бинарным значениям выполняется следующим образом

Значение импульса	Двоичное значение
1	0
−1	1

то исходный этап модуляции DS/BPSK может выполняться путем суммирования по модулю 2 двоичной информационной последовательности и двоичной расширяющей последовательности.

Демодуляция сигнала DS/BPSK производится с помощью вычисления корреляции или повторной модуляции принятого сигнала синхронизированной копией расширяющего сигнала $g(t-\hat{T}_d)$ (рис. 12.9, *в*), где $\hat{T}_d$ — оценка приемником задержки распространения T_d между передатчиком и приемником. При отсутствии шумов и интерференции выходной сигнал коррелятора может быть записан следующим образом:

$$A\sqrt{2P}x(t-T_d)g(t-T_d)g(t-\hat{T}_d)\cos[\omega_0(t-T_d)+\phi], \tag{12.11}$$

где постоянная A — коэффициент усиления системы, ϕ — случайное значение фазового угла из диапазона $(0, 2\pi)$. Поскольку $g(t)=\pm 1$, произведение $g(t-T_d)g(t-\hat{T}_d)$ будет равно единице, если $\hat{T}_d = T_d$, т.е. если кодовый сигнал в приемнике точно синхронизирован с кодовым сигналом в передатчике. При такой синхронизации выход принимающего коррелятора — это суженный сигнал, модулированный данными (за исключением случайной фазы ϕ и времени T_d). После этого для восстановления исходных данных используется обычный демодулятор.

12.3.1. Пример схемы прямой последовательности

На рис. 12.10 приводится пример процессов модуляции и демодуляции DS/BPSK, выполняемых в соответствии с блок-схемами рис. 12.9, *б* и *в*. На рис. 12.10, *а* показана двоичная информационная последовательность (1, 0) и ее эквивалент в виде биполярного импульсного сигнала $x(t)$. Присвоение двоичных значений импульсам выполняется аналогично случаю, описанному в предыдущем разделе. Примеры двоичной расширяющей последовательности и ее биполярного эквивалента $g(t)$ приводятся на рис. 12.10, *б*. Результат суммирования по модулю 2 информационной и кодовой последовательностей, а также произведение $x(t)g(t)$ представлены на рис. 12.10, *в*.

Как показано на рис. 12.10, *г*, при модуляции BPSK (см. уравнения (12.8) и (12.10)) фаза несущей $\theta_x(t) + \theta_g(t)$ равна π, если произведение сигналов $x(t)g(t)$ равно −1 (или сумма по модулю 2 данных и кода является двоичной единицей). Подобным образом фаза несущей равна нулю, если значение $x(t)g(t)$ равно +1 (или сумма по модулю 2 данных и кода равна двоичному нулю). При сравнении рис. 12.10, *б* и *в* легко заметить, что важной особенностью сигналов расширенного спектра является их *скрывающее свойство*. График на рис. 12.10, *в* содержит "скрытый" сигнал $x(t)$. Глядя на график, сложно выделить медленно меняющийся информационный сигнал из быстро меняющейся кодовой последовательности. Аналогичная сложность возникает при восстановлении приемником сигнала, если отсутствует точная копия кодового сигнала.

Как видно из рис. 12.10, *в*, демодуляция DS/BPSK проходит в два этапа. Первый этап — сужение полученного сигнала — выполняется путем определения корреляции этого сигнала с синхронизированной копией кодового сигнала. Второй этап — демодуляция данных — производится с помощью обычного демодулятора. На рис. 12.10, *д*

представлена копия кода $\hat{\theta}_g(t)$ в виде сдвига фазы (0 или π), который осуществляется приемником с целью сужения кода. На рис. 12.10, *е* представлен процесс вычисления фазы несущей $\hat{\theta}_x(t)$ после сужения либо после суммирования $\hat{\theta}_g(t)$ и $\theta_x(t) + \theta_g(t)$. После указанных преобразований исходные данные фактически уже восстановлены и представлены в виде значений фазы несущей. Завершающий этап, показанный на рис. 12.10, *ж*, предполагает восстановление информационного сигнала $\hat{x}(t)$ с помощью демодулятора BPSK.

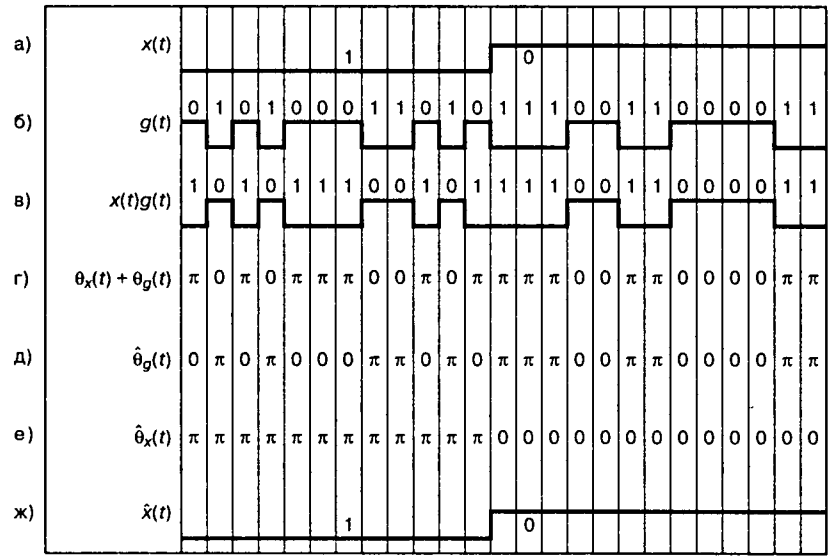


Рис. 12.10. Пример расширения спектра методом прямой последовательности: а) исходные двоичные данные; б) кодовая последовательность; в) переданная последовательность; г) фаза переданной несущей; д) фазовый сдвиг, выполненный кодом приемника; е) фаза принятой несущей после сдвига фаз (сужения); ж) демодулированный информационный сигнал

12.3.2. Коэффициент расширения спектра и производительность

Фундаментальным вопросом в использовании систем расширенного спектра является предлагаемая ими *степень защиты* сигнала от помех ограниченной мощности. Методы расширения спектра расширяют относительно низкоразмерный сигнал в многомерное сигнальное пространство. Сигнал "скрыт" в этом сигнальном пространстве, поскольку предполагается, что станции-постановщику преднамеренных помех неизвестны координаты передачи сигнала в каждый момент времени. Связь можно нарушить путем создания помех во всем диапазоне, используя при этом всю ограниченную мощность генератора. В этом случае в каждой точке диапазона будут присутствовать помехи ограниченной мощности. Еще одним способом нарушения связи может быть создание помех в некоторых точках диапазона. Соответственно, весь остальной диапазон будет свободен от преднамеренных шумов.

Рассмотрим набор из D ортогональных сигналов $s_i(t)$, $1 \le i \le D$, в N-мерном пространстве. Будем считать, что в общем случае $D << N$. В соответствии с выкладками, приведенными в разделе 3.1.3, можно записать следующее.

$$s_i(t) = \sum_{j=1}^{N} a_{ij}\psi_j(t) \quad i = 1, 2, \ldots, D; \quad 0 \le t \le T \qquad D << N \tag{12.12}$$

где

$$a_{ij} = \int_0^T s_i(t)\psi_j(t)dt, \tag{12.13}$$

а также

$$\int_0^T \psi_j(t)\psi_k(t)dt = \begin{cases} 1 & \text{при } j = k \\ 0 & \text{при } j \neq k \end{cases}. \tag{12.14}$$

Линейно независимые функции $\{\psi_j(t)\}$ *охватывают* или характеризуют N-мерное ортогональное пространство; их называют *базисными* функциями пространства. При передаче каждого информационного символа, чтобы скрыть D-мерный сигнал в N-мерном пространстве с помощью псевдослучайного расширяющего кода, независимо выбирается набор коэффициентов $\{a_{ji}\}$. Набор случайных переменных $\{a_{ij}\}$ может с вероятностью 1/2 иметь значение $\pm a$. Для корректного сужения сигнала приемник, разумеется, должен иметь доступ к каждому набору коэффициентов. Характерно, что даже если передача одного и того же i-го символа многократно повторяется, набор $\{a_{ij}\}$ выбирается заново для каждого процесса передачи. Предположим, что энергия всех сигналов набора D одинакова. Тогда среднюю энергию сигнала можно записать в следующем виде:

$$E_s = \int_0^T \overline{s_i^2(t)}dt = \sum_{j=1}^{N} \overline{a_{ij}^2} \quad i = 1, 2, \ldots, D, \tag{12.15}$$

где черта над выражением означает математическое ожидание по ансамблю большого числа процессов передачи символов. Независимые коэффициенты имеют нулевое среднее и корреляцию.

$$\overline{a_{ij}a_{ik}} = \begin{cases} \dfrac{E_s}{N} & \text{при } j = k \\ 0 & \text{при } j \neq k \end{cases} \tag{12.16}$$

Обычно считается, что станция умышленных помех не обладает априорной информацией о наборе коэффициентов $\{a_{ji}\}$. С точки зрения станции помех коэффициенты равномерно распределены по N базисным координатам. Если помехи создаются равномерно по всему диапазону, сигнал помех $w(t)$ может быть записан в следующем виде.

$$w(t) = \sum_{j=1}^{N} b_j\psi_j(t) \tag{12.17}$$

Полная энергия такого сигнала равна следующему.

$$E_w = \int_0^T w^2(t)dt = \sum_{j=1}^{N} b_j^2 \tag{12.18}$$

Станция умышленных помех может выработать стратегию выбора частей b_j^2 полной (фиксированной) энергии E_w таким образом, чтобы свести к минимуму отношение сигнал/шум (signal-to-noise ratio — SNR) в приемнике после демодуляции.

Выходной сигнал детектора в приемнике

$$r(t) = s_i(t) + w(t) \tag{12.19}$$

коррелирует с набором переданных сигналов (собственными шумами приемника пренебрегаем), так что выход i-го коррелятора можно записать в следующем виде.

$$z_i = \int_0^T r(t)s_i(t)dt = \sum_{j=1}^{N}(a_{ij}^2 + b_j a_{ij}) \tag{12.20}$$

Усредненное значение второго члена правой части уравнения (12.20) по ансамблю всех возможных псевдослучайных кодовых последовательностей равно нулю, поскольку считается, что элементы множества случайных переменных $\{a_{ij}\}$ с вероятностью 1/2 принимают значения $\pm a$. Следовательно, если считать, что передан был сигнал $s_m(t)$, математическое ожидание выхода i-го коррелятора может быть записано в следующем виде [6, 7].

$$\mathbf{E}(z_i|s_m) = \sum_{j=1}^{N}\overline{a_{ij}^2} = \begin{cases} E_s & \text{при } i = m \\ 0 & \text{при } i \neq m \end{cases} \tag{12.21}$$

Для случая $i = m$ член $\mathbf{E}(z_i|s_m)$ можно интерпретировать следующим образом. Пусть требуется передать сигнал $s_i(t)$. Выбирается N псевдослучайных коэффициентов a_{ij} ($1 \leq j \leq N$). При этом считается, что при восстановлении исходных данных приемник имеет доступ к каждому набору a_{ij}. Таким образом, хотя при вычислении $\mathbf{E}(z_i|s_m)$ i-й информационный символ задается в передатчике, передается набор коэффициентов, которые кажутся случайными для постороннего приемника. Отметим, что в уравнении (12.21) не учитывается возможность использования станцией умышленных помех изощренных, усложненных методов (описанных в разделе 12.6).

Предположим, что D сигналов равновероятны. Тогда математическое ожидание для выхода любого из D корреляторов можно записать следующим образом.

$$\mathbf{E}(z_i) = \frac{E_s}{D} \tag{12.22}$$

Подобным образом с помощью уравнений (12.15)–(12.21) вычисляем $\mathrm{var}(z_i|s_i)$, дисперсию выхода i-го коррелятора, считая что передан i-й сигнал.

$$\mathrm{var}(z_i|s_i) = \sum_{j,k} b_j b_k \overline{a_{ij}a_{ik}} = \tag{12.23}$$

$$= \sum_{j=1}^{N} b_j^2 \overline{a_{ij}^2} =$$

$$= \sum_{j=1}^{N} b_j^2 \frac{E_s}{N} = \tag{12.24}$$

$$= \frac{E_w E_s}{N}$$

Для полноты рассмотрения можно подобным образом вычислить дисперсию выхода i-го коррелятора после передачи m-го сигнала ($i \neq m$).

$$\mathrm{var}(z_i | s_m) = \frac{E_w E_m}{N} + \frac{E_s^2}{N} \tag{12.25}$$

Отношение мощности сигнала к мощности преднамеренной помехи (signal-to-jammer ratio — SJR) на выходе i-го коррелятора может быть определено следующим образом.

$$SJR = \sum_{m=1}^{D} \frac{\mathbf{E}^2(z_i | s_m)}{\mathrm{var}(z_i | s_m)} P(s_m) = \frac{E_s^2 / D}{E_w E_s / N} = \frac{E_s N}{E_w D} \tag{12.26}$$

Поскольку считается, что вероятность передачи каждого из сигналов одинакова, вероятность передачи m-го сигнала $P(S_m)$ равна $1/D$. Энергия сигнала и помехи обозначается, соответственно, $\mathbf{E}^2(z_i)$ и $\mathrm{var}(z_i)$. В соответствии с уравнением (12.21) члены суммы в (12.26) не равны нулю только при $i = m$. Таким образом, результат не зависит от распределения энергии станции умышленных помех. Какими бы ни были коэффициенты b_j в сумме $\sum_j b_j^2 = E_w$, значение SJR в уравнении (12.26) свидетельствует о том, что при расширении спектра энергия сигнала превосходит энергию помех в N/D раз. Данное отношение N/D называют *коэффициентом расширения спектра* (processing gain) G_p.

Если считать размерность сигнала с шириной полосы W и длительностью T приблизительно равной $2WT$, коэффициент расширения спектра можно записать в следующем виде.

$$G_p = \frac{N}{D} \approx \frac{2W_{ss}T}{2W_{min}T} = \frac{W_{ss}}{R}, \tag{12.27}$$

где W_{ss} — ширина полосы расширенного спектра (полная ширина полосы, используемая в методе расширения), W_{min} — минимальная ширина полосы данных (считается равной скорости передачи данных, R). Для систем с использованием метода прямой последовательности W_{ss} и W_{min} приблизительно равны, соответственно, скорости передачи элементарных сигналов R_{ch} и скорости передачи данных R. В результате можно записать следующее.

$$G_p = \frac{R_{ch}}{R} \tag{12.28}$$

В данном случае под *элементарным сигналом* (chip) подразумевается наименьший непрерывный сигнал в системе. Для систем расширения спектра методом прямой последовательности элементарный сигнал представляет собой импульс (или элемент сигнала) псевдослучайного кода.

В любом случае использования расширенного спектра (например, для подавления интерференции или достижения высокого временного разрешения) коэффициент расширения спектра — это параметр, описывающий преимущество системы расширенного спектра перед узкополосной системой. В общем случае для модуляции сигнала в системе расширения спектра методом прямой последовательности используется схема BPSK или QPSK. Предположим, что двоичный символ состоит из 1000 элементарных кодовых сигналов BPSK. В соответствии с уравнением (12.28) коэффициент расширения спектра в данном случае будет равен 1000. Для демонстрации того, что такая система расширенного спектра позволяет более устойчивую передачу (относительно узкополосной системы), рассмотрим следующий пример. Представим, что в процессе обнаружения решение относительно значения принятого символа принимается для каждого из 1000 элементарных сигналов. Разумеется, в действительности такое не происходит; 1000 элементарных сигналов собираются, и проверяется их корреляция с кодом, что порождает единое решение относительно значения бита. Но даже если принять такую схему, то бит будет обнаружен правильно, даже если 499 решений из 1000 будут неверными.

12.4. Системы со скачкообразной перестройкой частоты

В данном разделе рассматривается метод *скачкообразной перестройки частоты* (frequency hopping — FH). Для модуляции в данной схеме обычно используется M-арная частотная манипуляция (M-ary frequency shift keying — MFSK). При этой модуляции $k = \log_2 M$ информационных бит используются для определения одной из M передаваемых частот. Положение M-арного множества сигналов скачкообразно изменяется синтезатором частот на псевдослучайную величину, принадлежащую полосе W_{ss}. На рис. 12.11 представлена блок-схема системы FH/MFSK наиболее распространенного типа. В обычной системе MFSK несущая с *фиксированной частотой* модулируется символом данных; в системе FH/MFSK частота несущей является *псевдослучайной*. В обоих случаях передается один тон. Систему FH на рис. 12.11 можно рассматривать как двухэтапный процесс модуляции — модуляции информации и модуляции с перестройкой частоты — хотя он может быть реализован и как один этап, когда синтезатор частот производит тон передачи, основываясь на псевдослучайном коде и информационной последовательности. При каждом скачке генератор псевдослучайного сигнала передает синтезатору частот частотное слово (последовательность из l элементарных сигналов), которое определяет одну из 2^l позиций множества символов. Минимальное разнесение по частоте между последовательными скачками Δf и шириной полосы перестройки частот W_{ss} определяет минимальное количество элементарных сигналов частотного слова.

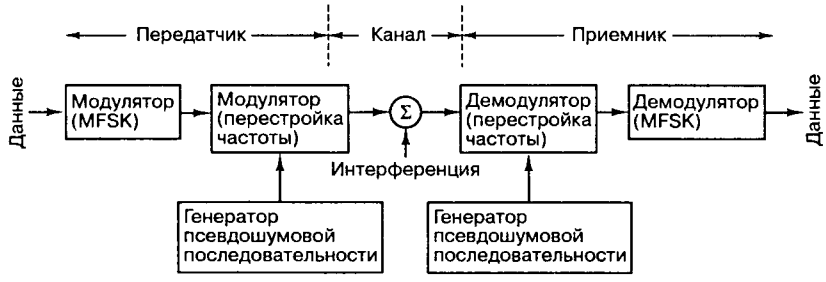


Рис. 12.11. Система FH/MFSK

Для данного скачка ширина полосы, необходимая для передачи, будет такой же, как и в обычной схеме MFSK, что, как правило, намного меньше W_{ss}. В то же время при усреднении по множеству скачков спектр FH/MFSK будет занимать всю полосу расширенного спектра. Метод расширенного спектра позволяет для перестройки частоты использовать полосы шириной порядка несколько гигагерц, что намного превышает аналогичные показатели систем DS [8]. Следовательно, коэффициент расширения спектра сигнала систем FH будет значительно больше. Из-за использования в случае FH полос значительной ширины сохранение фазовой когерентности от скачка к скачку является нелегкой задачей. Поэтому обычно в таких системах применяется некогерентная демодуляция. Рассмотрение когерентных систем с скачкообразной перестройкой частоты представлено в работе [9].

Как видно из рис. 12.11, приемник повторяет все операции передатчика в обратной последовательности. Полученный сигнал демодулируется путем наложения той же псевдослучайной тоновой последовательности, что использовалась для перестройки частоты. После этого сигнал обрабатывается стандартным набором из M некогерентных детекторов энергии с целью выбора наиболее вероятного символа.

Пример 12.1. Размер частотного слова

Ширина полосы системы W_{ss} равна 400 МГц; минимальное изменение частоты $\Delta f = 100$ Гц. Определите минимальное число элементарных сигналов псевдослучайного кода, необходимое для создания частотного слова.

Решение

$$\text{Число тонов, содержащихся в } W_{ss}, \text{ равно } \frac{W_{ss}}{\Delta f} = \frac{400\,\text{МГц}}{100\,\text{Гц}} = 4 \times 10^6$$

$$\text{Минимальное число элементарных сигналов} = \lceil \log_2(4 \times 10^6) \rceil = 22,$$

где $\lceil x \rceil$ — наименьшее целое, не превышающее x.

12.4.1. Пример использования скачкообразной перестройки частоты

Рассмотрим пример системы с перестройкой частоты, приведенный на рис. 12.12. Входные данные состоят из двоичной последовательности, характеризуемой скоростью передачи данных $R = 150$ бит/с. Модуляция — 8-FSK. Таким образом, скорость передачи символов равна $R_s = R/(\log_2 8) = 50$ символов/с (длительность передачи одного символа $T = 1/50 = 20$ мс). Изменение частоты происходит после передачи отдельного символа, причем скачки синхронизированы во времени с границами символов. Следовательно, скорость скачкообразной перестройки частоты равна 50 скачков/с. На рис. 12.12 представлен график зависимости ширины полосы частот (ось ординат, W_{ss}) от времени (ось абсцисс). Приведенные условные обозначения иллюстрируют присвоение восьмеричных символов FSK частотным тонам. Следует отметить, что разнесение тонов, определенное как $1/T = 50$ Гц, соответствует минимальному значению, которое необходимо для передачи ортогональных сигналов для данной некогерентной системы FSK (см. раздел 4.5.4).

Типичная двоичная информационная последовательность представлена в верхней части рис. 12.12. При использовании модуляции 8-FSK символы формируются из трех бит. При *обычной* модуляции 8-FSK производится передача однополосного тонового сигнала, полученного в соответствии с представленной на рисунке

схемой присвоения. Тоновый сигнал сдвинут по отношению к f_0, *фиксированному* центру частотного диапазона данных. Единственным отличием метода FH/MFSK от MFSK является то, что f_0 *не фиксирована*. При передаче очередного символа f_0 перескакивает на новую частоту, и вместе с ней перемещается вся структура диапазона данных. На рис. 12.12 первый символ последовательности данных, 0 1 1, соответствует тоновому сигналу, который на 25 Гц выше по отношению к f_0. На рисунке пунктирная линия соответствует f_0, непрерывная — тоновому сигналу. Во время передачи второго символа f_0 переходит в новое положение, обозначенное пунктиром. Второй символ, 1 1 0, задает тоновый сигнал на 125 Гц ниже по отношению к f_0. Подобным образом последний символ последовательности (0 0 1) соответствует сигналу, смещенному вверх на 125 Гц по отношению к центру диапазона. Центр частотного диапазона в последнем случае смещается, однако относительное расположение тонов остается прежним.

12.4.2. Устойчивость

В повседневной жизни под *устойчивостью* (robustness) подразумевают силу и выносливость. В контексте систем связи значение этого слова практически не отличается от обыденного. Уровень устойчивости определяет способность сигнала выдерживать искажения в канале (шумы, намеренные помехи, замирание сигнала и т. п.). Вероятность получения сигнала, несколько копий которого передаются на разных частотах, выше, чем в случае единичного сигнала, равного по мощности сумме всех копий. Чем выше разнесение сигнала (разнесенные во времени множественные передачи на разных частотах), тем выше его устойчивость к случайным помехам.

Следующий пример позволит лучше понять смысл сказанного выше. Рассмотрим сообщение, состоящее из четырех символов: s_1, s_2, s_3, s_4. Разнесение можно начать с N-кратного повторения сообщения. Пусть N равно 8. Тогда последовательность символов, называемых *элементарными сигналами* (chips), можно записать в следующем виде.

$$s_1s_1s_1s_1s_1s_1s_1s_1s_2s_2s_2s_2s_2s_2s_2s_2s_3s_3s_3s_3s_3s_3s_3s_3s_4s_4s_4s_4s_4s_4s_4s_4$$

Каждый из элементарных сигналов передается на отдельной частоте (центр диапазона данных сдвигается при передаче каждого символа). Серия сигналов на частотах f_i, f_j, f_k, ... более устойчива к помехам, чем сигнал без такого разнесения. Простым аналогом данного примера может быть сравнение выстрела дробью с выстрелом пулей. Вероятность того, что одна из множества дробинок попадет в цель, выше, чем для одной крупной пули.

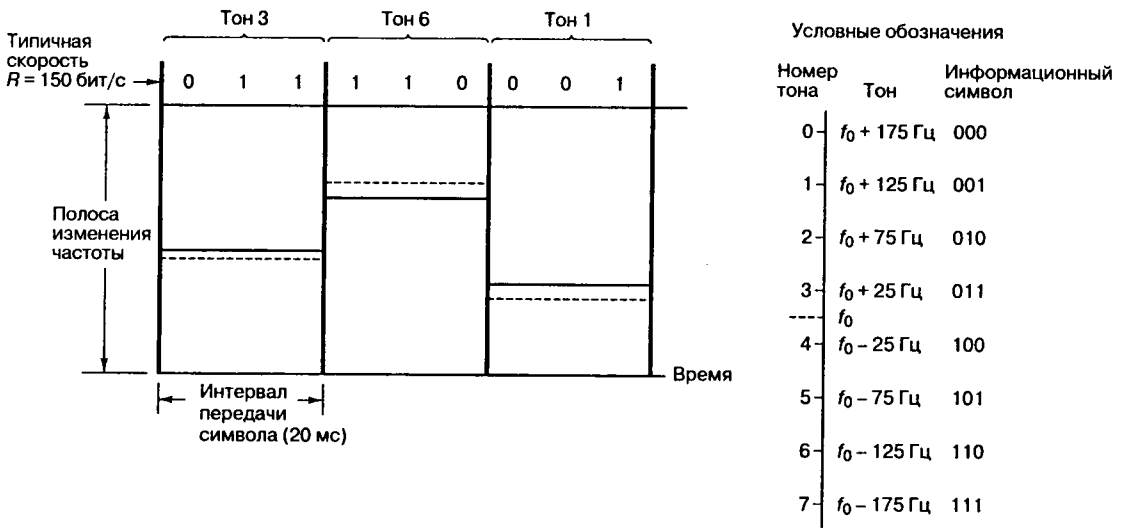


Рис. 12.12. Пример системы связи с использованием скачкообразной перестройки частоты и модуляции 8-FSK

12.4.3. Одновременное использование скачкообразной перестройки частоты и разнесения сигнала

В примере, изображенном на рис. 12.13, каждый из элементарных сигналов передается четыре раза ($N = 4$), в остальном данный случай аналогичен представленному на рис. 12.12. Каждый из интервалов передачи символа (20 мс) разбит на четыре части, которые соответствуют количеству передаваемых элементарных сигналов. Последовательность данных остается такой же, как и для рис. 12.12, и характеризуется скоростью $R = 150$ бит/с. Прежним остается и трехбитовое разбиение с целью формирования 8-ричных символов. Каждый символ передается четырежды, причем для каждого сеанса передачи генератор псевдослучайного кода изменяет центральную частоту диапазона передачи. Следовательно, для данного случая время передачи элементарного сигнала T_c равно $T/N = 20$ мс/4 = 5 мс. Скорость перестройки частоты равна следующему.

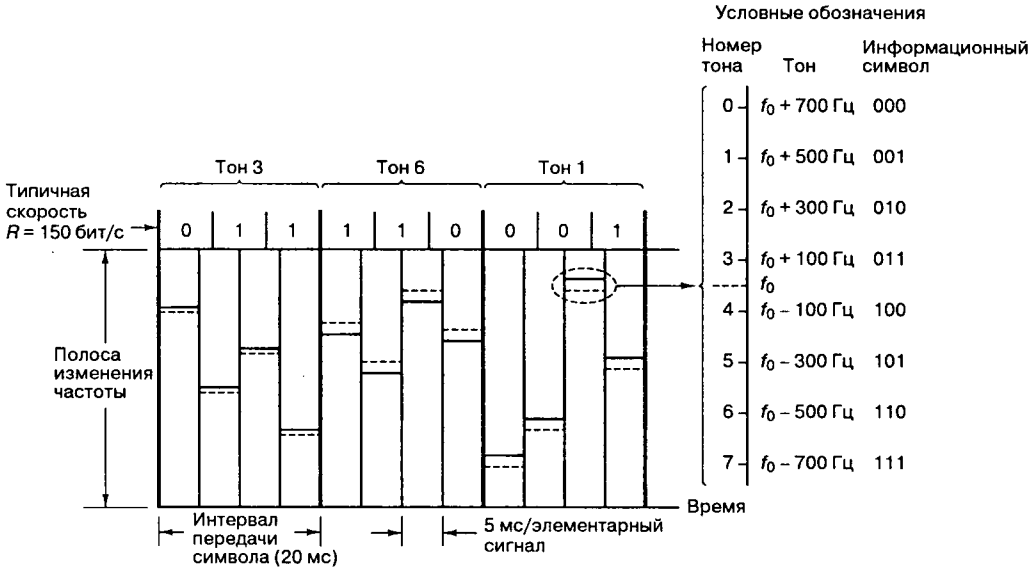


Рис. 12.13. Пример одновременного использования скачкообразной перестройки частоты и разнесения ($N = 4$)

$$\frac{NR}{\log_2 8} = 200 \text{ скачков/с}$$

Следует отметить, что разнесение тонов должно изменяться таким образом, чтобы удовлетворялось требование ортогональности. Поскольку длительность тонов FSK в данном примере равна длительности передачи элементарного сигнала ($T_c = T/N$), минимальное расстояние между тонами $1/T_c = N/T = 200$ Гц. Как и в предыдущем примере, на рис. 12.13 показано смещение центра диапазона передачи данных (и модулирующей структуры) при передаче каждого из элементарных сигналов. Частота передачи (сплошная линия) и центр диапазона передачи данных (пунктир) соотносятся между собой так же, как для каждого из элементарных сигналов, соответствующих определенному символу (рис. 12.12).

12.4.4. Быстрая и медленная перестройка частоты

В системах расширения спектра методом прямой последовательности термин "элементарный символ" означает символ псевдослучайного кода (наиболее короткий символ системы DS). В системе с перестройкой частоты тот же термин обозначает кратчайший непрерывный сигнал. Различают системы связи *медленной* (slow-frequency hopping — SFH) и *быстрой* (fast-frequency hopping — FFH) *перестройки частоты*. Для системы SFH кратчайший непрерывный сигнал — это информационный символ. В случае FFH — это скачок частоты. На рис. 12.14, *а* представлена система FFH со скоростью передачи данных 30 символов/с и скоростью изменения частоты 60 скачков/с. На рисунке показан сигнал $s(t)$ в течение времени передачи одного символа (1/30 с). Изменение формы сигнала в центре графика $s(t)$ связано с очередной скачкообразной перестройкой частоты. В данном примере элементарный сигнал соответствует изменению частоты, поскольку время перестройки меньше длительности символа. Каждый элементарный сигнал соответствует половине символа. На рис. 12.14, *б* иллюстрируется использование системы SFH. Скорость передачи данных по-прежнему равна 30 символов/с; скорость изменения частоты — 10 скачков/с. Сигнал $s(t)$ изображен на протяжении времени передачи трех элементарных сигналов (1/10 с). В данном примере скачки частоты происходят в начале и конце последовательности из трех символов. Форма сигнала меняется вследствие изменений режима модуляции. Теперь элементарный сигнал соответствует информационному символу, длительность которого меньше интервала между изменениями частоты.

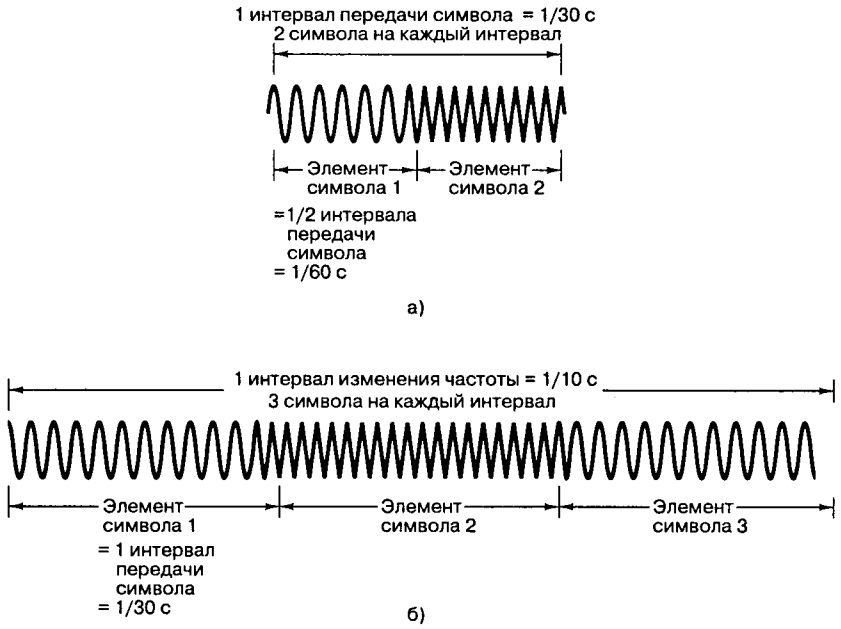


Рис. 12.14. Элементарный сигнал в системах FH/MFSK: а) система MFSK с скачкообразной перестройкой частоты, скорость передачи данных 30 символов/с, скорость изменения частоты 60 скачков/с, 1 элементарный сигнал = 1 интервал между скачками частоты; б) то же, но скорость изменения частоты 10 скачков/с, 1 элементарный сигнал = 1 символ

На рис. 12.15, *а* представлен пример двоичной системы FSK с использованием FFH. Сигнал разделен на $N = 4$ части, т.е. 4 элементарных сигнала соответствуют одному биту. Как и на рис. 12.13, пунктир показывает центр диапазона передачи данных, а непрерывная линия — частоту символа. В данном случае длительность элементарного сигнала равна интервалу между скачками частоты. На рис. 12.15, *б* представлен пример системы FSK с использованием SFH. В этом случае в течение промежутка между скачками частоты производится передача трех бит. В данной схеме SFH длительность элементарного сигнала равна времени передачи одного бита. Каким было бы время передачи элементарного сигнала, если бы в последнем примере система была не двоичной, а восьмеричной, т.е. каждые 3 бит передавались бы как один информационный символ? В этом случае временны*е* границы символа и интервала между скачками частоты совпадали бы. Таким образом, длительность передачи элементарного сигнала, интервал между скачками частоты и время передачи символа были бы одинаковы.

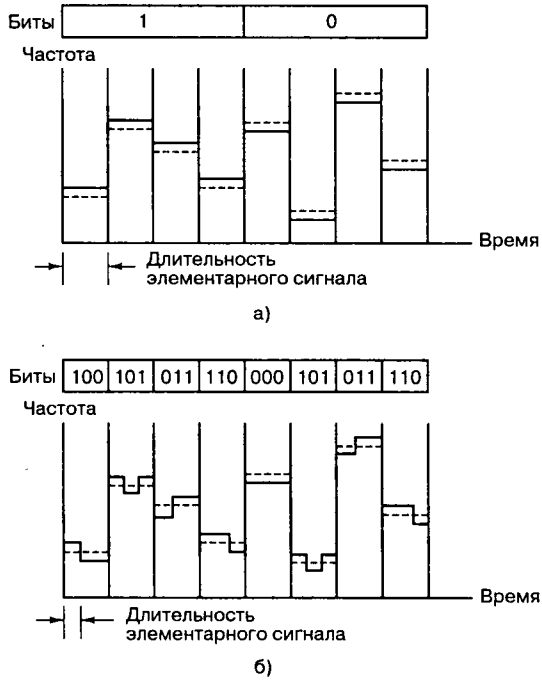


Рис. 12.15. Двоичные системы связи с использованием быстрой и медленной перестройки частоты: а) быстрая перестройка частоты: 4 скачка/бит; б) медленная перестройка частоты: 3 бит/скачок

12.4.5. Демодулятор FFH/MFSK

На рис. 12.16 приводится схема стандартного демодулятора MFSK в системе с быстрой скачкообразной перестройкой частоты (FFH/MFSK). Обработка сигнала начинается с обращения скачков частоты. Для этого используется генератор псевдослучайной последовательности, аналогичный существующему в передатчике. После прохождения через фильтр нижних частот ширина полосы сигнала становится равной ширине полосы данных. Затем сигнал демодулируется с использованием блока из M детекторов

энергии (или детекторов огибающей). За каждым детектором следует схема одностороннего ограничения и накопитель. Схемы ограничения играют важную роль при наличии намеренных помех; их применение будет подробно рассмотрено ниже. Следует отметить, что демодулятор *не принимает* решения относительно значения символов на основе изучения отдельных элементарных сигналов. Вместо этого после получения энергии N элементарных сигналов и после того, как энергия N-го сигнала сложится с энергиями предыдущих $N-1$ сигналов, демодулятор принимает решение, выбирая символ, соответствующий накопителю z_i ($i = 1, 2, \ldots, M$) с максимальной энергией.

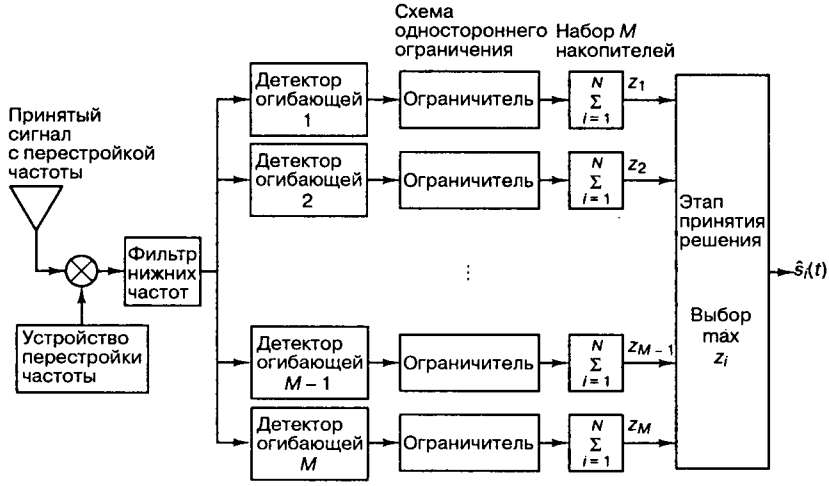


Рис. 12.16. Демодулятор FFH/MFSK

12.4.6. Коэффициент расширения спектра сигнала

В уравнении (12.27) приводится общее выражение для коэффициента расширения спектра сигнала: $G_p = W_{ss}/R$. Для системы расширения спектра методом прямой последовательности величина W_{ss} равна скорости передачи элементарных сигналов R_{ch}. При использовании скачкообразной перестройки частоты уравнение (12.27) также выражает коэффициент расширения спектра, однако значение W_{ss} равно ширине полосы частот, в пределах которой может происходить изменение частоты. Данную полосу называют *полосой перестройки* (hopping band) W_h. Таким образом, коэффициент расширения спектра сигнала для системы со скачкообразной перестройкой частоты можно записать в следующем виде.

$$G_p = \frac{W_h}{R} \tag{12.29}$$

12.5. Синхронизация

В системах расширенного спектра (DS и FH) для успешной демодуляции принятого сигнала приемник должен обладать *синхронизированной* копией расширяющего или кодового сигнала. Процесс синхронизации сгенерированного приемником расширяющего сигнала и полученного сигнала расширенного спектра обычно проходит в

два этапа. На первом этапе два сигнала приводятся в *грубое* соответствие друг другу (процесс *первоначальной синхронизации*). В ходе второго этапа обработки (*этап сопровождения*) с помощью контура обратной связи последовательно выбирается сигнал, наиболее *точно* соответствующий полученному.

12.5.1. Первоначальная синхронизация

Задача данного этапа — синхронизировать полученный сигнал расширенного спектра и локально сгенерированный сигнал расширения путем поиска в двухмерной области временной и частотной неопределенности. Различают когерентные и некогерентные схемы первоначальной синхронизации. В большинстве случаев используется некогерентный метод. Это связано с тем, что обычно сужение сигнала производится до синхронизации несущей. Следовательно, фаза несущей на данном этапе неизвестна. При определении неопределенности по частоте и времени необходимо учитывать следующее.

1. Неопределенность в расстоянии между приемником и передатчиком переходит в неопределенность во времени задержки распространения сигнала.
2. Несоответствия в работе тактовых генераторов приемника и передатчика приводят к разности фаз между соответствующими расширяющими сигналами, которая имеет тенденцию к росту как функция времени, затраченного на синхронизацию.
3. Неопределенность в скорости движения приемника относительно передатчика переходит в неопределенность значения доплеровского сдвига частоты в полученном сигнале.
4. Относительное несоответствие между частотными генераторами приемника и передатчика приводит к сдвигам частот между двумя сигналами.

12.5.1.1. Структуры корреляторов

Общая особенность всех методов синхронизации — определение корреляции полученного и сгенерированного сигналов с целью создания меры их схожести. Затем эта мера сравнивается с пороговой величиной для определения, синхронны ли сигналы. Если да, приемник переходит к этапу сопровождения[1]. В противном случае он изменяет частоту или фазу сгенерированного кода (что фактически является поиском во временной и частотной областях), после чего снова проверяется корреляция.

Рассмотрим простой пример синхронизации в системе расширения спектра методом прямой последовательности с использованием *параллельного поиска* (рис. 12.17). Сгенерированный приемником код $g(t)$ передается с задержками, которые вводятся через половину периода передачи элементарного сигнала ($T_c/2$). Если неопределенность во времени между полученным сигналом и локальным кодом равна времени передачи N_c элементарных сигналов, а полный параллельный поиск в области временной неопределенности должен быть произведен в течение одного непрерывного временного интервала, то используется $2N_c$ корреляторов. Все корреляторы одновременно изучают последовательность из λ элементарных сигналов, после чего сравниваются выходы всех корреляторов. В завершение выбирается локальный код, соответствующий коррелятору с максимальным выходом. Концептуально — это простейший метод поиска; в нем одновременно анализируются все возможные позиции кода (или

[1] Довольно часто для снижения вероятности появления ложных тревог пороговая величина дополнительно проверяется соответствующим алгоритмом до начала этапа сопровождения [4].

фрагментов кода) и для выбора нужного кода используется алгоритм максимального правдоподобия. Выходной сигнал каждого детектора является суммой полученного сигнала и шума. По мере возрастания λ вероятность возникновения ошибки синхронизации (т.е. неверного согласования кода) уменьшается. Следовательно, величину λ следует выбирать таким образом, чтобы одновременно минимизировать время поиска и вероятность возникновения ошибок синхронизации.

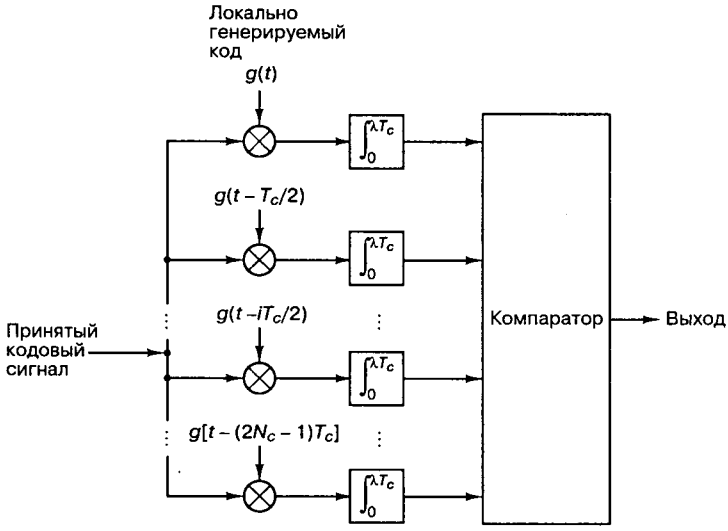


Рис. 12.17. Получение синхронизации в схеме прямой последовательности с использованием метода параллельного поиска

На рис. 12.18 приводится схема синхронизации системы связи со скачкообразной перестройкой частоты. Предположим, что в качестве шаблона синхронизации (без модуляции данных) используется последовательность из N частот, являющаяся частью последовательности скачков частоты. Для первичной обработки полученного сигнала применяется N некогерентных согласованных фильтров, каждый из которых состоит из смесителя частот, полосового фильтра и квадратичного детектора огибающей (последовательно соединенного детектора огибающей и квадратичного устройства). Если процесс скачкообразной перестройки частоты можно описать последовательностью $f_1, f_2, \ldots, f_N$, времена задержки фильтров подбираются таким образом, что при появлении искомой серии скачков частоты система дает выходной сигнал значительной мощности, который и указывает на обнаружение нужной последовательности. Процесс синхронизации может выполняться довольно быстро, поскольку все возможные отклонения кода анализируются одновременно. Следует отметить, что наличие на рис. 12.18 полосовых фильтров указывает, что частоты локального генератора $f_1, f_2, \ldots, f_N$ выбраны таким образом, чтобы их отклонение от ожидаемой частоты сигнала было равно определенной промежуточной частоте (intermediate frequency — IF). Та же система может быть реализована так, что частоты, полученные генератором приемника, будут выбираться без сдвига. Тогда на выходе смесителей будут образовываться узкополосные сигналы. В этом случае фильтры должны быть фильтрами нижних частот (low-pass filter — LPF). В процессе смешивания обычно получается комплексный сигнал, состоящий из синфазного и квадратурного компонентов.

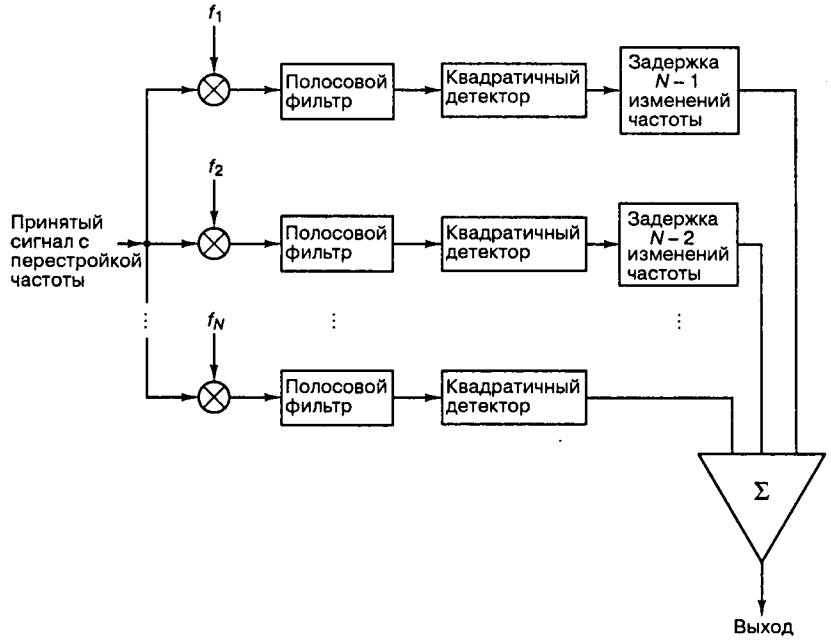


Рис. 12.18. Получение синхронизации для системы связи со скачкообразной перестройкой частоты

Если в течение каждого процесса определения корреляции обрабатываются λ элементарных сигналов длительностью T_c каждый, максимальное время полного параллельного поиска можно записать в следующем виде.

$$(T_{\text{acq}})_{\max} = \lambda T_c \tag{12.30}$$

Среднюю длительность процесса синхронизации можно оценить с помощью параметра *вероятности обнаружения* P_D. P_D характеризует вероятность правильного завершения процесса после обработки λ элементарных сигналов. Если полученный результат неверен, будут обработаны последующие λ элементарных сигналов. Следовательно, средняя длительность процесса обнаружения может быть записана следующим образом [4].

$$\begin{aligned}\overline{T}_{\text{acq}} &= \lambda T_c P_D + 2\lambda T_c P_D (1-P_D) + 3\lambda T_c P_D (1-P_D)^2 + \ldots = \\ &= \frac{\lambda T_c}{P_D}\end{aligned} \tag{12.31}$$

Поскольку число корреляторов или согласованных фильтров, необходимых для полного выполнения процесса параллельного обнаружения, может быть чрезвычайно большим, указанный метод на практике, как правило, не применяется. Вместо схем, изображенных на рис. 12.17 и 12.18, может быть использован единичный коррелятор или согласованный фильтр, производящий *последовательный поиск* до достижения синхронизации. Как и следовало ожидать, компромисс между методами параллельного и последовательного поиска — это компромисс между сложной технической реализацией с быстрой синхронизацией и простой технической реализацией с большим временем синхронизации (при равных скорости передачи данных и неопределенности).

12.5.1.2. Последовательный поиск

Для синхронизации довольно часто используется единичный коррелятор или согласованный фильтр в совокупности с методом последовательного поиска нужной фазы (сигнал DS) или последовательности скачков частоты (сигнал FH). Последовательное повторение процедуры определения корреляции позволяет значительно снизить сложность, размер и стоимость системы. На рис. 12.19 и 12.20 представлены основные конфигурации данной схемы в системе связи расширенного спектра методом прямой последовательности (DS) и скачкообразной перестройки частоты (FH). При пошаговом последовательном получении синхронизации в системе DS устанавливается период синхронизации псевдослучайного локального кода и определяется корреляция данного кода с полученным псевдослучайным сигналом. В течение интервалов поиска λT_c, где $\lambda >> 1$, выходной сигнал сравнивается с заданным пороговым значением. Если порог не достигнут, выходной сигнал увеличивается на установленную часть (обычно 1/2) элементарного сигнала и проверка повторяется. По достижении порогового значения считается, что псевдослучайный код синхронизирован; в результате увеличение фазы кода приемника прекращается, и система переходит к этапу сопровождения. Для системы FH (рис. 12.20) генератор псевдослучайного кода управляет устройством скачкообразной перестройки частоты. Процесс получения синхронизации считается завершенным, когда последовательность скачков частоты локального сигнала совпадает со скачками частоты полученного сигнала.

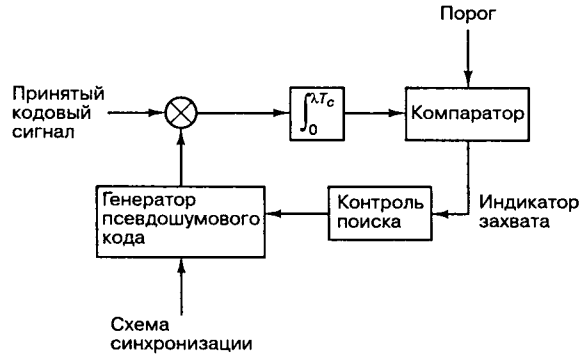


Рис. 12.19. Процесс последовательного поиска для системы, использующей метод прямой последовательности

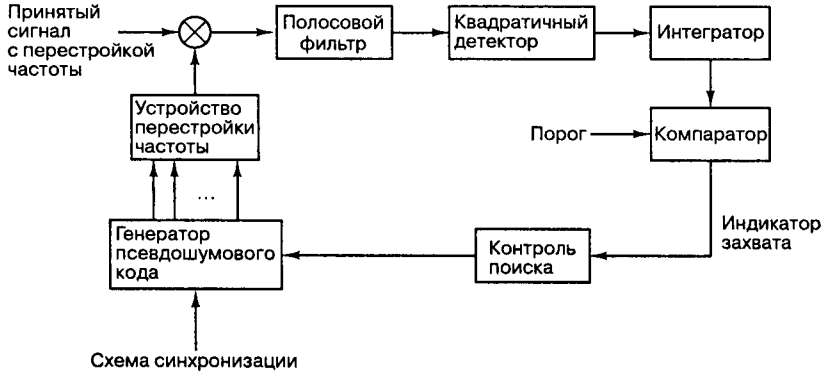


Рис. 12.20. Процесс последовательного поиска для системы с перестройкой частоты

Максимальное время последовательного поиска для системы DS с шагом увеличения 1/2 элементарного сигнала равно следующему.

$$(T_{acq})_{max} = 2N_c\lambda T_c \tag{12.32}$$

Здесь размер области неопределенности, в которой выполняется поиск, равен длительности N_c элементарных сигналов. Среднее время получения синхронизации при использовании последовательного поиска для системы DS при $N_c >> \frac{1}{2}$ будет следующим [10]:

$$\overline{T}_{acq} = \frac{(2 - P_D)(1 + KP_{FA})}{P_D}(N_c\lambda T_c), \tag{12.33}$$

где λT_c — интервал поиска, P_D — вероятность правильного обнаружения, P_{FA} — вероятность ложной тревоги. Определим время, необходимое для проверки правильности обнаружения, равным $K\lambda T_c$, где $K >> 1$. Таким образом, при ложной тревоге будет потеряно $K\lambda T_c$ секунд. При $N_c >> \frac{1}{2}$ и $K << 2N_c$ дисперсия времени синхронизации будет равна следующему.

$$(\text{var})_{acq} = (2N_c\lambda T_c)^2(1 + KP_{FA})\left(\frac{1}{12} + \frac{1}{P_D^2} - \frac{1}{P_D}\right) \tag{12.34}$$

12.5.1.3. Последовательная оценка

Схема использования еще одного метода поиска, *быстрой синхронизации путем последовательной оценки* (rapid acquisition by sequential estimation — RASE), приводится на рис. 12.21. Впервые данный метод был использован Р. Уордом (R. Ward) [10]. Изначально переключатель находится в положении "1". Система вводит свою лучшую оценку первых n элементов полученного сигнала в n разрядов генератора псевдослучайной последовательности. Заполненный регистр определяет начальное состояние генератора. Одним из свойств псевдослучайной последовательности является то, что каждое последующее состояние разрядов зависит только от предыдущего. Следовательно, если оценка первых n элементарных сигналов выполнена верно, все последующие сигналы генератора псевдослучайной последовательности будут правильными. Когда анализ первой последовательности элементарных сигналов закончен, переключатель устанавливается в положение "2". Если начальное состояние регистра было определено верно, генератор приемника создает сигналы, идентичные принятым (при отсутствии шумов). Если выходной сигнал коррелятора после λT_c превышает установленный пороговый уровень, считается, что синхронизация выполнена успешно. В противном случае переключатель возвращается в положение "1", данные регистра обновляются и вся последовательность операций повторяется. Как только система синхронизируется, полученная последовательность элементарных сигналов больше не оценивается. Определим *минимальное* время синхронизации, считая, что шумы отсутствуют. Первые n элементарных сигналов корректно загружены в регистр, поэтому можем записать следующее.

$$T_{acq} = nT_c \tag{12.35}$$

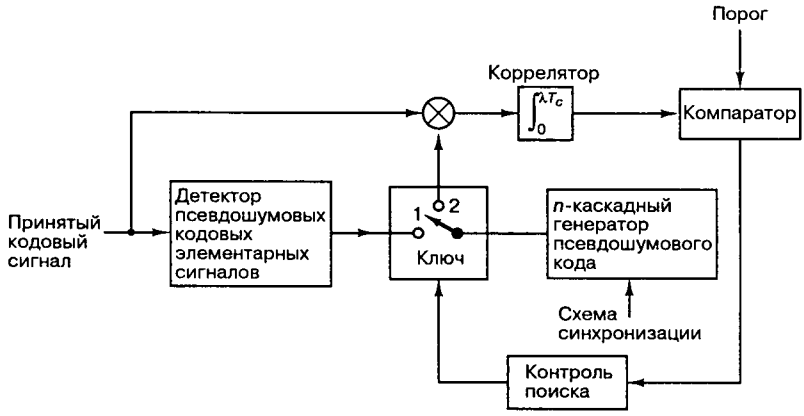


Рис. 12.21. Быстрая синхронизация путем последовательной оценки

Если скорость синхронизации является главным преимуществом системы RASE, ее основной недостаток — высокая чувствительность к помехам и интерферирующим сигналам. Причина такой чувствительности состоит в том, что процесс оценки включает поэлементную демодуляцию по принципу жесткого решения, что не позволяет воспользоваться помехоустойчивыми свойствами псевдослучайного кода. Более подробное описание систем последовательной оценки приводится в работе [4].

12.5.2. Сопровождение

По окончании этапа (грубой) синхронизации начинается этап сопровождения, или достижения идеальной синхронизации. Различают когерентные и некогерентные контуры сопровождения. Когерентным называется контур, где известны частота и фаза несущей волны, а контур сопровождения может работать с узкополосным сигналом. Если же частоту несущей точно определить невозможно (например, из-за доплеровского эффекта) — имеем некогерентный контур. Поскольку в большинстве случаев фаза и частота несущей априори не известны точно, для сопровождения полученного псевдослучайного кода используются именно некогерентные контуры. Кроме того, различают *контуры постоянного сопровождения с задержкой и опережением* (full-time early-late tracking loop), часто называемые контурами *автоподстройки по задержке* (delay-locked loop — DLL), и контуры *сопровождения с задержкой и опережением с разделением времени* (time-shared early-late tracking loop), часто именуемые контурами внесения искусственных флуктуаций (tau-dither loop — TDL). Простой пример применения некогерентного контура DLL в системе расширения спектра методом прямой последовательности при использовании двоичной фазовой манипуляции (binary phase-shift keying — BPSK) представлен на рис. 12.22. Несущая модулируется информационным сигналом $x(t)$ и кодовым сигналом $g(t)$ с использованием схемы BPSK. Как и ранее, считаем, что шумы и интерференция отсутствуют, поэтому можем записать следующее.

$$r(t) = A\sqrt{2P}x(t)g(t)\cos(\omega_0 t + \phi) \tag{12.36}$$

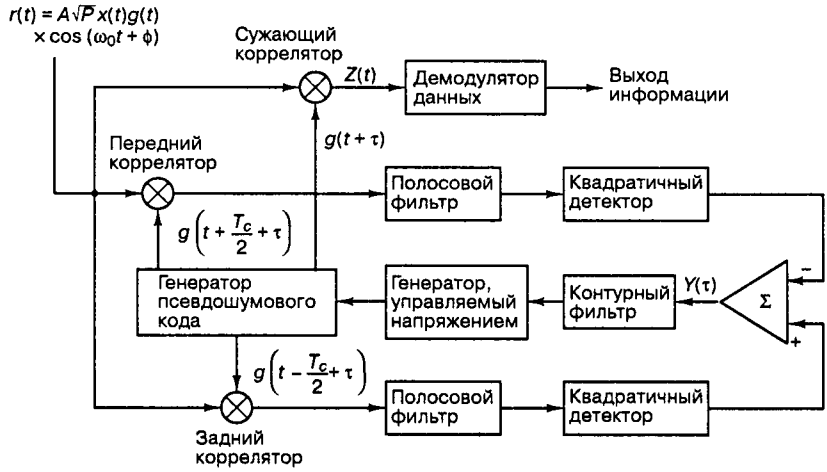


Рис. 12.22. Использование контура DLL для сопровождения сигналов системы DS/SS

Здесь A — коэффициент усиления системы; ϕ — случайный угол сдвига фаз в диапазоне $(0, 2\pi)$. Сгенерированный контуром сопровождения кодовый сигнал сдвинут по отношению к полученному сигналу $g(t)$ на время τ, причем $\tau < T_c/2$. Для проведения точной синхронизации контур генерирует две псевдослучайные последовательности: $g(t + T_c/2 + \tau)$ и $g(t - T_c/2 + \tau)$, одна из которых отстает от другой на время передачи элементарного сигнала. Два узкополосных фильтра предназначаются для пропускания данных, а также для усреднения произведения $g(t)$ и двух псевдослучайных последовательностей $g(t \pm T_c/2 + \tau)$ (в работе [4] указывается оптимальная ширина полосы для данного типа фильтров). Квадратичный детектор огибающей исключает данные, поскольку $|x(t)| = 1$. Выход каждого детектора огибающей можно приблизительно записать следующим образом.

$$E_D \approx \mathbf{E}\left\{\left|g(t)g\left(t \pm \frac{T_c}{2} + \tau\right)\right|\right\} = \left|R_g\left(\tau \pm \frac{T_c}{2}\right)\right| \qquad (12.37)$$

Оператор $\mathbf{E}\{\cdot\}$ обозначает *математическое ожидание*, а $R_g(x)$ — это автокорреляционная функция псевдослучайного сигнала, как показано на рис. 12.8. Сигнал обратной связи $Y(\tau)$ представлен на рис. 12.23. Если τ больше нуля, $Y(\tau)$ указывает генератору, управляемому напряжением, (ГУН) увеличить частоту, что приводит к уменьшению τ. Если значение τ отрицательно, частота ГУН уменьшается, в результате τ возрастает. Если τ — это достаточно малая величина, $g(t)g(t + \tau) \approx 1$, что дает в итоге суженный сигнал $Z(t)$. Впоследствии $Z(t)$ подается на вход обычного демодулятора данных. Подробное описание использования контуров DLL приводится в работах [4, 12–14].

Недостатком контура DLL является то, что цепи опережения и запаздывания должны быть точно синхронизированы, иначе $Y(\tau)$ будет сдвинут по фазе и, соответственно, его значение будет ненулевым при нулевой ошибке. Данная проблема решается с помощью контура с разделением времени. В таком контуре опережающий и запаздывающий корреляторы используются в разное время. Очевидным преимуществом является то, что для работы контура достаточно одного коррелятора. Кроме того, снижается актуальность проблемы смещения постоянной составляющей.

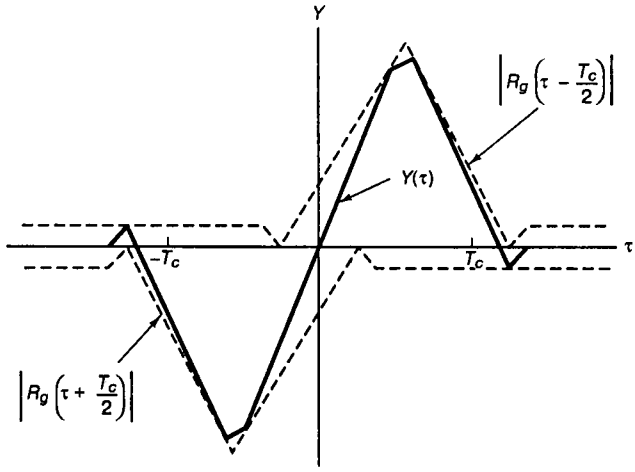


Рис. 12.23. $Y(\tau)$ — сигнал обратной связи контура DLL

При нормальной работе многих управляющих контуров контрольный сигнал практически равен нулю. С этим связан один из недостатков таких систем — нулевой сигнал часто приводит к тому, что контур становится неуправляемым. Особенно остро эта проблема проявляется в сложных контурах сопровождения, которые изменяют коэффициент усиления в зависимости от внешних условий. На рис. 12.24 представлен контур TDL; это одна из разновидностей схем сопровождения с разделением времени. Для решения проблемы нулевого сигнала в данном контуре вводится небольшая намеренная погрешность. В результате выходной сигнал контура как бы "вибрирует" вокруг точного сигнала. Обычно отклонение от нормы невелико, поэтому потери в производительности минимальны. Преимущество контура TDL состоит в том, что для выполнения функций *сопровождения и сужения* кодовой последовательности достаточно одного коррелятора. Как и в случае DLL, проверяется корреляция полученного сигнала с опережающей и запаздывающей версиями псевдослучайного кода приемника. Как показано на рис. 12.24, генератором псевдослучайного кода управляет синхронизирующий сигнал, в фазу которого добавляются псевдослучайные флуктуации, лежащие в пределах квадратичной коммутационной функции. Постоянные изменения фазы позволяют избежать нарушений в работе контура, устраняя необходимость слежения за идентичностью функций в опережающем и запаздывающем контурах. Если боковые фильтры контура TDL спроектированы должным образом, отношение сигнал/шум в этом контуре будет меньше приблизительно на 1,1 дБ по сравнению с контуром DLL [4]. Более подробное описание синхронизации псевдослучайных кодов приводится в работах [4, 15, 16].

12.6. Учет влияния преднамеренных помех

12.6.1. "Состязание" с помехами

При постановке преднамеренных помех основная задача состоит в том, чтобы лишить противника надежной связи и при этом свести материальные затраты к минимуму. Задача приемника и передатчика — создать систему связи, устойчивую к помехам, ос-

новываясь на следующих предположениях: (1) абсолютная устойчивость к помехам невозможна; (2) станция-постановщик преднамеренных помех обеспечена информацией об основных параметрах системы (частотный диапазон, время сеансов связи, объем передаваемой информации и т.д.); (3) станция-постановщик преднамеренных помех не имеет априорной информации о последовательности скачков частоты или псевдослучайных кодах. Передаваемый сигнал должен быть сформирован таким образом, чтобы единственной возможностью для подавления сигнала было создание широкополосного гауссова шума. Другими словами, необходимо, чтобы применение усложненных методов подавления сигнала не давало никаких преимуществ. Основное правило при создании помехоустойчивой системы связи — сделать процесс подавления сигнала максимально дорогостоящим.

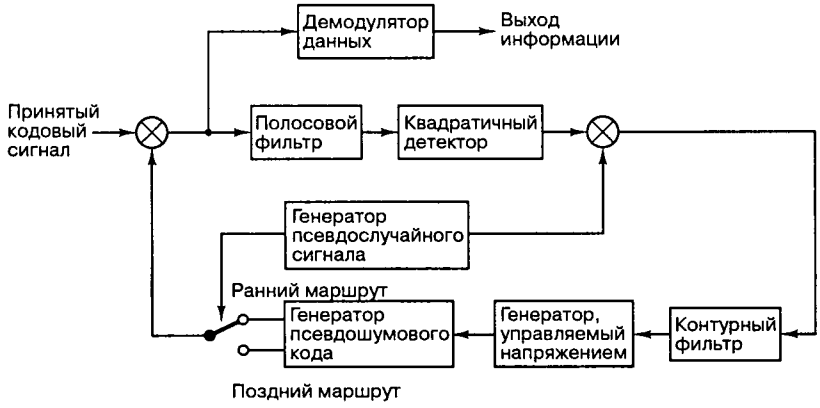


Рис. 12.24. Контур TDL

12.6.1.1. Типы преднамеренных помех

Для подавления связи возможно использование различных сигналов. Выбор зависит от системы связи, сигнал которой требуется подавить. На рис. 12.25 изображены графики спектральной плотности мощности различных типов преднамеренных помех, наложенных на тоновые сигналы системы связи с M-арной частотной манипуляцией и скачкообразной перестройкой частоты (FH/MFSK). Область по оси абсцисс представляет собой полосу расширенного спектра W_{ss}. Три столбца графиков соответствуют трем моментам времени передачи символов (скачкам частоты), в которые происходит передача символов со спектрами G_1, G_2 и G_3. Рис. 12.25, *а* иллюстрирует работу станции преднамеренных помех сравнительно малой мощности, создающей шумы по всей области расширенного спектра. На рис. 12.25, *б* широта покрытия диапазона преднамеренными помехами уменьшается, но при этом увеличивается мощность самих помех (при этом площадь, которую ограничивает кривая мощности шумов, остается постоянной). В данном случае область шумов не всегда совмещается с сигналом. Однако если это все же происходит, негативное влияние на сигнал может быть значительным. На рис. 12.25, *в* помехи создаются в отдельных частях диапазона в случайно выбранные отрезки времени. Использование такого метода не позволяет системе связи адаптироваться к наличию помех. В двух оставшихся случаях для подавления связи используется уже не непрерывная полоса частот, а набор тоновых сигналов, передаваемых в определенных точках диапазона (рис. 12.25, *г*), которые могут размещаться с определенным шагом (рис. 12.25, *д*). Последний метод обычно применяется для по-

давления связи в системах со скачкообразной перестройкой частоты. Еще один метод, не представленный на рис. 12.25, — создание импульсно-модулированного шума с ограниченной шириной полосы. В дальнейшем будем считать (если не оговорено противное), что для подавления связи используется широкополосный шум, который постоянно покрывает всю полосу W_{ss}. Воздействие на сигнал узкополосного шума и ступенчатых помех будет рассмотрено позже.

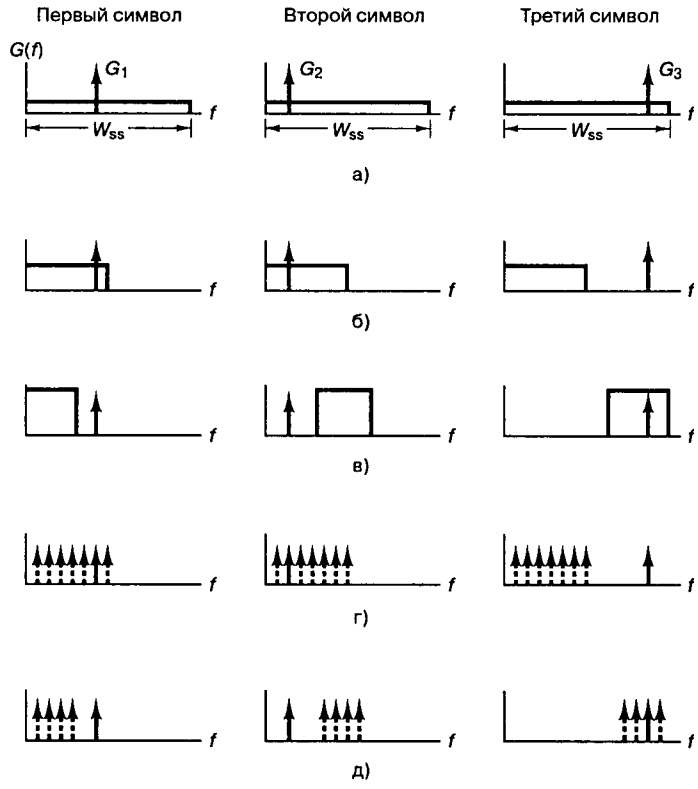


Рис. 12.25. Типы преднамеренных помех: а) широкополосный шум; б) узкополосный шум; в) ступенчатый шум; г) узкополосные тоновые помехи; д) ступенчатые тоновые помехи

12.6.1.2. Защита от помех

Задача помехоустойчивой (anti-jam — AJ) системы связи — добиться истощения ресурсов станции преднамеренных помех с помощью (1) использования широкого диапазона частот (2) в течение длительного времени (3) при передаче из разнесенных точек. Для повышения устойчивости к помехам необходимо использовать (1) разнесение частот посредством расширения спектра методами прямой последовательности и скачкообразной перестройки частоты; (2) разнесение во времени, посредством переключения временных интервалов; (3) пространственное разделение с помощью узконаправленной антенны (в этом случае постановщик преднамеренных помех сможет эффективно использовать лишь боковой лепесток антенны, что дает системе связи дополнительный выигрыш в 20–25 дБ); и (4) различные сочетания первых трех вариантов.

12.6.1.3. Отношение J/S

В главе 5 уровень ошибок в канале связи рассматривался как функция помех со стороны теплового шума. Основное внимание уделялось различию требуемого и фактически имеющегося отношений сигнал/шум E_b/N_0. В данном разделе вероятность ошибок в канале по-прежнему будет рассматриваться как функция помех (суммы теплового шума и широкополосного гауссова шума, созданного станцией преднамеренных помех). Следовательно, отношение сигнал/шум можно записать следующим образом: $E_b/(N_0 + J_0)$, где J_0 — спектральная плотность мощности преднамеренных помех. Будем считать (если не оговорено иное), что J_0 равно J/W_{ss}, где J — средняя мощность преднамеренных помех, полученная приемником; W_{ss} — ширина полосы расширенного спектра. В общем случае мощность станции преднамеренных помех значительно выше мощности теплового шума. Поэтому суммарную величину отношения сигнал/шум обычно считают равной E_b/J_0. Таким образом, подобно случаю теплового шума обозначим через "$(E_b/J_0)_{\text{треб}}$" отношение энергии бита данных к спектральной плотности мощности шума, *требуемое* для поддержания заданного уровня вероятности ошибок в канале. Параметр E_b может быть выражен следующим образом.

$$E_b = ST_b = \frac{S}{R}$$

В данном случае S — мощность полученного сигнала, T_b — время передачи бита, R — скорость передачи данных (бит/с). Тогда $(E_b/J_0)_{\text{треб}}$ может быть записано следующим образом.

$$\left(\frac{E_b}{J_0}\right)_{\text{треб}} = \left(\frac{S/R}{J/W_{ss}}\right)_{\text{треб}} = \frac{W_{ss}/R}{(J/S)_{\text{треб}}} = \frac{G_p}{(J/S)_{\text{треб}}}, \tag{12.38}$$

где $G_p = W_{ss}/R$ — коэффициент расширения спектра сигнала. Отношение сигнал/шум может быть выражено в следующем виде.

$$\left(\frac{J}{S}\right)_{\text{треб}} = \frac{G_p}{(E_b/J_0)_{\text{треб}}} \tag{12.39}$$

Отношение $(J/S)_{\text{треб}}$ — это критерий качества, который определяет степень *невосприимчивости* системы связи к помехам. Какая система имеет больший иммунитет к преднамеренным помехам: система с большим или меньшим $(J/S)_{\text{треб}}$? Чем *больше* $(J/S)_{\text{треб}}$, тем *устойчивее* система к помехам, поскольку данный параметр характеризует мощность шумов, *требуемую* для искажения сеанса связи. Естественно, наиболее желательным для системы связи была бы передача сигнала *вообще* без искажений.

Уравнение (12.39) можно интерпретировать следующим образом. Пытаясь подавить сигнал, противник максимально увеличивает значение $(E_b/J_0)_{\text{треб}}$. Для этого вместо широкополосного шума могут генерироваться тоновые, импульсные или узкополосные помехи. Из большого отношения $(E_b/J_0)_{\text{треб}}$ следует малое значение $(J/S)_{\text{треб}}$ в фиксированном участке полосы. Для увеличения $(J/S)_{\text{треб}}$ сообщающиеся стороны могут увеличить коэффициент расширения спектра сигнала. При проектировании систем связи необходимо выбирать такие сигналы передачи данных, чтобы единственной выигрышной стратегией для генератора помех было создание широкополосного гауссова шума.

12.6.1.4. Порог сопротивляемости помехам

В некоторых случаях соотношение $(J/S)_{\text{треб}}$ называют *порогом сопротивляемости помехам* (anti-jam (AJ) margin), поскольку данный параметр описывает устойчивость системы к попыткам подавления сигнала. Однако использование данного термина не всегда корректно, в общем случае он применяется для обозначения запаса прочности против *конкретной угрозы*. Воспользуемся вычислениями для энергетического резерва системы против теплового шума (глава 5) и определим энергетический резерв системы против преднамеренных помех.

$$M_{\text{AJ}}(\text{дБ}) = \left(\frac{E_b}{J_0}\right)_{\text{прин}}(\text{дБ}) - \left(\frac{E_b}{J_0}\right)_{\text{треб}}(\text{дБ}), \tag{12.40}$$

где $(E_b/J_0)_{\text{прин}}$ — фактическое значение принятого E_b/J_0. По аналогии с уравнением (12.38) $(E_b/J_0)_{\text{прин}}$ можно записать в следующем виде.

$$\left(\frac{E_b}{J_0}\right)_r = \frac{G_p}{(J/S)_r}, \tag{12.41}$$

где $(J/S)_{\text{прин}}$, или просто J/S, — это отношение мощности полученных приемником помех к мощности сигнала. Позднее будет выведено уравнение для E_b/I_0, подобное (12.41), где I_0 характеризует спектральную плотность мощности интерференции, возникающей между несколькими пользователями сотовой системы связи CDMA. Принцип вычисления отношения удельной энергии к мощности помех не изменяется вне зависимости от механизма возникновения шумов: случайная интерференция, преднамеренное подавление сигнала или интерференция между сигналами пользователей в одной спектральной области.

Подставив в уравнение (12.40) выражения (12.38) и (12.41), получим следующее.

$$M_{\text{AJ}}(\text{дБ}) = \frac{G_p}{(J/S)_{\text{прин}}}(\text{дБ}) - \frac{G_p}{(J/S)_{\text{треб}}}(\text{дБ}) \tag{12.42}$$

$$= \left(\frac{J}{S}\right)_{\text{треб}}(\text{дБ}) - \left(\frac{J}{S}\right)_{\text{прин}}(\text{дБ}) \tag{12.43}$$

Пример 12.2. Подавление спутникового сигнала

На рис. 12.26 изображено подавление спутникового сигнала станцией умышленных помех. Устройство связи, расположенное на самолете, оборудовано системой расширения спектра методом скачкообразной перестройки частоты с эффективной изотропно-излучаемой мощностью EIRP_T = 20 дБВт. Скорость передачи данных R = 100 бит/с. Станция преднамеренных помех непрерывно генерирует широкополосный гауссов шум с уровнем EIRP_J = 60 дБВт. Предположим, что $(E_b/J_0)_{\text{треб}}$ = 10 дБ. Также будем считать, что потери мощности при распространении радиоволн одинаковы для устройства, находящегося на самолете, и станции преднамеренных помех.

а) В каком случае помехи представляют большую опасность: при передаче на спутник или при передаче со спутника?

б) Каким должно быть значение ширины полосы системы со скачкообразной перестройкой частоты W_{ss} для получения резерва против помех 20 дБ?

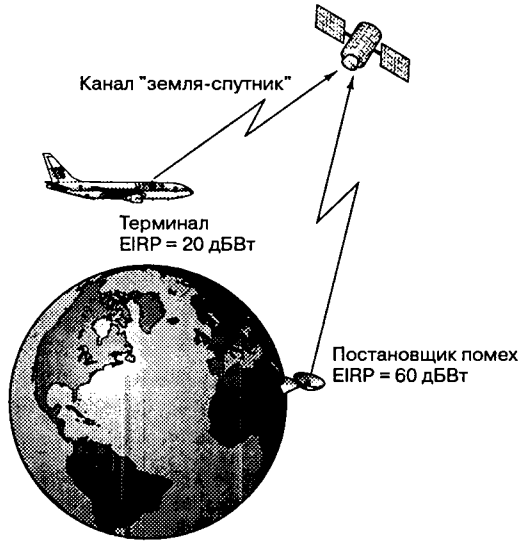


Рис. 12.26. Подавление спутникового канала связи

Решение

а) Большую опасность представляет подавление передачи на спутник, поскольку данная помеха может нарушить связь множества наземных терминалов, использующих спутниковый транспондер. Для достижения аналогичного результата при передаче со спутника пришлось бы создавать помехи для каждого из множества терминалов. Подавление сигналов со спутника может быть нежелательным при проведении определенных военных операций, однако состояние передачи на спутник намного важнее.

б) В соответствии с предположением, что потери мощности при распространении радиоволн одинаковы для устройства, находящегося на самолете, и станции преднамеренных помех, в уравнении (12.43) $(J/S)_{\text{прин}}$ можно заменить отношением мощности *переданных* помех и сигнала ($\text{EIRP}_J/\text{EIRP}_T$). Таким образом, можем записать следующее.

$$M_{\text{AJ}}(\text{дБ}) = (J/S)_{\text{треб}}(\text{дБ}) + \text{EIRP}_T(\text{дБВт}) - \text{EIRP}_J(\text{дБВт}) =$$

$$= G_p(\text{дБ}) - \left(\frac{E_b}{J_0}\right)_{\text{треб}}(\text{дБ}) + \text{EIRP}_T(\text{дБВт}) - \text{EIRP}_J(\text{дБВт})$$

$$G_p = 20\ \text{дБ} + 10\ \text{дБ} - 20\ \text{дБВт} + 60\ \text{дБВт} = 70\ \text{дБ}$$

$$W_{\text{ss}} = G_p(\text{дБ}) + R(\text{дБГц}) = 70\ \text{дБ} + 20\ \text{дБГц} =$$

$$= 90\ \text{дБГц} = 1\ \text{ГГц}$$

Пример 12.3. Подавление сигнала со спутника

В примере 12.2 предполагалось, что расстояние от спутника до самолета и станции преднамеренных помех одинаково. Однако следует учесть, что чем ближе будет находиться источник помех к приемнику, тем большим будет его негативное влияние. Рассмотрим сеанс связи "спутник-земля" при наличии помех. Эффективная изотропно-излучаемая мощность спутника и станции помех равна, соответственно, EIRP_S = 35 дБВт; EIRP_J = 60 дБВт. Потери мощности сигнала равны L_S = 200 дБ при передаче от спутника к приемнику и L_S' = 160 дБ при передаче от станции помех к приемнику. Каким должен быть коэффициент расширения спектра сигнала для закрытия канала с нулевым резервом против помех? Допустим, что $(E_b/J_0)_{\text{треб}}$ = 10 дБ.

Решение

При описанном подавлении сигнала со спутника расстояние от станции помех до самолета намного меньше, чем от спутника до самолета. Разница в расстоянии непосредственно влияет на пространственные потери мощности сигнала. Использовав уравнение (12.43), можно записать следующее.

$$M_{AJ}\,(\text{дБ}) = \left(\frac{J}{S}\right)_{\text{треб}}(\text{дБ}) - \left(\frac{J}{S}\right)_{\text{прин}}(\text{дБ}),$$

где

$$\left(\frac{J}{S}\right)_r(\text{дБ}) = \text{EIRP}_J\,(\text{дБВт}) - L_s'(\text{дБ}) - \text{EIRP}_s\,(\text{дБВт}) + L_s\,(\text{дБ}),$$

а также

$$\left(\frac{J}{S}\right)_{\text{треб}}(\text{дБ}) = \frac{W_{ss}}{R}(\text{дБ}) - \left(\frac{E_b}{J_0}\right)_{\text{треб}}(\text{дБ})$$

Найдя из записанных уравнений $G_p = W_{ss}/R$, получим следующее.

$$G_p = 75\ \text{дБ}$$

12.6.2. Подавление сигнала широкополосным шумом

Рассмотрим создание преднамеренных помех, которые могут быть смоделированы с помощью стационарного гауссова шума с нулевым средним и равномерным распределением спектральной плотности мощности (по крайней мере, в рассматриваемой области частот). Тогда при постоянной мощности полученного сигнала J спектральная плотность мощности сигнала помех J_0' равна J/W, где W — ширина полосы диапазона, в которой создаются помехи. Если генератор, используя всю свою мощность, создает помехи во всем диапазоне расширенного спектра W_{ss}, его называют *широкополосным постановщиком помех* (broadband jammer). Спектральная плотность мощности энергии такой станции равна следующему.

$$J_0 = \frac{J}{W_{ss}} \tag{12.44}$$

В главе 4 было показано, что вероятность битовой ошибки P_B для передачи сигналов BPSK с когерентной демодуляцией (без канального кодирования) равна следующему.

$$P_B = Q\left(\sqrt{\frac{2E_b}{N_0}}\right), \tag{12.45}$$

где функция $Q(x)$ определена в уравнениях (3.43) и (3.44). Табулированные значения данной функции приводятся в табл. Б.1. Однополосная спектральная плотность мощности шума N_0 соответствует тепловому шуму на входе РЕЙК-приемника. Из-за наличия умышленных помех полная спектральная плотность мощности увеличивается от N_0 до $(N_0 + J_0)$. Таким образом, средняя вероятность битовой ошибки в когерентной системе связи BPSK при наличии широкополосного шума равна следующему.

$$P_B = Q\left(\sqrt{\frac{2E_b}{N_0 + J_0}}\right) = Q\left[\sqrt{\frac{2E_b/N_0}{1 + (E_b/N_0)(J/S)/G_p}}\right] \tag{12.46}$$

Графики зависимости P_B от E_b/N_0 при заданном значении J/S приведены на рис. 12.27 [6, 21]. Кривизна графиков уменьшается по мере увеличения E_b/N_0. Это свидетельствует о том, что при заданном отношении мощностей сигнал/шум всегда будет существовать неснижаемая вероятность возникновения ошибки, вызванной наличием помех. Единственная возможность снизить эту вероятность состоит в увеличении коэффициента расширения спектра сигнала.

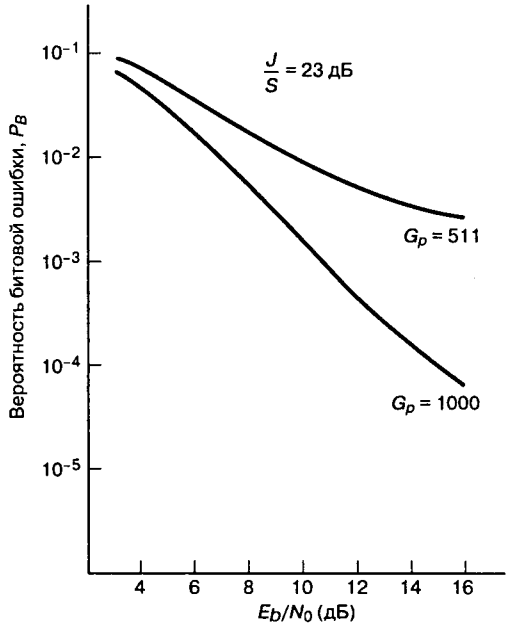


Рис. 12.27. Вероятность битовой ошибки в зависимости от E_b/N_0 при заданном значении J/S. (Перепечатано с разрешения авторов из Pickholtz R. L., Schilling D. L. and Milstein L. B. Theory of Spread-Spectrum Communications — A Tutorial, IEEE Trans. Commun., vol. COM30, n. 5, May, 1982, Fig. 11, p. 866 © 1982, *IEEE.)*

12.6.3. Подавление сигнала узкополосным шумом

Негативное влияние постановщика помех на систему связи со скачкообразной перестройкой частоты чаще всего может быть увеличено за счет использования *узкополосных помех.* Если для модуляции применяется двоичная частотная манипуляция с некогерентным обнаружением, вероятность битовой ошибки будет равна следующему (см. уравнение (4.96)).

$$P_B = \frac{1}{2}\exp\left(-\frac{E_b}{2N_0}\right) \tag{12.47}$$

Определим параметр ρ ($0 < \rho \leq 1$), указывающий часть полосы сигнала, в которой присутствуют помехи. Покрывая меньшую часть диапазона, генератор имеет возможность увеличивать в ней мощность помех. Например, покрывая полосу $W = \rho W_{ss}$, генератор

увеличивает спектральную плотность энергии шумов до уровня J_0/ρ. В таком случае средняя полученная мощность помех будет постоянной; она равна $J = J_0 W_{ss}$.

При подавлении связи узкополосными помехами вероятность корректного получения одного символа равна $(1-\rho)$. С другой стороны, при спектральной плотности мощности помех J_0/ρ вероятность подавления передачи одного символа равна ρ. Используя уравнение (12.47), можно выразить среднюю вероятность битовой ошибки в следующем виде.

$$P_B = \frac{1-\rho}{2}\exp\left(-\frac{E_b}{2N_0}\right) + \frac{\rho}{2}\exp\left[-\frac{E_b}{2(N_0 + J_0/\rho)}\right] \tag{12.48}$$

В большинстве случаев постановки преднамеренных помех справедливо предположение $J_0 >> N_0$. В результате, уравнение (12.48) упрощается до следующего вида.

$$P_B \approx \frac{\rho}{2}\exp\left(-\frac{\rho E_b}{2J_0}\right) \tag{12.49}$$

На рис. 12.28 представлены графики зависимости вероятности битовой ошибки от отношения E_b/J_0 при различных значениях ρ. Из рисунка видно, что для постановщика помех наиболее предпочтительно выбрать $\rho = \rho_0$, которое максимизирует P_B. Следует отметить, что ρ_0 уменьшается по мере возрастания E_b/J_0 (см. геометрическое место точек ρ_0 на рис. 12.28). Функция ρ_0 находится, если продифференцировать выражение (12.49) и приравнять $dP_B/d\rho$ к нулю. В результате это приводит к следующему.

$$\rho_0 = \begin{cases} \dfrac{2}{E_b/J_0} & \text{для } \dfrac{E_b}{J_0} > 2 \\ 1 & \text{для } \dfrac{E_b}{J_0} \le 2 \end{cases} \tag{12.50}$$

В данном случае максимальное значение P_B равно следующему.

$$(P_B)_{\max} = \begin{cases} \dfrac{e^{-1}}{E_b/J_0} & \text{для } \dfrac{E_b}{J_0} > 2 \\ \dfrac{1}{2}\exp\left(-\dfrac{E_b}{2J_0}\right) & \text{для } \dfrac{E_b}{J_0} \le 2 \end{cases}, \tag{12.51}$$

где e — основание натурального логарифма (2,71828...). Результат вычислений впечатляет. В наихудшем случае воздействие узкополосных помех на систему связи расширенного спектра *без использования кодирования* превращает экспоненциальную зависимость (12.49) в линейную (уравнение (12.51)). Геометрическое место точек ρ_0 на рис. 12.28 описывает отношение P_B к E_b/J_0 при максимально неблагоприятном воздействии узкополосных шумов на сигнал. Для значения вероятности битовой ошибки 10^{-6} разница между широкополосными и узкополосными помехами (в случае максимально неблагоприятного воздействия) составляет более 40 дБ при одинаковой мощности постановщика помех [4, 22]. Следовательно, негативное влияние на сигнал значительно выше при использовании узкополосных шумов по сравнению с широкополосными. Уменьшить это влияние можно с помощью метода прямого исправления ошибок (forward error correction — FEC) путем чередования [9]. Фактически для кодов с достаточно низкой интенсивностью метод FEC может

привести к тому, что постановщик узкополосных помех будет наносить максимальный вред только при работе в широкополосном режиме [23, 24].

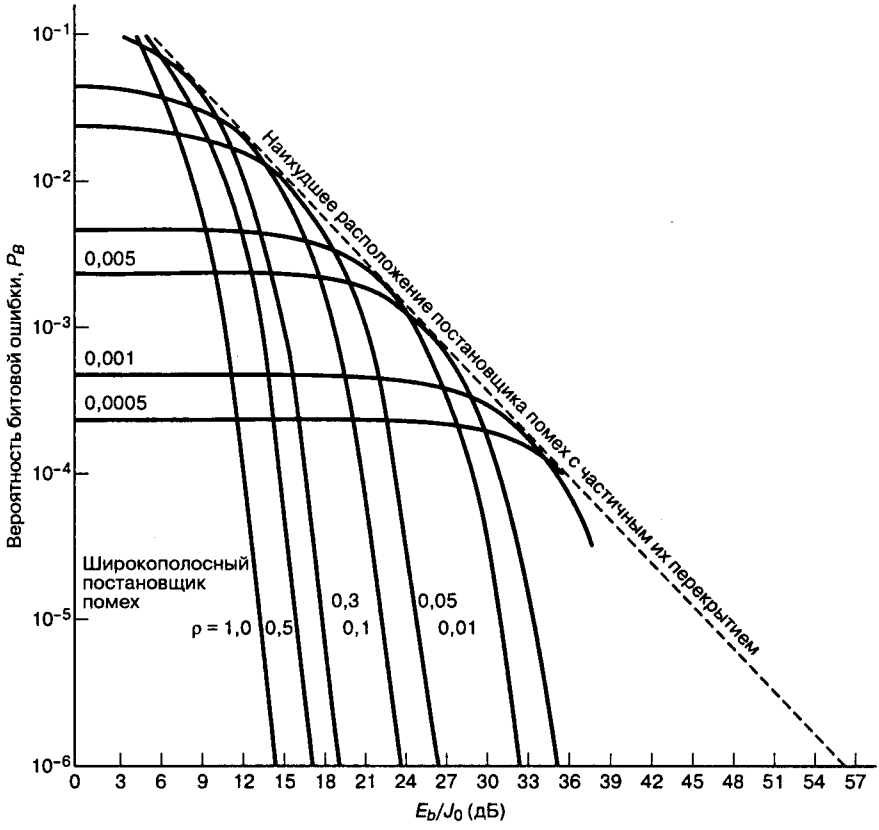


Рис. 12.28. Постановщик узкополосных помех (подавление сигнала FH/BFSK). (Перепечатано с разрешения издателя, Computer Science Press, Inc., 1803 Research Blvd., Rockville, MD., 20850, USA, из работы Simon M. K., Omura J. K., Scholtz R. A. and Levitt B. K., Spread Spectrum Communications, Vol. 1, Fig. 3.24, p. 173. © 1985.*)*

12.6.4. Подавление сигнала разнотонными помехами

При создании разнотонных помех станция-постановщик делит полную полученную мощность J между непрерывными тонами, имеющими случайную фазу и равными по мощности. Эти сигналы распределяются в диапазоне расширенного спектра W_{ss} в определенном порядке [9]. Анализ влияния тоновых помех на сигнал значительно сложнее, чем в случае шумов, в особенности для систем DS. Часто тоновые помехи рассматривают как гауссов шум. Хороший анализ системы DS при наличии разнотонных помех представлен в работе [25]. Производительность некогерентной системы связи FH/FSK считается одинаковой как при узкополосных тоновых помехах, так и при узкополосном шуме [26]. Однако применение узкополосных тоновых помех для подавления сигнала FH/FSK более эффективно. Причина в том, что использование непрерывных тоновых помех позволяет более эффективно ввести энергию в некогерентные детекторы [8]. Подробное описание производительности различных систем связи при наличии помех разного типа приводится в работах [8, 9, 26, 27].

Рассмотрим демодулятор FFH/MFSK, изображенный на рис. 12.16. Между каждым детектором огибающей и накопителем расположена схема одностороннего ограничения элементарных сигналов. Опишем работу схемы ограничения при воздействии на систему тоновых помех. На рис. 12.29 представлена восьмеричная схема FSK со скачкообразной перестройкой частоты и без разнесения сигнала (12.29, *а*), а также система с *быстрой* скачкообразной перестройкой частоты с использованием многократной ($N = 4$) передачи данных и ограничения элементарных сигналов (12.29, *б*). Обе части рисунка изображают состояние одного из $M = 8$ накопителей, представленных на рис. 12.16. Поступивший в накопитель сигнал обозначается вектором. Как видно из рис. 12.29, *а*, при отдельном скачке частоты полоса данных занята полученным символом с мощностью S. Если тоновая помеха с полученной мощностью J ($J \geq S$) случайно попадет в диапазон данных, детектор будет не в состоянии правильно определить полученный символ.

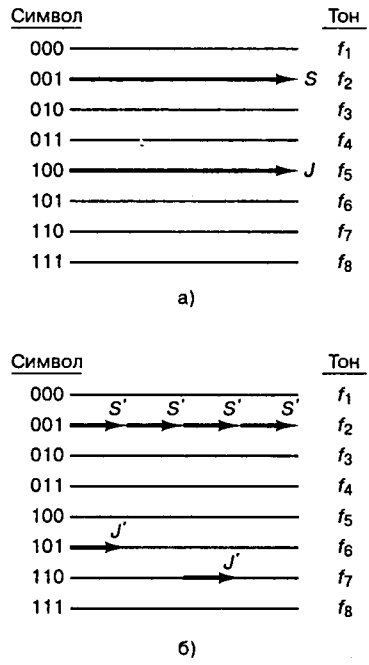


Рис. 12.29. Многократная передача символов с быстрыми скачками при наличии тоновых помех: а) отдельный скачок частоты; б) четыре скачка частоты

На рис. 12.29, *б* четыре элементарных сигнала (длина каждого вектора является мерой мощности ограниченного элементарного сигнала S') суммируются и полностью заполняют накопитель. Если тоновые помехи случайно попадут в спектральную область сигнала, это не повлияет на работу детектора, поскольку мощность помех ограничивается до одного уровня с элементарными сигналами связи ($J' = S'$). В примере, приведенном на рис. 12.29, *б*, два сигнала тоновых помех попадают в диапазон данных. Однако благодаря ограничению мощности никаких сомнений при определении полученного символа не возникает.

12.6.5. Подавление сигнала импульсными помехами

Рассмотрим работу системы связи DS/BPSK при подавлении сигнала импульсными помехами. Станция преднамеренных помех генерирует импульсы белого гауссова шума в узкой полосе частот. Средняя мощность шумов при получении равна J, хотя суммарная мощность генератора во время передачи импульса превышает это значение. Предположим, что генератор шумов может определить центральную частоту и полосу, которые используются для передачи данных. Допустим также, что мощность помех может быть увеличена за счет уменьшения времени передачи (другими словами, использовать часть $0 < \rho < 1$ полного времени передачи). Тогда в течение используемого времени спектральная плотность мощности постановщика возрастет до J_0/ρ, а усредненное по времени значение мощности J будет постоянным (где $J = J_0 W_{ss}$; W_{ss} — ширина полосы системы расширенного спектра).

Определение вероятности битовой ошибки для системы BPSK с когерентной демодуляцией и без канального кодирования было представлено в уравнении (12.45).

$$P_B = Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$$

Однополосная спектральная плотность мощности шума N_0 представляет тепловой шум на входе приемника. Из-за преднамеренных помех это значение возрастает до $(N_0 + J_0/\rho)$. Поскольку время передачи помех характеризуется коэффициентом ρ, средняя вероятность битовой ошибки равна следующему.

$$P_B = (1-\rho)Q\left(\sqrt{\frac{2E_b}{N_0}}\right) + \rho Q\left(\sqrt{\frac{2E_b}{N_0 + J_0/\rho}}\right) \tag{12.52}$$

При наличии преднамеренных помех значением N_0 можно пренебречь. Тогда выражение для P_B примет следующий вид.

$$P_B \approx \rho Q\left(\sqrt{\frac{2E_b\rho}{J_0}}\right) \tag{12.53}$$

Очевидно, что для генератора помех необходимо выбрать такое значение ρ, при котором P_B будет максимальным. На рис. 12.30 представлены кривые P_B для разных значений ρ. Аналогично созданию узкополосных помех, значение $\rho = \rho_0$, при котором P_B максимально, уменьшается по мере увеличения E_b/J_0. Продифференцировав уравнение (12.53), получим следующее.

$$\rho_0 = \begin{cases} \dfrac{0{,}709}{E_b/J_0} & \text{для } \dfrac{E_b}{J_0} > 0{,}709 \\ 1 & \text{для } \dfrac{E_b}{J_0} \le 0{,}709 \end{cases} \tag{12.54}$$

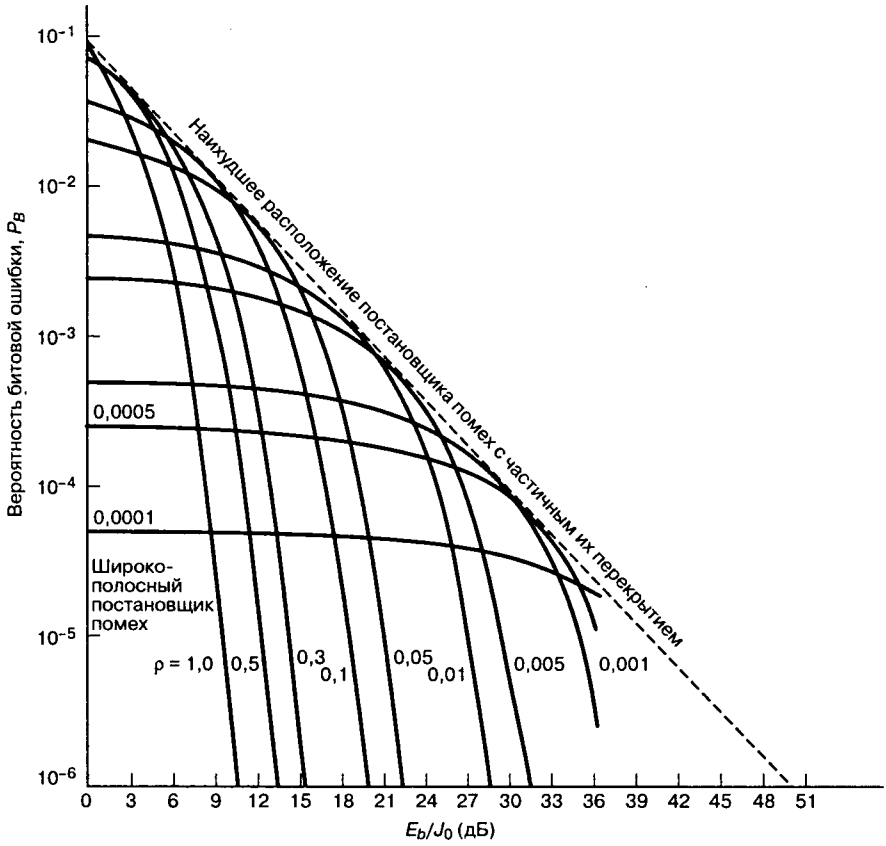


Рис. 12.30. Постановщик импульсных помех (подавление сигнала DS/BPSK). (Перепечатано с разрешения издателя, Computer Science Press, Inc., 1803 Research Blvd., Rockville, Md. 20850 USA, из работы Simon M. K., Omura J. K., Scholtz R. A. and Levitt B. K., Spread Spectrum Communications, Vol. 1, Fig. 3.7, p. 150 © 1985.*)*

Следовательно, максимальная вероятность битовой ошибки равна следующему.

$$(P_B)_{\max} = \begin{cases} \dfrac{0{,}083}{E_b/J_0} & \text{для } \dfrac{E_b}{J_0} > 0{,}709 \\ Q\left(\sqrt{\dfrac{2E_b}{J_0}}\right) & \text{для } \dfrac{E_b}{J_0} \le 0{,}709 \end{cases} \tag{12.55}$$

При максимально неблагоприятном воздействии помех на систему расширенного спектра *без использования кодирования* дополнительная функция ошибок (12.53) переходит в линейную зависимость (12.55). При вероятности ошибки 10^{-6} существует разница в 40 дБ между значениями E_b/J_0 для наиболее неблагоприятного постановщика импульсных помех и для постановщика широкополосных помех (рис. 12.30). Следовательно, негативное воздействие на систему DS/BPSK (без применения кодирования) при одинаковой выходной мощности будет значительно больше при использовании импульсных помех, чем в случае шумов постоянной мощности. Результат такого воздействия аналогичен влиянию узкополосных помех на систему связи FH/BFSK без

использования кодирования (см. раздел 12.6.3). В обоих случаях эффективное подавление сигнала достигается с помощью концентрации мощности генератора помех для "глушения" определенной части переданных символов. Кодирование с прямым исправлением ошибок и использованием чередования может практически полностью восстановить исходное качество сигнала [8, 23–25, 28].

12.6.6. Создание ретрансляционных помех

Вернемся к примерам 12.2 и 12.3, в которых рассматривался уровень устойчивости системы расширенного спектра со скачкообразной перестройкой частоты к широкополосному гауссову шуму. При определении уровня устойчивости не учитывалась скорость перестройки частоты. Интуитивно можно предположить, что чем чаще происходят скачки частот, тем проще "скрыть" сигнал от преднамеренных помех. Ведь если скорость изменения частоты не влияет на чувствительность к помехам, то почему же не применяются системы, в которых частота меняется один раз в день или раз в неделю? Ответ на этот вопрос скрывается в исходных предположениях, которые мы приняли в начале рассмотрения. В ходе вычисления коэффициента расширения спектра сигнала G_p предполагалось, что генератор помех не может предугадать положение сигнала в любой момент времени, имея в то же время информацию о ширине полосы расширенного спектра W_{ss}. Считалось, что скорость перестройки частоты *достаточно велика*, так что генератор помех не успевает проследить за процессом передачи и, соответственно, изменить свою тактику. При каких условиях это предположение может быть неверным? Кроме уже рассмотренных, существуют "интеллектуальные" постановщики помех, так называемые *постановщики ретрансляционных помех* (repeat-back jammer), способные проследить процесс передачи сигнала, что, как правило, делается с помощью бокового луча передающей антенны. Такие генераторы характеризуются высокой скоростью обработки сигнала, а также способностью приема сигналов в широкой области спектра. Это позволяет сконцентрировать мощность помех в непосредственной близости от сигнала системы FH/FSK. Преимущество постановщика помех такого типа перед широкополосным очевидно, поскольку помехи могут быть сконцентрированы в той полосе диапазона, которая используется для связи в каждый момент времени. Следует отметить, что такой метод подавления сигнала эффективен только по отношению к системам расширенного спектра со скачкообразной перестройкой частоты, поскольку в системах, использующих метод прямой последовательности, не существует мгновенного узкополосного сигнала, который можно было бы запеленговать.

Каким образом можно уменьшить негативное влияние постановщика ретрансляционных помех? Одним из возможных путей может быть увеличение скорости перестройки частоты до такой степени, чтобы в течение времени, нужного генератору помех для обработки полученного сигнала и создания помех, система перестраивалась на новую частоту. Естественно, в таком случае помехи не смогут повлиять на качество связи. Более подробно данный метод рассматривается в приведенном ниже примере.

Пример 12.4. Защита от постановщика ретрансляционных помех с помощью быстрой перестройки частоты

Предположим, что постановщик ретрансляционных помех расположен на расстоянии $d = 30$ км от наземной станции связи и способен обнаружить любой сигнал, передаваемый на спутник, который находится на небольшом расстоянии от обеих станций (рис. 12.31). Насколько быстро должна изменяться частота, используемая для передачи сигнала, чтобы избежать подавления сеанса связи? Допустим, что перестройка постановщика помех на вы-

бранную частоту происходит мгновенно. Время задержки сигнала постановщика помех относительно сигнала станции связи равно задержке распространения сигнала между станцией связи и постановщиком.

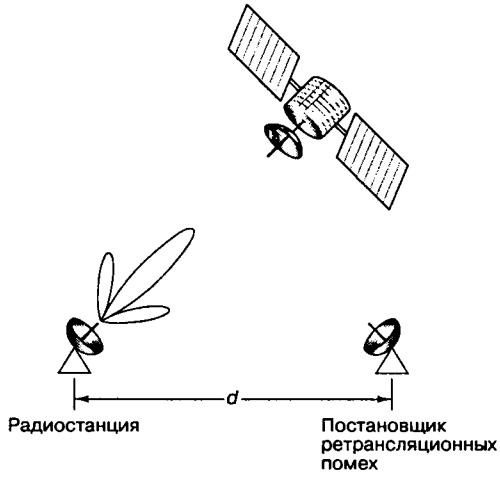


Рис. 12.31. Использование быстрой перестройки частоты для предотвращения подавления связи постановщиком ретрансляционных помех

Решение

Чтобы сигнал связи и помехи передавались в разное время, для интервала между двумя скачками частоты должно выполняться следующее условие:

$$T_{\text{hop}} \le \frac{d}{c} = \frac{3\times10^4\,\text{м}}{3\times10^8\,\text{м/с}} = 10^{-4}\,\text{с},$$

где c — скорость света. Тогда $R_{\text{hop}} \ge 10\,000$ скачков/с.

12.6.7. Система BLADES

Еще одна схема, позволяющая избежать подавления сигнала постановщиком ретрансляционных помех, была создана в середине 1950-х годов и получила название *BLADES* (Buffalo Laboratories Application of Digitally Exact Spectra). Перед передачей каждого бита генератор кода выбирает две частоты. *Окончательный выбор* частоты, которая будет использоваться, выполняется в зависимости от значения бита. На рис. 12.32 представлен типичный поток данных, состоящий из двоичных нулей и единиц, называемых *паузами* и *метками*. На рисунке также изображена последовательность пар частот (f_1 и f_1', f_2 и f_2' и т.д.). Для передачи метки выбирается частота f_i, для паузы — f_i'. Как видно из рисунка, поток данных преобразуется в последовательность тоновых сигналов f_1', f_2, f_3', f_4', f_5, ... В чем же преимущество такого метода передачи данных при постановке ретрансляционных помех? Постановщик помех обнаруживает передачу битов и создает помехи в спектральной области, близкой к частоте сигнала. Модуляция данных системой BLADES не имеет структуры в обычном понимании этого слова: с равной вероятностью сигнал может *или* присутствовать, *или* отсутствовать на определенной частоте. Поэтому помехи, создаваемые в спектральной области, близкой к частоте сигнала, не влияют на структуру данных. При некогерентной сис-

теме связи помехи только усиливают сигнал связи. Единственной возможностью для подавления связи остается создание широкополосных помех во всей области расширенного спектра.

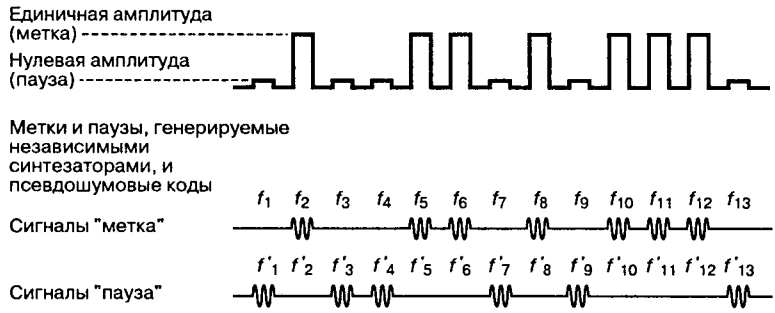


Рис. 12.32. Система BLADES

Следует отметить, что для передачи бита данных достаточно одной частоты. В таком случае для передачи двоичной единицы используется псевдослучайная частота, а передача нуля не производится. Приемник использует идентичный генератор кода для отслеживания псевдослучайной последовательности частот. Двоичная единица определяется при наличии сигнала на указанной частоте, двоичный нуль — при его отсутствии. Разумеется, данный метод менее устойчив к помехам, чем метод передачи пауз и меток с использованием двух независимо выбранных частот.

12.7. Использование систем связи расширенного спектра в коммерческих целях

12.7.1. Множественный доступ с кодовым разделением

Применение расширенного спектра в системах связи множественного доступа позволяет использовать одну частотную полосу для одновременной передачи нескольких сигналов без взаимной интерференции. В главе 11 использование расширенного спектра для задач множественного доступа рассматривалось на примере систем FH/CDMA. Данный раздел посвящен системам CDMA, использующим метод прямой последовательности (DS/CDMA). Итак, N пользователей получают индивидуальный код $g_i(t)$, где $i = 1, 2, ..., N$. Коды являются приблизительно ортогональными, так что взаимную корреляцию двух кодов считают приближенно равной нулю. Основное преимущество такой системы связи — возможность асинхронной передачи данных по всему диапазону различными пользователями. Другими словами, моменты переходов в символах различных пользователей не должны совпадать.

Блок-схема стандартной системы DS/CDMA приведена на рис. 12.33. Первый блок схемы соответствует модуляции данными несущей волны, $A\cos\omega_0(t)$. Выход модулятора, принадлежащего пользователю из группы 1, можно записать в следующем виде.

$$s_1(t) = A_1(t)\cos[\omega_0 t + \phi_1(t)] \tag{12.56}$$

Вид полученного сигнала может быть произвольным, поскольку процесс модуляции не ограничивается дополнительными требованиями.

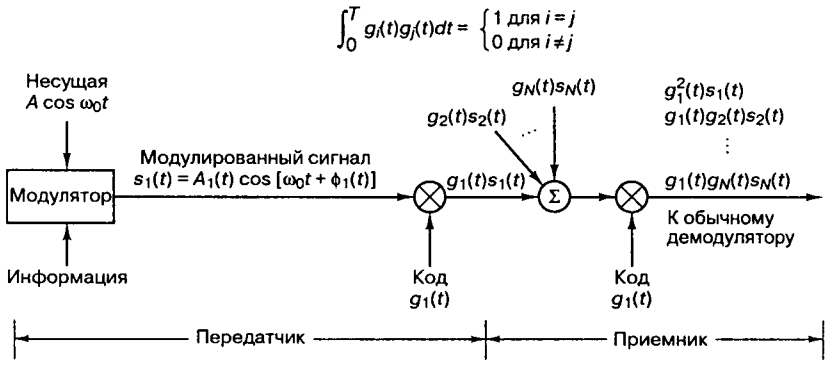


Рис. 12.33. Множественный доступ с кодовым разделением

Модулированный сигнал умножается на расширяющий сигнал $g_1(t)$, закрепленный за группой 1; результат $g_1(t)s_1(t)$ передается по каналу. Аналогичным образом для пользователей групп от 2 до N берется произведение кодовой функции и сигнала. Довольно часто доступ к коду ограничен четко определенной группой пользователей. Результирующий сигнал в канале является линейной комбинацией всех передаваемых сигналов. Пренебрегая задержками в передаче сигналов, указанную линейную комбинацию можно записать следующим образом.

$$g_1(t)s_1(t) + g_2(t)s_2(t) + \ldots + g_N(t)s_N(t) \tag{12.57}$$

Как указывалось ранее, умножение $s_1(t)$ на $g_1(t)$ дает в результате функцию, спектр которой является сверткой спектров $s_1(t)$ и $g_1(t)$. Поскольку сигнал $s_1(t)$ можно считать узкополосным (по сравнению с кодовым или расширяющим сигналом $g_1(t)$), полосы $g_1(t)s_1(t)$ и $g_1(t)$ можно считать приблизительно равными. Рассмотрим приемник, настроенный на получение сообщений от группы пользователей 1. Предположим, что полученный сигнал и код $g_1(t)$, сгенерированный приемником, полностью синхронизированы между собой. Первым шагом приемника будет умножение полученного сигнала в форме (12.57) на $g_1(t)$. В результате будет получена функция

$$g_1^2(t)s_1(t)$$

и набор побочных сигналов.

$$g_1(t)g_2(t)s_2(t) + g_1(t)g_3(t)s_3(t) + \ldots + g_1(t)g_N(t)s_N(t) \tag{12.58}$$

Подобно уравнению (12.14), если кодовые функции $\{g_i(t)\}$ взаимно ортогональны, полученный сигнал может быть идеально извлечен при отсутствии шумов, поскольку $\int_0^T g_i^2(t)dt = 1$. Побочные сигналы легко отсеиваются системой, так как $\int_0^T g_i(t)g_j(t)dt = 0$ при $i \neq j$. На практике кодовые функции не всегда идеально ортогональны между собой. Следовательно, взаимная корреляция кодов приводит к ухудшению качества связи и ограничивает максимальное число одновременно работающих пользователей.

Рассмотрим частотное представление приемника DS/CDMA. На рис. 12.34, *а* представлен широкополосный входной сигнал приемника, включающий в себя сигналы пользователей и побочные (нежелательные) сигналы. Каждый сигнал расширен от-

дельным кодом со скоростью передачи данных R_{ch} и характеризуется функцией спектральной плотности мощности вида $\text{sinc}^2(f/R_{ch})$. На графике также представлен полученный приемником тепловой шум, который равномерно распределен по всему диапазону. Суммарный сигнал, описанный выражением (12.58), поступает на вход корреляторa приемника, управляемого синхронизированной копией $g_1(t)$. На рис. 12.34, *б* представлен спектр, полученный после корреляции (сужения) с кодом $g_1(t)$. В дальнейшем пользовательский сигнал, расположенный в информационной полосе частот (центрированной на промежуточной частоте), обрабатывается обычным демодулятором, который должен иметь ширину полосы, достаточную для передачи расшифрованного сигнала. Побочные сигналы (см. уравнение (12.58)) не проходят процесс сужения спектра. Поэтому интерферировать с желаемым сигналом будут только сигналы, расположенные в его информационной полосе частот.

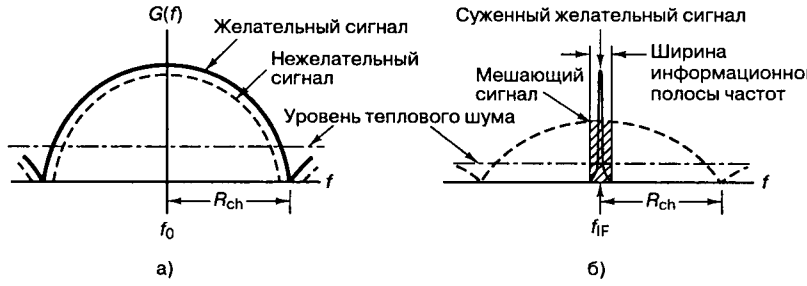


Рис. 12.34. Обнаружение сигнала расширенного спектра: а) спектр на входе приемника; б) спектр после корреляции с точным и синхронизированным псевдослучайным кодом

В работе [17] приводится превосходный анализ систем связи DS/SSMA с учетом корреляционных свойств кодовых последовательностей. В работах [18–20] анализируется производительность систем множественного доступа DS и FH при наличии интерференции.

12.7.2. Каналы с многолучевым распространением

Рассмотрим систему связи DS с двоичной фазовой манипуляцией при использовании канала, имеющего более одного маршрута распространения сигнала от передатчика к приемнику. Данный эффект может быть вызван отражением сигнала, преломлением его атмосферой либо отражением от зданий или других объектов. В итоге многолучевое распространение может вызывать флуктуации мощности сигнала на входе приемника. Маршрут прохождения сигнала может включать несколько дискретных траекторий, имеющих различные характеристики поглощения и времени задержки. На рис. 12.35 приводится пример двулучевого канала связи. Время задержки прямого сигнала по отношению к отклоненному равно τ. Подобное расхождение во времени может приводить к появлению "фантомных изображений" на экране телевизора, а в особо неблагоприятных случаях и к полной потере синхронизации изображения.

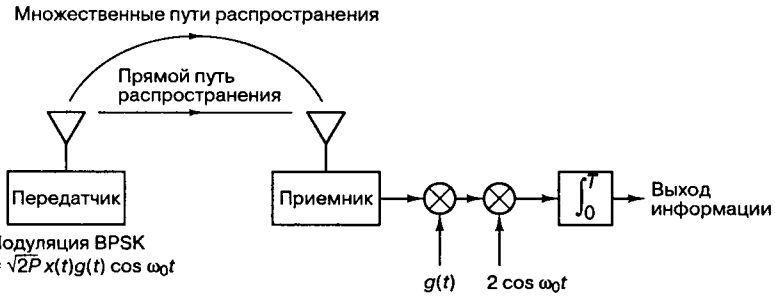


Рис. 12.35. Работа системы связи BPSK, использующей метод прямой последовательности, при многолучевом распространении сигнала

В случае системы связи расширенного спектра, в которой использован метод прямой последовательности, предположим, что приемник синхронизирован по времени задержки и фазе неотклоненного сигнала. Тогда полученный сигнал может быть выражен следующим образом.

$$r(t) = Ax(t)g(t)\cos\omega_0 t + \alpha Ax(t-\tau)g(t-\tau)\cos(\omega_0 t + \theta) + n(t) \qquad (12.59)$$

Здесь $x(t)$ — информационный сигнал, $g(t)$ — кодовый сигнал, $n(t)$ — гауссов процесс шума с нулевым средним, τ — разница во времени задержки для двух траекторий прохождения ($\theta < \tau < T$), θ — случайная фаза, равномерно распределенная в промежутке $(0, 2\pi)$, α — потери мощности многолучевого сигнала относительно прямого распространения. Для приемника, синхронизированного с прямым сигналом, выход коррелятора может быть представлен следующим образом.

$$z(t=T) = \int_0^T [Ax(t)g(t)\cos\omega_0 t + \\ + \alpha Ax(t-\tau)g(t)g(t-\tau)\cos(\omega_0 t + \theta) + n(t)g(t)]2\cos\omega_0 t\,dt \qquad (12.60)$$

где $g^2(t) = 1$. Для $\tau > T_c$, $g(t)g(t-\tau) \approx 0$ (для кодов с большими периодами), где T_c — длительность элементарного сигнала. Следовательно, если значение T_c меньше разницы во времени задержки между сигналами с разной траекторией распространения, можно записать следующее.

$$z(t=T) = \int_0^T [2Ax(t)\cos^2\omega_0 t + 2n(t)g(t)\cos\omega_0 t]dt = Ax(T) + n_0(T), \qquad (12.61)$$

где $n_0(T)$ — случайная гауссова переменная с нулевым средним. Таким образом, система связи с расширенным спектром (подобно системе CDMA) эффективно устраняет интерференцию, вызванную многолучевым распространением сигнала, с помощью приемника, скореллированного по коду.

Улучшить производительность системы связи при наличии многолучевого распространения сигнала можно и с помощью скачкообразной перестройки частоты. Быстрое изменение частоты позволяет приемникам избежать потерь мощности сигнала из-за многолучевого распространения. Поскольку рабочая частота приемника изменяется до того, как отклоненный сигнал поступает на вход, интерференция между двумя версиями сигнала невозможна.

12.7.3. Стандартизация систем связи расширенного спектра

В соответствии с требованиями Федеральной комиссии связи США (Federal Communications Commission — FCC), эксплуатация радиоустановок без приобретения лицензии допускается только для маломощного оборудования (мощностью ниже 1 мВт), за исключением некоторых частот ограниченного использования. В 1985 году сотрудник FCC, доктор Майкл Маркус (Michael Marcus), предложил разрешить применение систем радиосвязи расширенного спектра большей мощности (до 1 Вт) на частотах *ISM* (Industrial, Scientific and Medical — радиочастотные диапазоны для промышленного, научного и медицинского применения). Допустимые уровни электромагнитного излучения для устройств, не требующих лицензирования, определяются в томе 47, части 15 Свода федеральных постановлений США (Code of Federal Regulations — CFR). Для простоты их называют правилами "Part-15". Требования относительно систем расширенного спектра содержатся в разделе 15.247.

Частоты ISM могут использоваться по прямому назначению (например, оборудованием для диатермии) или же для правительственных нужд в экстренных случаях (к примеру, системами обнаружения). В обоих случаях используемое оборудование является источником мощных электромагнитных полей, которые могут интерферировать с обычными каналами связи. Частоты ISM чрезвычайно "зашумлены". Нелицензированное устройство радиосвязи может вызвать нежелательные эффекты для пользователя, имеющего лицензию. Необходимым требованием для указанных устройств является устойчивость к интерференции. В то же время создание помех для других пользователей запрещено.

В соответствии с правилами Part-15 среднее время использования частот для систем FH не должно превышать 0,4 с (скорость перестройки частоты должна быть не ниже 2,5 скачков/с). Для систем, использующих метод прямой последовательности, минимальное значение коэффициента расширения спектра сигнала должно составлять 10 дБ. Для смешанных систем связи, использующих одновременно метод прямой последовательности и метод перестройки частоты, это значение составляет 17 дБ. Для систем связи, которые не подлежат лицензированию, были выделены три спектральные области ISM. Некоторые параметры, связанные с использованием данных областей, приводятся в табл. 12.1.

Таблица 12.1. Требования к использованию систем связи расширенного спектра в соответствии с правилами Part-15

Полоса ISM	Полная ширина полосы	Максимальная ширина полосы на канал (FH)*	Минимальное количество скачков частоты на канал	Минимальная ширина полосы на канал (DS)*
902–928 МГц	26 МГц	500 кГц	25–50**	500 кГц
2,4000–2,4835 ГГц	83,5 МГц	1 МГц	75	500 кГц
5,7250–5,8500 ГГц	125 МГц	1 МГц	75	500 кГц

*Максимальная ширина полосы на канал для систем со скачкообразной перестройкой частоты равна 20 дБ. Минимальная ширина полосы на канал для системы, использующей метод прямой последовательности, равна 6 дБ.

**Каналам FH с шириной полосы менее 250 кГц требуется, по крайней мере, 50 скачков частоты на канал; каналам с шириной полосы более 250 кГц — как минимум, 25 частот.

В результате послабления требований относительно максимально допустимых уровней мощности, коммерческими компаниями было разработано множество устройств радиосвязи расширенного спектра. Данные устройства значительно превосходят по возможностям узкополосное радиооборудование низкой мощности, которое использовалось ранее. Среди новых коммерческих применений технологии расширенного спектра можно назвать устройства связи офисной техники (например, совместное использование принтера или создание беспроводных локальных сетей), телефонную радиосвязь, торговое оборудование (кассовые аппараты, сканеры штрих-кода).

12.7.4. Сравнительные характеристики систем DS и FH

Теоретически системы связи, использующие метод прямой последовательности (direct sequence — DS) и скачкообразную перестройку частоты (frequency hopping — FH), могут обладать равной производительностью (например, при полном отсутствии помех или в открытом пространстве). Для мобильных устройств связи со значительными задержками многолучевого распространения, метод прямой последовательности наиболее приемлем, так как все побочные версии сигнала, время отставания которых превышает время передачи элементарного сигнала, являются "невидимыми" для приемника (см. раздел 12.7.2). Системы FH могут быть эффективны в такой же степени, только если скорость перестройки частоты выше скорости передачи данных, а ширина используемой полосы достаточно велика (см. главу 15).

Использование системы радиосвязи со скоростной перестройкой частоты (fast frequency hopping — FFH) может быть связано со значительными материальными затратами (в основном, из-за необходимости применения высокоскоростных частотных синтезаторов). Скорость изменений частоты коммерческих систем FH, как правило, ниже скорости передачи данных, и поэтому такие системы связи обладают свойствами узкополосных радиоустройств. Отметим, что интерференция при использовании медленной перестройки частоты (slow frequency hopping — SFH) и метода прямой последовательности несколько отличается. Для устройств SFH характерно случайное появление мощных пакетов ошибок. При использовании DS появление помех более равномерно распределено во времени, причем шумы являются непрерывными и менее мощными по сравнению с устройствами SFH. При высокой скорости передачи данных негативное влияние многолучевого распространения сигнала более значительно для систем SFH. Для уменьшения этого влияния необходимо на протяжении длительного времени использовать чередование битов сигнала (см. главу 15). Сфера применения SFH ограничивается обеспечением разнесения в стационарных (или имеющих низкую скорость передвижения) системах радиосвязи. Кроме того, SFH может использоваться просто для удовлетворения стандарта Part-15. Создание радиосистем DS с большим значением коэффициента расширения спектра также может быть достаточно дорогостоящим (из-за применения высокоскоростных контуров). Чтобы избежать использования высокоскоростных контуров, значение коэффициента расширения обычно выбирают не более 20 дБ [29].

Пример 12.5. Обнаружение сигналов, скрытых шумами

В разделе 12.1.1.1 было показано, что расширение спектра не дает преимуществ при наличии тепловых шумов. В данном примере будет доказано, что любое значение E_b/N_0, доступное для узкополосной системы, остается неизменным после расширения спектра. Иными словами, применение расширенного спектра не дает определенных преимуществ при наличии тепловых шумов, однако и не ухудшает качество связи. Следовательно, расширение

спектра может быть использовано как для удовлетворения требований Part-15, так и для создания систем связи множественного доступа (например, систем CDMA, соответствующих стандарту IS-95).

Расширение спектра методом прямой последовательности позволяет обнаружить сигнал, уровень спектральной плотности мощности которого меньше аналогичного параметра шума. На рис. 12.36, *а* представлен график спектральной плотности мощности полученного сигнала с интенсивностью $S_0(f) = 10^{-5}$ Вт/Гц и шириной полосы 1 МГц. Поверхность, ограниченная графиком, представляет собой прямоугольник. Скорость передачи данных R будем считать равной 10^6 бит/с. Рассмотрим шум AWGN (изображен без соблюдения масштаба), который характеризуется спектральной плотностью мощности $N_0(f) = 10^{-6}$ Вт/Гц и присутствует на всех частотах диапазона. Требуется найти значение E_b/N_0 полученного сигнала для рассматриваемого случая узкой полосы частот. После этого рассмотрим расширение описанного выше сигнала (ширина полосы расширенного спектра $W_{ss} = 10^8$ Гц), как показано на рис. 12.36, *б*. При этом полная усредненная мощность сигнала не изменяется по сравнению со случаем узкой полосы. Докажите, что при использовании широкополосного приемника E_b/N_0 полученного сигнала не изменится по сравнению с узкополосным сигналом, а следовательно, не изменится и уровень возникновения ошибок.

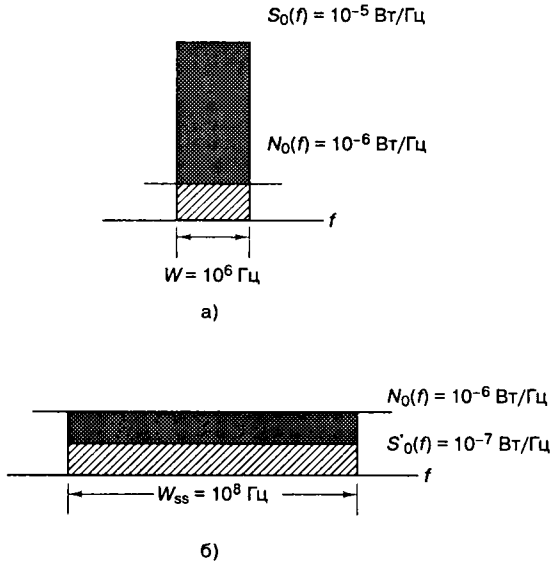


Рис. 12.36. Спектральная плотность мощности сигнала и шума: а) до расширения спектра; б) после расширения спектра

Решение

До расширения спектра полная усредненная мощность сигнала равна $S = 10^{-5}$ Вт/Гц × 10^6 Гц = 10 Вт. Определим полную среднюю мощность шума: $N = 10^{-6}$ Вт/Гц × 10^6 Гц = 1 Вт. E_b/N_0 полученного сигнала может быть записано в следующем виде.

$$\frac{E_b}{N_0} = \frac{S/R}{N_0} = \frac{10\text{Вт}/10^6\text{бит/с}}{10^{-6}\text{Вт/Гц}} = 10 \text{ или } 10 \text{ дБ}$$

После расширения спектра спектральная плотность мощности сигнала $S_0(f)$ уменьшается во столько же раз, во сколько возрастает ширина полосы (в данном случае, на 2 порядка). Следовательно, полная усредненная мощность сигнала после расширения по-прежнему равна 10 Вт.

Спектральная плотность мощности шума AWGN не снижается после расширения спектра. Полная усредненная мощность шума равна $N' = 10^{-6}$ Вт/Гц $\times$ 10^8 Гц = 100 Вт. Таким образом, E_b/N_0 полученного сигнала после расширения может быть выражено в следующем виде.

$$\frac{E_b}{N_0} = \frac{S/R}{N'/W_{ss}} = \frac{S}{N'}\left(\frac{W_{ss}}{R}\right) = \frac{S}{N'}G_p = \frac{10\,\text{Вт}}{100\,\text{Вт}} \times 100 = 10 \text{ или } 10 \text{ дБ},$$

где коэффициент расширения спектра сигнала $G_p = W_{ss}/R = 100$. Процесс обнаружения скрытых в шуме сигналов расширенного спектра с использованием прямой последовательности не позволяет привести интуитивно понятную иллюстрацию (рис. 12.36, *б*). Подобным образом в выражении для принятого E_b/N_0 после расширения спектра мощность сигнала связи равна 10 Вт, а мощность шума — 100 Вт, и снова интуитивно ничего нельзя сказать о возможности обнаружения сигнала. Значение E_b/N_0, аналогичное случаю с узкой полосой частот, позволяет получить коэффициент расширения спектра сигнала (который затруднительно представить визуально).

12.8. Сотовые системы связи

Беспроводные системы связи, в частности сотовые, используются для персональной связи сравнительно недолго. Наиболее важные моменты развития этой отрасли представлены ниже.

Годы

- 1921 Начало работы радиодиспетчерской полицейской службы в Детройте, штат Мичиган.
- 1934 Применение систем мобильной связи с использованием амплитудной модуляции (amplitude modulation — AM) сотрудниками государственной и муниципальной полиции США.
- 1946 Для абонентов коммутируемой телефонной сети общего пользования (public-switched telephone network — PSTN) стало возможным использование радиотелефонов.
- 1968 Начало разработок концепции сотовой связи в лабораториях корпорации Bell.
- 1981 Стандарт NMT (Nordic Mobile Telephone — северная мобильная связь), разработанный Ericsson Corporation для трех скандинавских стран, становится первой системой сотовой связи, работающей в реальных условиях.
- 1983 Корпорация Ameritech (Чикаго, США) начинает использование стандарта AMPS (Advanced Mobile Phone System — усовершенствованная система мобильной радиотелефонной связи) с применением частотной модуляции.
- 1990-е Во всем мире начинается использование цифровой сотовой связи второго поколения. Система GSM (Global System for Mobile — глобальная система мобильной связи) получает распространение по всей Европе. Множество различных стандартов, применяемых ранее, становятся непрактичными в использовании.
- 1990-е В США используются системы цифровой связи второго поколения IS-54, а также их модификации IS-136 (TDMA) и IS-95 (CDMA).
- 2000-е Международная стандартизация цифровых систем связи третьего поколения позволит сделать роуминг доступным практически во всем мире. Среди дополнительных преимуществ нового стандарта сотовой связи — возможность подключаться к разным системам PSTN, используя один телефон, а также доступ к системам высокоскоростной пакетной передачи данных (например, IP-сети).

12.8.1. CDMA/DS

На рис. 11.3 и 11.7 иллюстрируется совместное использование ресурса связи для схем FDMA и TDMA. При FDMA различные полосы частот являются взаимно ортогональными (предполагается идеальная фильтрация). Для TDMA взаимно ортогональными являются различные временные интервалы (предполагается идеальная синхронизация). Аналогичный случай ортогональности различных каналов для системы CDMA со скачкообразной перестройкой частоты представлен на рис. 11.14, причем подразумевается, что коды управления частотными скачками позволяют всем абонентам использовать разные временные интервалы и частоты. Графически несложно изобразить процесс передачи данных со скачкообразной перестройкой частоты и переключением временных интервалов при отсутствии конфликтных ситуаций. Однако при использовании системы расширения спектра методом прямой последовательности (direct-sequence spread-spectrum — DS/SS) графическое представление необходимых условий ортогональности для многих пользователей, одновременно работающих в одном спектре, будет нелегкой задачей. На рис. 12.37 представлены три различных сигнала DS/SS, расширенных по широкому диапазону частот, находящемуся ниже уровня мощности шумов и интерференции. Считается, что шумы и интерференция являются гауссовыми и широкополосными; их спектральная плотность мощности равна $N_0 + I_0$. В связи с примером, приведенным на рис. 12.37, наиболее часто возникает вопрос, как один из этих сигналов может быть обнаружен, если все они находятся по соседству в спектральной области и скрыты в шумах и помехах, вызванных интерференцией. Детектор DS/SS проверяет корреляцию полученного сигнала с псевдослучайным кодом определенного пользователя. Если псевдослучайные коды взаимно ортогональны, то в течение длительного времени приема средняя мощность всех сигналов других пользователей будет равна нулю. Если же условие взаимной ортогональности не выполняется, в процессе обнаружения будет происходить интерференция между сигналами разных пользователей.

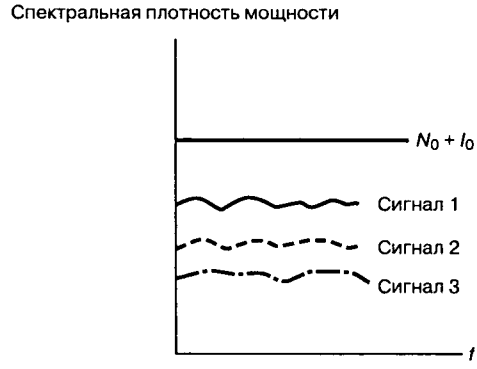


Рис. 12.37. Три сигнала DS/SS в одной спектральной области

В системе мобильной телефонной связи с использованием CDMA сигналы разных пользователей *интерферируют* между собой. Это происходит по следующим причинам.

1. Корреляция двух различных расширяющих кодов, принадлежащих одному семейству идеально ортогональных *длинных кодов*, может не равняться нулю в течение короткого времени, такого как длительность передачи одного символа.

2. Для обслуживания большого числа пользователей, как правило, необходимы длинные коды. При разработке таких кодов можно добиться малой взаимной корреляции, но при этом сложно получить идеальную взаимную ортогональность.
3. Многолучевое распространение сигнала и неидеальная синхронизация приводят к интерференции элементарных сигналов различных пользователей.

Рассмотрим канал обратной связи (от мобильного устройства к базовой станции), работающий в перегруженной сотовой ячейке. Интерференция в данном случае вызвана одновременным присутствием многих сигналов CDMA и превосходит по мощности помехи, вызванные тепловым шумом. Следовательно, влиянием тепловых шумов при наличии взаимной интерференции сигналов можно пренебречь. Тогда при $N_0 << I_0$ для отношения E_b/I_0 принятого сигнала, обозначенного как $(E_b/I_0)_{прин}$, можно записать следующее.

$$\left(\frac{E_b}{N_0 + I_0}\right)_{прин} \approx \left(\frac{E_b}{I_0}\right)_{прин} = \frac{S/R}{I/W_{ss}} = \frac{W_{ss}/R}{I/S} = \frac{G_p S}{I} \tag{12.62}$$

Здесь $G_p = W_{ss}/R$ — коэффициент расширения спектра сигнала, W_{ss} — ширина полосы расширенного спектра, S — полученная мощность сигнала одного из пользователей, I — мощность помех, вызванных интерференцией со всеми остальными пользователями. Из уравнения (12.62) следует, что даже если полученные помехи значительно превосходят по мощности сигнал пользователя, необходимую величину E_b/I_0 можно получить за счет коэффициента расширения спектра (посредством механизма проверки корреляции с кодом). Если базовая станция связи управляет мощностью сигнала и, следовательно, полученная мощность сигнала каждого из пользователей сбалансирована, то можно записать $I = S \times (M - 1)$, где M — полное число пользователей, вносящих вклад в интерференцию на входе приемника. Теперь можно выразить $(E_b/I_0)_{прин}$ через коэффициент расширения спектра и число активных пользователей в ячейке.

$$\left(\frac{E_b}{I_0}\right)_r \approx \frac{G_p S}{I} = \frac{G_p S}{S \times (M - 1)} = \frac{G_p}{M - 1} \tag{12.63}$$

Следует отметить, что $(E_b/I_0)_{треб}$ в уравнении (12.63) аналогично E_b/J_0 для приемника, получающего подавляемый сигнал в уравнении (12.41), причем J_0 и J соответствуют I_0 и I. Системы CDMA подвержены интерференции (шумы считают широкополосными и гауссовыми) независимо от того, чем она вызвана — преднамеренными помехами, случайными источниками сигналов или же самими пользователями. Будем считать, что G_p и необходимое значение E_b/I_0 (обозначим как $(E_b/I_0)_{треб}$) известны. Используя уравнение (12.63), можно записать максимально допустимое количество пользователей (источников интерферирующих сигналов) в сотовой ячейке для заданного уровня ошибок.

$$M_{max} \approx \frac{G_p}{(E_b/I_0)_{треб}} \tag{12.64}$$

Отметим, что уравнение (12.63) показывает, что для перегруженной ячейки интерференция накладывает ограничения на использование технологии CDMA. К примеру, если количество активных пользователей в ячейке внезапно возрастет вдвое, то полученное E_b/I_0 уменьшится в два раза. Аналогично из уравнения (12.63) следует, что уменьшение $(E_b/I_0)_{треб}$ позволяет увеличить максимально допустимое количество поль-

зователей. Ниже приводится список других факторов, от которых зависит число пользователей в ячейке.

- **Разделение по секторам или коэффициент усиления антенны** G_A. Ячейка может быть разделена на три сектора по 120° с помощью трех направленных антенн с коэффициентом усиления порядка 2,5 (или 4 дБ). Данный коэффициент определяет, во сколько раз может быть увеличено количество пользователей.
- **Фактор активности речи** G_V. В среднем в процессе разговора около 60% времени занимают паузы между словами и фразами, а также время слушания. Следовательно, для непосредственной передачи сигнала необходимо лишь 40% общего времени связи, т.е. время, когда один из собеседников говорит. Для каналов передачи речи данный факт позволяет увеличить количество пользователей в число раз, равное коэффициенту G_V, 2,5 (или 4 дБ).
- **Фактор интерференции от внешних ячеек** H_0. При технологии CDMA может применяться 100%-ное повторное использование частоты (см. раздел 12.8.2). Все соседние ячейки могут использовать один и тот же спектр. Тогда, кроме заданного уровня интерференции I_x, внутри ячейки существует дополнительная внешняя интерференция. Если потери сигнала описываются функцией четвертой степени (см. раздел 15.2.1), мощность внешней интерференции можно считать равной 55% от полной мощности интерференции внутри ячейки [30, 31]. Следовательно, полная интерференция может быть записана в виде 1,55 I_x. Число пользователей уменьшается в соответствии с коэффициентом H_0, который равен 1,55 (или 1,9 дБ).
- **Фактор несинхронной интерференции** γ. При оценке уровня интерференции пользователей, находящихся внутри и снаружи ячейки, было сделано предположение, что все используемые каналы идентичны (т.е. рабочие характеристики одинаковы для всех пользователей, передающих голосовые сигналы). Предположим также, что интерференция, связанная с сужением, может аппроксимироваться случайной гауссовой переменной. Будем считать, что пользователи равномерно распределены по площади ячейки, а управление мощностью в каждой из ячеек идеально. Наихудший случай — когда все интерферирующие между собой сигналы синхронизованы по фазе и элементарному сигналу. Для несинхронного канала связи ситуация будет лучше. В данном случае в уравнение (12.64) вводится коэффициент γ, описывающий интерференцию, вследствие чего максимально возможное количество пользователей увеличивается по сравнению с наихудшим сценарием. Если считать, что элементарный сигнал можно графически представить в виде идеального прямоугольника, значение γ равно 1,5 [31–34]. Вообще, данное значение зависит от формы функции, описывающей элементарный сигнал [31].

Используя коэффициенты G_A, G_V, H_0 и γ (а также их значения, приведенные выше), вычислим максимально возможное количество активных пользователей M' в ячейке.

$$M' = \frac{\gamma G_A G_V}{H_0} \times M_{\max} = \frac{\gamma G_p G_A G_V}{\left(E_b / I_0\right)_{\text{треб}} H_0} \approx 6 \times M_{\max} \qquad (12.65)$$

Точный расчет возможностей системы CDMA намного сложнее, чем приведенный в уравнении (12.65). При выводе данной формулы считалось, что пользователи равномерно распределены по площади ячейки, а управление мощностью осуществляется

идеально. В то же время влияние теплового шума считалось ничтожно малым. Изменения информационного обмена внутри ячейки не учитывались. Не рассматривалась топология местности как фактор, влияющий на параметр n функции потерь сигнала. При уменьшении n интерференция может возрастать. Вообще, емкость системы CDMA рассматривается во многих работах, в частности на примере систем, соответствующих стандарту IS-95. Для более подробного ознакомления с этой темой стоит обратиться к работам [30–32, 35–38]. В следующем разделе приводится упрощенный сравнительный анализ трех методов множественного доступа, позволяющий охарактеризовать преимущества CDMA.

12.8.2. Сравнительный анализ аналоговой частотной модуляции, TDMA и CDMA

До использования сотовых систем связи, в 1976 году в Нью-Йорке (население которого на то время составляло более 10 миллионов человек) мобильной связью могли одновременно пользоваться лишь 543 пользователя, в то время как всего их было 3700. Концепция сотовой связи иллюстрируется на рис. 12.38. В данном примере рассматривается конфигурация из семи ячеек (одна из используемых на данный момент). Благодаря разбиению географической области на ячейки с возможностью использования одних и тех же частот в разных ячейках, была значительно увеличена эффективность применения частотных полос в радиотелефонных системах связи.

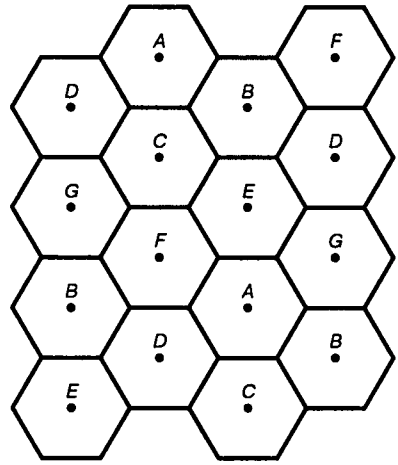


Рис. 12.38. Конфигурация из семи ячеек

В США частотный диапазон, используемый для передачи сигнала базовой станцией связи (869–894 МГц), принято называть *прямым* (forward), или *нисходящим* (downlink) каналом, а диапазон передачи данных мобильными устройствами (824–849 МГц) — *обратным* (reverse), или *восходящим* (uplink) каналом. Такая терминология используется для стандарта AMPS и других систем связи. Полосу, которую занимает один канал (30 кГц), иногда называют *поддиапазоном* (subband). Пара каналов, используемая для связи (прямой и обратный каналы), в сумме занимает 60 кГц и разделена полосой в 45 МГц. В пределах крупных городов США (всего около 750) Федеральная комиссия по средствам связи (FCC) выделила полосы по 25 МГц для передачи и приема сигналов. В целях поддержки конкуренции в пределах города обычно дается разрешение на ра-

боту двум компаниям. Каждая из них получает две полосы по 12,5 МГц — для приема и передачи сигналов.

Сравним количество доступных каналов в ячейке для трех сотовых систем связи (аналоговая FM, TDMA и CDMA) при широком географическом покрытии с множеством ячеек (рис. 12.38). Рассчитать количество аналоговых каналов, используемых в системе AMPS, можно довольно просто. Будем считать, что для связи выделена полоса в 12,5 МГц. Для предотвращения интерференции между пользователями, которые находятся в выделенном диапазоне 12,5 МГц и имеют приблизительно равную мощность, необходимо, чтобы в соседних ячейках использовались разные частоты. При конфигурации из семи ячеек (рис. 12.38) связь в ячейке *F* может осуществляться на полосе частот, которая отличается от диапазона ячеек *A*, *B*, *C*, *D*, *E* и *G*. Лишь одна седьмая часть полосы шириной 12,5 МГц может использоваться для связи в каждой ячейке. Следовательно, для каждой ячейки полоса шириной 1,78 МГц доступна для приема и передачи данных. При конфигурации из семи ячеек говорится, что *коэффициент повторного использования частоты* равен 1/7. Таким образом, при использовании аналоговой системы FM количество поддиапазонов шириной 30 кГц будет равно 1,78 МГц/30 кГц, или приблизительно 57 каналов в ячейке (без учета каналов, используемых для управления).

Североамериканский стандарт сотовой связи TDMA с использованием множественного доступа получил название IS-54 (последняя модификация этого стандарта — IS-136). Системы связи, соответствующие этим двум стандартам, должны удовлетворять требованиям использования частот, установленным для AMPS. Таким образом, ширина полосы канала TDMA равна 30 кГц. В 1950-х годах более эффективное применение кодирования исходного сигнала позволило увеличить количество используемых каналов. При *наземной* телефонной связи каждый голосовой сигнал кодируется со скоростью 64 Кбит/с. Возможно ли использование аналогичного стандарта для *сотовых* систем? Нет, поскольку сотовые системы связи ограничены шириной полосы. На данный момент кодирование голосовых сигналов позволяет достичь качества связи, аналогичного обычному телефонному разговору, при скорости передачи данных 8 Кбит/с. Даже при более низкой скорости этот метод позволяет получить приемлемое качество связи. Для вычислений значение скорости передачи данных принимается равным 10 Кбит/с. Сам процесс вычисления в этом случае достаточно прост. Одновременный доступ к каждому из каналов с шириной полосы 30 кГц может иметь 30 кГц/10 Кбит/с = 3 пользователя. Следовательно, количество пользователей, одновременно имеющих доступ к каналу в случае TDMA, в три раза больше, чем для аналоговой системы FM. Другими словами, количество каналов для каждой ячейки TDMA составляет $57 \times 3 = 171$.

Основным преимуществом систем CDMA по сравнению с аналоговыми FM или TDMA является возможность полного (100%) повторного использования частоты. Это значит, что вся ширина полосы, предусмотренная стандартом FCC (12,5 МГц), может одновременно использоваться для приема и передачи сигнала. Для сравнения CDMA, систем множественного доступа AMPS с использованием аналоговой частотной модуляции (другими словами, FDMA), а также TDMA стандарта IS-54 рассмотрим уравнение (12.65). Для корректности сравнения пренебрежем коэффициентом G_A, который характеризует разбиение ячейки на сектора. Данный коэффициент не используется в расчетах рабочих характеристик FDMA и TDMA, хотя в обоих случаях разбиение ячейки на сектора позволило бы улучшить параметры системы. Если ячейка не разбивается на сектора, количество активных пользователей в ячейке CDMA будет равно следующему.

$$M'' = \frac{\gamma G_p G_V}{(E_b/I_0)_{\text{треб}} H_0} \tag{12.66}$$

Из уравнения (12.28) получаем выражение для коэффициента расширения спектра сигнала.

$$G_p = \frac{R_{\text{ch}}}{R} = \frac{12{,}5 \text{ млн элементарных сигналов/с}}{10 \text{ Кбит/с}} = 1250 \tag{12.67}$$

Следует отметить, что такая скорость передачи (12,5 миллионов элементарных сигналов в секунду) не соответствует стандарту IS-95. В данном примере это значение используется для корректного сравнения CDMA, TDMA и аналоговой системы FM, имеющих ширину полосы 12,5 МГц.

Примем значение $(E_b/I_0)_{\text{треб}}$ равным 7 дБ (что аналогично умножению на 5) [30], а коэффициенты G_V, γ и H_0 равными 2,5, 1,5 и 1,55. Подставив указанные значения в уравнение (12.66), получим следующее.

$$M'' = \frac{1{,}5 \times 1250 \times 2{,}5}{5 \times 1{,}55} \approx 605 \tag{12.68}$$

Таким образом, системы FDMA с использованием аналоговой частотной модуляции, TDMA и CDMA могут поддерживать одновременное использование 57, 171 и 605 каналов в ячейке. Можно сказать, что при заданной ширине полосы CDMA превосходит AMPS по количеству активных пользователей приблизительно в 10 раз, а TDMA приблизительно в 3,5 раза. Следует отметить, что при выводе уравнения (12.68) не были учтены некоторые факторы (например, амплитудное замирание — см. главу 15), которые могут значительно уменьшить полученный результат. Следует также помнить, что анализ проводился для обратного канала CDMA, причем считалось, что применяются длинные коды, а сигналы пользователей не синхронизированы. В обратном направлении (канал-станция/мобильное устройство) может использоваться ортогональное распределение по каналам, что позволит улучшить результат (12.68).

Провести корректное сравнение CDMA и TDMA/FDMA достаточно сложно. При единичной ячейке рабочие характеристики TDMA/FDMA ограничиваются пространством, а параметры CDMA — интерференцией (см. следующий раздел). Если же используется множество ячеек, возможности всех указанных систем ограничиваются интерференцией. Улучшить отдельные характеристики каждой из систем можно следующим образом. Для TDMA/FDMA возможно повышение коэффициента повторного использования за счет увеличения интерференции. При использовании системы CDMA возможно увеличение нагрузки, но также за счет повышения интерференции.

12.8.3. Системы, ограниченные интерференцией и пространственными факторами

При правильном проектировании и эксплуатации системы CDMA интерференция в ней не играет значительной роли. Следовательно, весь рабочий спектр частот доступен для пользователей. Однако, исходя из уравнений (12.63) и (12.64), можно сказать, что интерференция накладывает определенные ограничения на системы CDMA. Использование кодирования с коррекцией ошибок чрезвычайно важно в случае CDMA, поскольку снижение значения $(E_b/I_0)_{\text{треб}}$ практически прямо сказывается на увеличении

допустимого числа активных пользователей. Увеличение эффективности кодирования на 1 дБ (что приводит к уменьшению отношения $(E_b/I_0)_{\text{треб}}$ на то же значение) позволяет повысить число активных пользователей ячейки CDMA на 25%.

При рассмотрении работы единичной ячейки, системы FDMA и TDMA можно назвать, соответственно, ограниченными частотным и временным диапазонами. Рассмотрим TDMA. В случае идеальной синхронизации распределения временных интервалов между растущим числом абонентов при получении сигнала базовой станцией, не происходит интерференции с сигналами других пользователей. Количество активных пользователей может увеличиваться до максимально возможного. Однако если все временные интервалы заполнены, увеличение числа активных пользователей приводит к чрезмерному возрастанию интерференции. Системы связи FDMA также являются ограниченными частотным диапазоном. Для таких систем увеличение количества пользователей после заполнения всех доступных полос влечет за собой чрезмерное возрастание интерференции.

Система CDMA — это система, ограниченная интерференцией, поскольку появление дополнительного пользователя ведет к увеличению общего уровня интерференции сигналов, принимаемых базовой станцией. Интерференция, вносимая отдельным мобильным радиоустройством, зависит от мощности, уровня синхронизации, а также от взаимной корреляции с другими сигналами CDMA. Допустимое количество каналов системы CDMA зависит от допустимого уровня интерференции. На рис. 12.39 представлено принципиальное различие между системами, возможности которых ограничиваются интерференцией (в данном случае CDMA) и пространством (TDMA). Предположим, что обе системы используют для связи полосу частот ограниченной ширины. В случае единичной ячейки при постепенном заполнении временных интервалов TDMA сигнал, поступающий на базовую станцию, не интерферирует с сигналами других мобильных радиоустройств. Количество активных пользователей TDMA может увеличиваться до полного заполнения всех доступных временных интервалов. После этого использование дополнительных интервалов приводит к возрастанию интерференции свыше допустимого уровня. Для систем CDMA при активизации каждого из пользователей уровень интерференции сигналов, получаемых базовой станцией, возрастает. Дополнительная интерференция, вносимая отдельным мобильным устройством, зависит от его мощности, синхронизации во времени, а также от взаимной корреляции с кодовыми сигналами других устройств. В пределах одной ячейки каналы предоставляются пользователям до достижения определенного предельного уровня интерференции [29]. Как видно из рис. 12.39, способность к адаптации системы, возможности которой ограничены интерференцией, значительно выше, чем в случае ограничений, связанных с пространственным фактором. К примеру, в праздничные дни, когда нагрузка телефонных сетей значительно возрастает, операционный центр системы CDMA может принять решение о повышении допустимого порога интерференции, чтобы увеличить количество активных пользователей. В случае системы, ограниченной пространством, такое просто невозможно.

Повторимся, пространственно-ограниченные системы (например, FDMA и TDMA) имеют жесткий порог производительности при применении одной ячейки. Если же используется множество ячеек, то путем изменения коэффициента повторного использования частот, а также отношения мощности сигнала к интерференции (S/I) можно добиться того, что указанные системы становятся ограниченными только интерференцией.

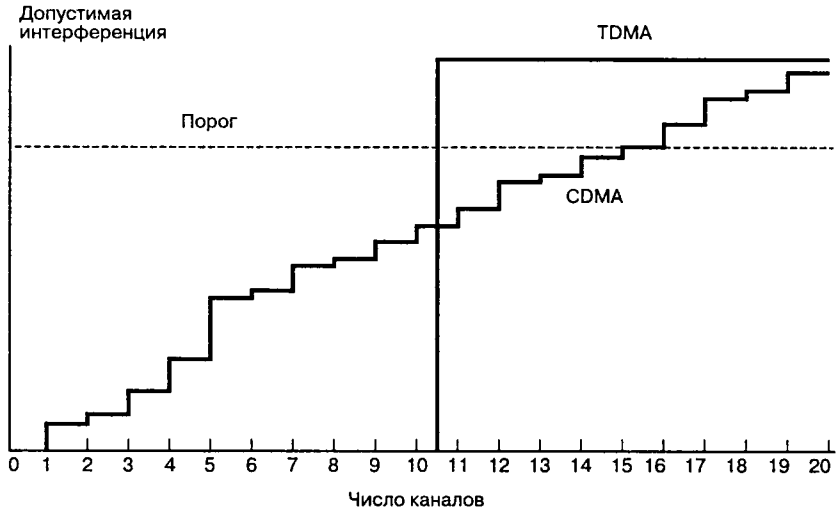


Рис. 12.39. Системы TDMA ограничены временной областью; возможности CDMA ограничены интерференцией

12.8.4. Цифровые сотовые системы связи CDMA стандарта IS-95

Interim Standard 95 (IS-95) определяет требования к радиотелефонным системам связи с применением сигналов расширенного спектра (метод прямой последовательности (DS/SS)) для обеспечения множественного доступа. Этот стандарт был разработан корпорацией Qualcomm для работы в спектре частот, используемом аналоговыми системами связи (AMPS) в США. Одновременная работа систем связи разных стандартов стала возможной благодаря технологии дуплексной передачи сигнала с использованием частотного разделения (frequency division duplexing — FDD). Системы AMPS используют полосу шириной 25 МГц для передачи сигнала от базовой станции к мобильному устройству (прямой канал) в диапазоне 869–894 МГц и полосу такой же ширины для обратной передачи сигнала (обратный канал) в диапазоне 824–849 МГц. При работе IS-95 в каждый отдельный момент времени используется система CDMA с шириной полосы 1,25 МГц, а мобильные устройства соответствуют одновременно двум стандартам (AMPS и CDMA). Возможности систем, соответствующих стандарту IS-95, ограничены интерференцией. Для снижения отношения $(E_b/N_0)_{\text{треб}}$ применяются различные методы обработки сигнала. Основные характеристики (форма сигнала, кодирование, методы подавления интерференции) рассматриваемых систем приводятся ниже.

- Каждый канал расширяется на полосу шириной 1,25 МГц, после чего фильтруется для ограничения спектра.
- Скорость передачи элементарных сигналов R_{ch} для псевдослучайного кода равна 1,2288 миллионов элементарных сигналов в секунду. Номинальная скорость передачи данных, называемая режимом RS1 (Rate Set 1), равна 9,6 Кбит/с и соответствует коэффициенту расширения $G_p = R_{\text{ch}}/R = 128$. В стандарте IS-95 возможно использование улучшенного скоростного режима RS2 (14,4 Кбит/с).
- Модуляция данных осуществляется с помощью двоичной фазовой манипуляции (BPSK) с применением расширения сигнала методом QPSK. При этом каждый

квадратурный компонент несущей является сигналом BPSK, модулированным данными.

- Используется сверточное кодирование с декодированием по алгоритму Витерби.
- Для разнесения по времени используется устройство временного уплотнения импульсных сигналов с интервалом 20 мс.
- Сигналы с многолучевым распространением обрабатываются RAKE-приемником. Для пространственного разделения используются две антенны в каждом секторе ячейки.
- Для разделения по каналам применяется ортогональное кодовое уплотнение.
- Регулирование мощности позволяет минимизировать энергию передаваемого сигнала и, следовательно, уменьшить интерференцию.

Передача сигнала от базовой станции к мобильному устройству может осуществляться с использованием четырех типов прямых каналов: контрольный, синхронизационный, поисковый и канал передачи данных. При обратной связи различают каналы доступа и передачи данных. Существует несколько модификаций стандарта IS-95: IS-95A, JSTD-008, IS-95B, IS-2000. IS-95B включает в себя использование сотовой полосы частот стандарта IS-95, а также полосы службы персональной связи (personal communication service — PCS). Этот стандарт позволяет передавать голосовые сигналы, а также данные со скоростью 115,2 Кбит/с при одновременном использовании до восьми каналов RS2. Стандарт IS-2000 описывает системы радиосвязи CDMA третьего поколения, также называемые системами с использованием множественных несущих. По сравнению с другими модификациями, IS-2000 имеет множество дополнительных возможностей. В данной главе рассматривается IS-95, структура которого сохраняется во всех последующих модификациях, поскольку все они построены на основе данного стандарта.

12.8.4.1. Прямой канал связи

Базовая станция использует 64 канала для передачи уплотненного сигнала. Для передачи данных пользователя применяется 61 канал. Один из каналов является контрольным, один — синхронизационным и, по крайней мере, один используется как поисковый. Стандарт IS-95 позволяет одновременную передачу голоса, данных и специальных сигналов. Скорость передачи голоса может быть равна 9600, 4800, 2400 или 1200 бит/с. Данные уровни скорости предусмотрены режимом RS1. В режиме RS2 поддерживается скорость до 14,4 Кбит/с. На рис. 12.40 представлена упрощенная блок-схема передатчика базовой станции, который использует стандартный канал данных со скоростью передачи 9,6 Кбит/с. С помощью кодирования методом линейного предсказания (linear predictive coding — LPC, см. раздел 13.4.2) производится черновая оцифровка голосового сигнала со скоростью 8 Кбит/с. После добавления битов обнаружения ошибок скорость передачи возрастает до 9,6 Кбит/с. Полученная последовательность данных разбивается на кадры длительностью 20 мс. Следовательно, при скорости передачи данных 9,6 Кбит/с один кадр содержит 192 бит. Следующий шаг, представленный на рис. 12.40, — сверточное кодирование (степень кодирования 1/2, $K = 9$), в ходе которого все биты данных в равной мере защищаются кодом. В результате скорость в канале возрастает до 19,2 Кбит/с и остается неизменной после обработки данных устройством временного уплотнения импульсных сигналов с рабочим интервалом, равным длительности кадра (20 мс). Следующие три шага включают

сложение по модулю 2 двоичных значений псевдослучайных кодов и ортогональных последовательностей (применяется для обеспечения конфиденциальности); распределение по каналам; и определение базовой станции. Каждое изменение кода можно образно представить как барьер, ограничивающий по тем или иным причинам доступ к определенному сообщению. В целях конфиденциальности используются псевдослучайные коды максимальной длины с 42-разрядным регистром сдвига. В системе со скоростью передачи 1,2288 миллионов элементарных сигналов в секунду такой код повторяется с периодом приблизительно в 41 день. Системы, соответствующие стандарту IS-95, используют идентичное оборудование для кодирования для всех базовых станций и мобильных устройств. В целях конфиденциальности каждое мобильное устройство получает уникальную модификацию кода со сдвигом по фазе или во времени. Пользователям, которые связываются между собой, не нужно знать кодовые модификации друг друга, поскольку базовая станция производит демодуляцию и повторную модуляцию всех обрабатываемых сигналов. Значение скорости передачи данных в канале (19,2 Кбит/с) перед кодированием не является окончательным. Код применяется для прореживания сигнала, поэтому используется только каждый 64-й бит последовательности (что не влияет на уникальность кода).

Следующий применяемый код называют *защитой Уолша* (Walsh cover). Данный код используется для распределения по каналам с последующим расширением спектра. Код является ортогональным и генерируется с помощью матрицы Адамара (Hadamard matrix) (правила получения кода приводятся в разделе 6.1.3.1). Используя указанный метод, можно создать код Уолша, размерность которого равна $2^k \times 2^k$ (k — положительное целое число). Набор кодов Уолша характеризуется матрицей 64 × 64, где каждая строка соответствует отдельному коду. Как показано на рис. 12.40, один из 64 кодов суммируется по модулю 2 с защищаемой двоичной последовательностью. Поскольку элементы набора кодов Уолша взаимно ортогональны, их применение позволяет разделить прямой канал связи на 64 ортогональных сигнала. Канал 0 используется для проверки когерентности получения данных мобильным устройством. Канал 32 применяется для синхронизации, а также, по крайней мере, один канал резервируется в качестве поискового. Следовательно, для передачи данных доступен 61 канал. Защита Уолша применяется в системах со скоростью передачи 1,2288 миллионов элементарных сигналов в секунду. Таким образом, в процессе связи "базовая станция-мобильное устройство" каждый бит в канале (скорость передачи 19,2 Кбит/с) преобразуется в 64 элементарных сигнала Уолша. Конечная скорость передачи составляет 1,2288 миллионов элементарных сигналов в секунду. На рис. 12.41 представлена последовательность из 64 сигналов Уолша, а на рис. 12.42 приведен простой пример распределения по каналам с использованием ортогональных кодов (к примеру, кодов Уолша). Выходной сигнал будет отличным от нуля только в том случае, если приемник использует правильный код для доступа к каналу пользователя. Применение правильного кода дает на выходе некоторое ненулевое значение A, которое позволяет "открыть дверь" канала.

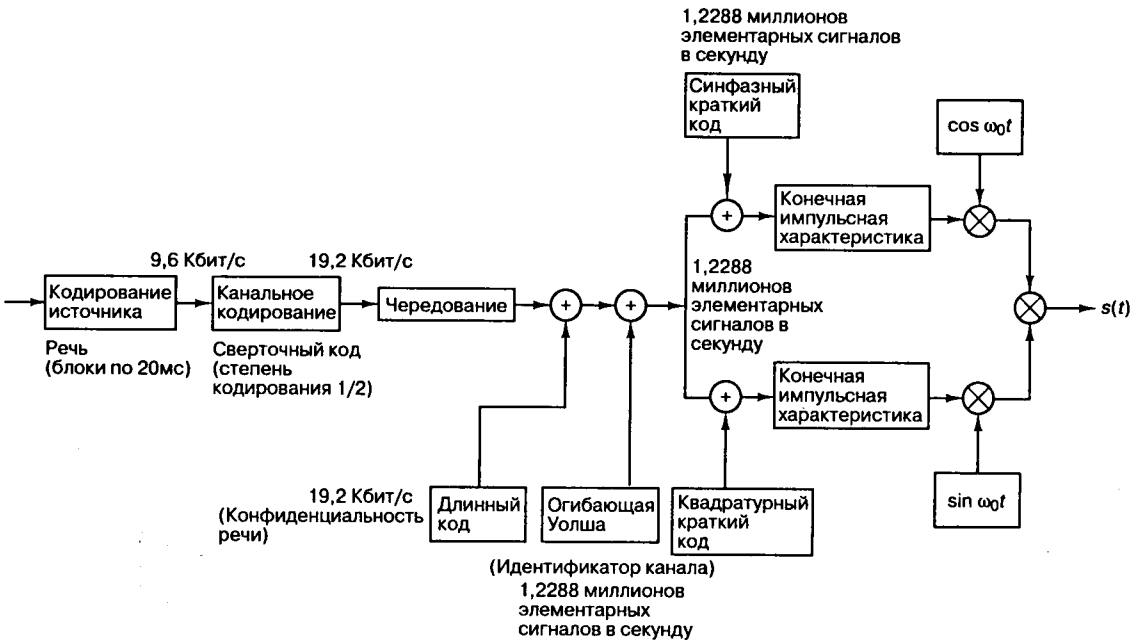


Рис. 12.40. Передача голоса с использованием прямого канала CDMA

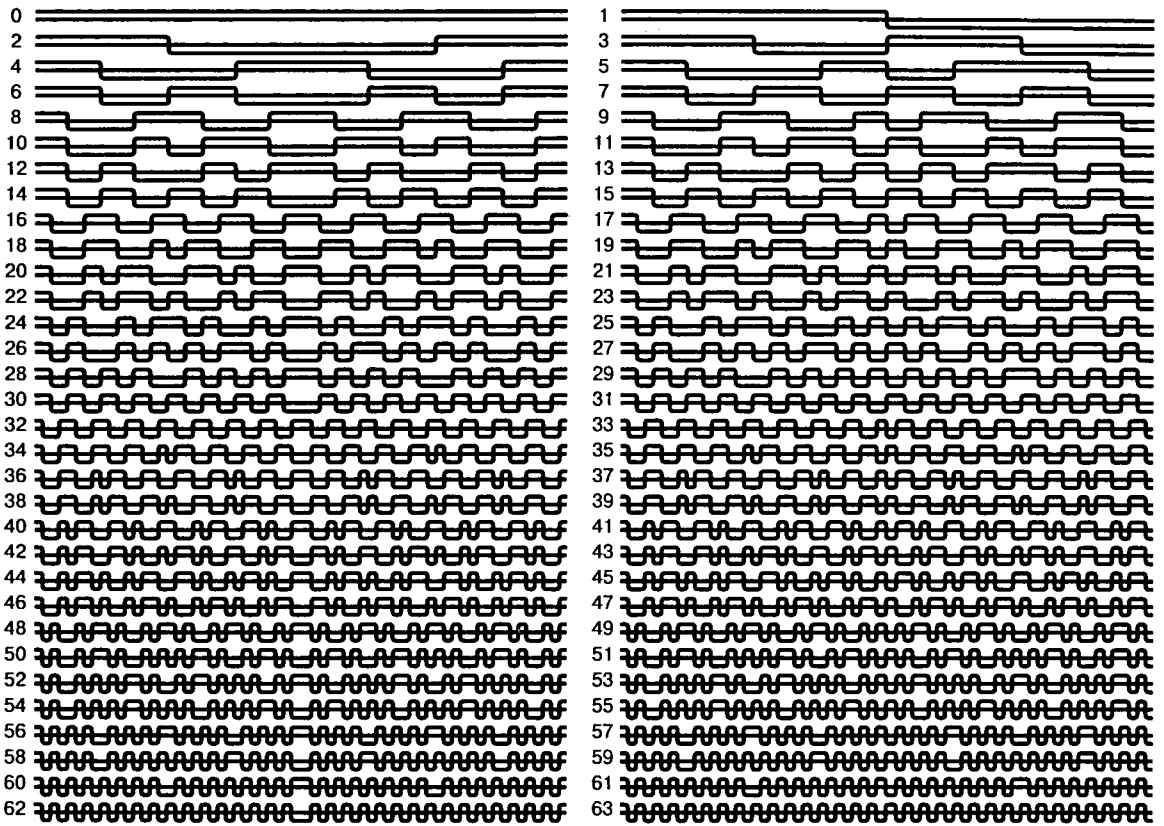


Рис. 12.41. Последовательность из 64 сигналов Уолша

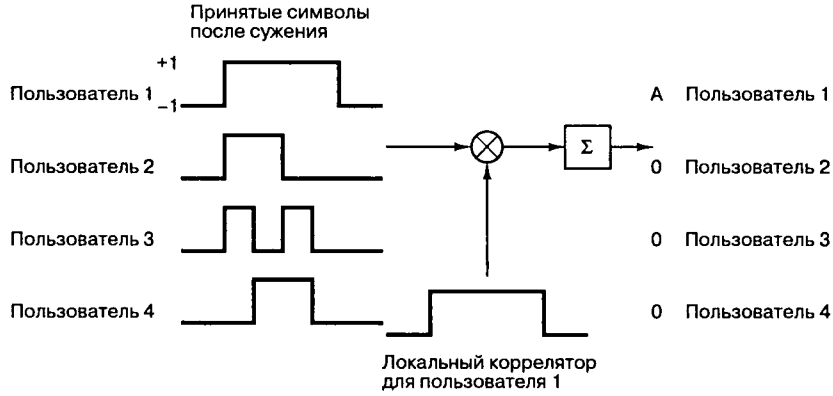


Рис. 12.42. Пример передачи сигнала с применением ортогональных функций для распределения по каналам

Следующий применяемый код (рис. 12.40) называют *коротким* (short), поскольку он основывается на 15-разрядном регистре сдвига и повторяется с интервалом $2^{15} - 1$ элементарных сигналов (один период длится 26,67 мс). Этот последний "барьер", используемый со сдвигом по фазе 90° со скоростью передачи 1,2288 миллионов элементарных сигналов в секунду, позволяет зашифровать сигнал. Поскольку все базовые станции используют идентичное распределение по каналам методом Уолша, при отсутствии шифрования их сигналы коррелировали бы между собой (что нежелательно). Короткий код можно представить в качестве "адреса" базовой станции. Использование этого кода требует наличия двух регистров сдвига: одного в синфазном канале(I), другого в квадратурном (Q). Каждая базовая станция для определения своего местоположения применяет особую модификацию (сдвиг) кодов I и Q, состоящих из 64 элементарных сигналов. Таким образом, использование данных кодов позволяет получить 512 уникальных адресов. Это число достаточно велико, поскольку станции, находящиеся достаточно далеко друг от друга, могут использовать одинаковые адреса.

Таким образом, код Уолша позволяет ортогонализировать сигналы пользователей, находящихся в одной ячейке; короткий псевдослучайный код делает сигналы пользователей разных ячеек независимыми друг от друга (присваивает адрес каждой станции); длинный псевдослучайный код позволяет сделать сигналы разных пользователей системы взаимно независимыми (используется для конфиденциальности связи). Чтобы после применения кода Уолша ортогональность каналов была идеальной, работа всех пользователей должна быть синхронизирована во времени с погрешностью, не превышающей малой доли длительности элементарного сигнала. Для прямого канала это теоретически возможно, поскольку передача сигнала на мобильные устройства производится базовой станцией. Однако, учитывая эффект многолучевого распространения, более корректно будет сказать, что применение кодов Уолша позволяет достичь неполной ортогональности. Для получения аналогичного результата в случае обратного канала необходимо использовать временную синхронизацию с обратной связью, что не предусмотрено в IS-95. Уменьшить сложность системы можно за счет увеличения интерференции

внутри ячейки. В широкополосном стандарте CDMA третьего поколения (wideband CDMA — WCDMA) такая возможность предусмотрена [39].

Последние шаги, представленные на рис. 12.40, соответствуют широкополосному (1,25 МГц) фильтрованию на фильтре с конечной импульсной характеристикой, а также смешиванию несущей с использованием расширения QPSK и модуляции BPSK. Идентичные кодированные сигналы одновременно присутствуют в синфазном и квадратурном каналах, однако из-за шифрования коротким кодом эти сигналы отличаются друг от друга.

12.8.4.2. Обратный канал связи

Уплотненный сигнал, передаваемый каждой базовой станцией, состоит из 64 каналов, причем для передачи данных могут использоваться лишь 61 из них (или меньше). Однако при связи в обратном направлении (мобильное устройство-базовая станция) передается единичный сигнал (канал), который может содержать данные или запрос на доступ к сети. На рис. 12.43 представлена упрощенная блок-схема передачи сигнала с использованием обратного канала. Общая структура аналогична существующей в прямом канале (рис. 12.40), однако существуют некоторые существенные отличия. В стандарте IS-95 не поддерживается применение обратных управляющих каналов, поскольку для каждого мобильного устройства был бы необходим отдельный канал такого типа. В стандарте IS-2000 для каждого обратного канала связи резервируется управляющий канал. Поскольку обратный канал менее устойчив по сравнению с прямым, для улучшения работы системы применяется более эффективный сверточный код (степень кодирования 1/3). Следует также отметить, что после обработки устройством временного уплотнения импульсных сигналов биты канала модулируют 64-разрядный код Уолша. Этот код аналогичен используемому для распределения по каналам при передаче по прямому каналу. Однако при обратной связи коды Уолша используются для прямо противоположной цели — они играют роль модулирующих волн. При скорости передачи данных, равной 9,6 Кбит/с, два бита данных (после кодирования трансформируются в шесть канальных битов, иногда называемых кодовыми символами) после разделения отображаются одним из 64 ортогональных сигналов Уолша, который впоследствии и передается. Скорость передачи символов Уолша равна следующему.

$$R_w = \frac{R_c}{\log_2 M} = \frac{R(n/k)}{\log_2 M} = \frac{9600 \times 3}{6} = 4800 \text{ символов Уолша / с} \qquad (12.69)$$

Здесь скорость передачи канальных битов R_c равна произведению скорости передачи данных и обратной интенсивности кода, или $R(n/k)$. Каждый из 64-разрядных кодов Уолша состоит из 64 элементов, называемых *элементарными сигналами Уолша*. Исходя из уравнения (12.69), скорость передачи элементарных сигналов Уолша составляет 64 × 4800 = 307 200 сигналов/с. Следовательно, результатом модуляции является расширение спектра (однако не до полной ширины полосы). Элементарные сигналы Уолша повторяются 4 раза, и окончательная скорость передачи данных составляет 1,2288 миллионов элементарных сигналов в секунду.

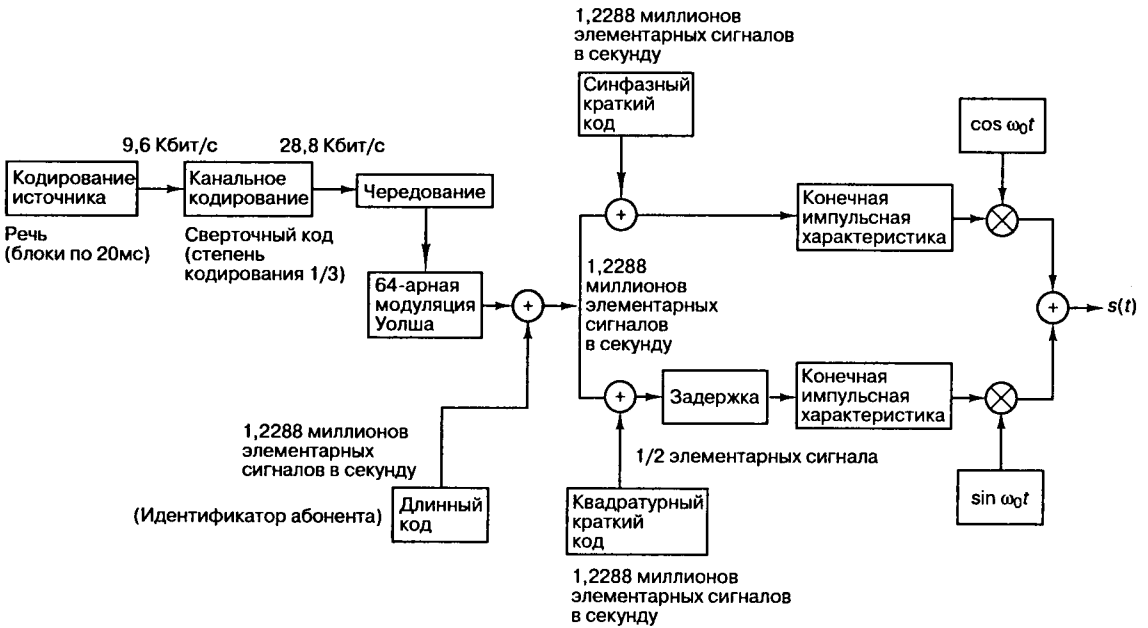


Рис. 12.43. Передача голоса с использованием обратного канала CDMA

Может возникнуть вопрос, почему в качестве модулирующих волн были выбраны 64-ричные функции Уолша. Вспомним компромиссы между параметрами каналов, ограниченных по мощности (раздел 9.7.3). Для сохранения мощности за счет уменьшения ширины полосы было бы логично использовать M-арную частотную манипуляцию, например, MFSK. По мере возрастания M ширина полосы будет увеличиваться и одновременно будет снижаться отношение E_b/N_0, необходимое для получения заданного уровня достоверности передачи. Использование подобного метода передачи сигнала для узкополосной системы является компромиссным решением, поскольку снижение необходимого уровня мощности достигается за счет увеличения ширины полосы. Однако для систем расширенного спектра, соответствующих стандарту IS-95, применение 64-ричных функций Уолша для модуляции можно описать как "бесплатное приобретение", поскольку система уже использует расширенную полосу в 1,25 МГц. Применение 64-ричных ортогональных функций не приводит к дальнейшему расширению полосы. Если представить, что форма импульсов на графике функций Уолша (рис. 12.41) несколько округлена, то не напомнило бы вам это форму сигналов MFSK? Да, графики этих двух функций весьма похожи. В общем случае базовая станция обнаруживает 64-ричные функции Уолша некогерентно, что аналогично обнаружению 64-ричных тонов FSK. Нужно отметить, что некоторые типы приемников базовых станций используют когерентный метод обнаружения, что позволяет выиграть 1–2 дБ.

Для обратной связи необходимо распределение по каналам, поскольку пользователи должны быть разделены. При использовании обратного канала пользователи отличаются друг от друга длинным кодом (кодом конфиденциальности). В прямом канале связи этот код применяется для прореживания сигнала, что позволяет обеспечить конфиденциальность. При связи мобильное устройство-базовая станция (рис. 12.43) код используется со скоростью 1,2288 миллионов элементарных сигналов в секунду для распределения по каналам (адресации), а также для шифрования сигнала, достижения конфиденциальности и расширения спектра. После расширения длинным кодом, спектр сигнала расширяется еще раз с помощью двух коротких псевдослучайных кодов, что обеспечивает отсутствие корреляции между синфазными и квадратурными символами. Последние шаги, приведенные на рис. 12.43, соответствуют фильтрованию на фильтре с конечной импульсной характеристикой, а также преобразованию несущей с помощью модуляции BPSK в сигнал OQPSK. Модуляция OQPSK применяется, чтобы избежать возможности изменения фазы несущей на 180° (см. раздел 9.8.1). Этот метод позволяет уменьшить соотношение пиковой и средней мощности усилителя передатчика, что упрощает проектирование системы. OQPSK не применяется для прямых каналов, поскольку в этом случае базовая станция передает уплотненный сигнал 64 каналов. Каждый процесс прямой передачи может быть описан вектором, который характеризует весь уплотненный сигнал. Вектор принимает значение из множества возможных соотношений фаза/амплитуда. Следовательно, посредством сдвига синфазного и квадратурного каналов невозможно добиться положительного результата, поскольку невозможно избежать переходов несущей через нуль. После фильтрования полученного сигнала образуется спектр с двухсторонней шириной полосы по уровню 3 дБ, равной 1,25 МГц.

12.8.4.3. Типы приемников

Приемник мобильного устройства. Данный приемник когерентно демодулирует сигналы QBPSK прямого канала, используя контрольный сигнал в качестве эталона. Схема приемника включает трехкомпонентный RAKE-приемник, который позволяет расшифровать три наиболее сильных компонента многолучевого сигнала (минимальное требование IS-95). RAKE-приемник разрешает и разделяет многолучевые компоненты сигнала расширенного спектра при условии, что разница во времени распространения между отдельными лучами больше длительности одного элементарного сигнала. Сигналы FDMA не могут быть разделены подобным образом, поскольку они по определению являются узкополосными. Многолучевые компоненты сигнала TDMA можно разделить, поскольку пользователи передают данные в виде пакетов. Однако при заданном времени задержки полосы сигналов-пакетов стандартной системы TDMA недостаточно широки для разрешения многолучевого сигнала. При использовании CDMA ширина полосы превышает 1 МГц и любые многолучевые компоненты, характеризующиеся временем задержки более 1 мкс, могут быть разрешены. RAKE-приемник быстро отслеживает многолучевые компоненты и эффективно сочетает их (в случае приемника мобильного устройства — когерентно). Принцип работы RAKE-приемника описывается в разделе 15.7.2. Выходные сигналы демодулятора обрабатываются декодером Витерби (мягкая схема принятия решений). Последний шаг восстановления информации — определение скорости передачи данных передатчика (9600, 4600, 2400 или 1200 бит/с); это осуществляется путем четырехкратного декодирования выходного демодулированного сигнала. Другими словами, проводится проверка для всех четырех возможных скоростей передачи данных. В процессе декодирования сигнала и анализа битов обнаружения ошибок регистрируется несколько дополнительных параметров, которые используются для выбора окончательной декодированной последовательности.

Приемник базовой станции. Базовая станция резервирует отдельный канал для получения сигналов каждого из активных пользователей ячейки. Сигналы пользователей, модулированные 64-ричным кодом Уолша, во время приема являются некогерентными (аналогично случаю приема некогерентных ортогональных сигналов MFSK). В схеме приемника обычно используется четырехкомпонентный RAKE-приемник, позволяющий демодулировать четыре наиболее мощных многолучевых компонента выходного сигнала двух антенн (см. раздел 15.7.2), которые с целью разнесения пространственно разделены между собой на расстояние, равное нескольким длинам волн. Выходные сигналы демодулятора обрабатываются декодером Витерби (мягкая схема принятия решений). Последним шагом восстановления информации является четырехкратная демодуляция сигнала с помощью процедуры, аналогичной используемой в случае мобильного устройства. Для выбора окончательной последовательности данных проводится сравнение параметров, полученных при расшифровке сигнала и анализе битов обнаружения ошибок.

12.8.4.4. Регулировка мощности

В системах, пользователи которых одновременно передают сигналы базовой станции, используя одну и ту же частоту, необходима регулировка мощности. При отсутствии такой регулировки сигналы пользователей, находящихся недалеко от базовой станции, будут приняты с гораздо бо́льшим уровнем мощности, чем сигналы пользователей, которые находятся около границы ячейки. Основная задача

процедуры регулировки — изменить процесс передачи каждого мобильного устройства таким образом, чтобы входная мощность полученных базовой станцией сигналов была равной (и по возможности постоянной). В соответствии с основным принципом работы регулирующего алгоритма уровень мощности сигналов пользователей должен быть обратно пропорционален мощности, полученной от базовой станции. Стандартом IS-95 описываются три метода регулировки мощности: управление обратным каналом; управление прямым и обратным каналами по принципу обратной связи; прямое управление каналом.

Прямое управление обратным каналом. Предположим, что потери сигнала во время распространения одинаковы для прямого и обратного каналов (на самом деле это не совсем так, поскольку рабочие частоты этих каналов разделены полосой в 45 МГц). Базовая станция постоянно передает калибровочную постоянную (которая определяется уровнем EIRP), используя синхронизационный канал. Эта информация позволяет мобильному устройству регулировать выходную мощность таким образом, чтобы мощность сигнала, полученного базовой станцией, не отличалась от сигналов других пользователей. Рассмотрим пример использования такого алгоритма. Мощность передачи сигнала мобильным устройством выбирается так, чтобы сумма мощностей переданного и полученного базовой станцией (с учетом потерь при распространении) сигналов была равна определенному значению (например, –73 дБмВт[2]), которое передается с помощью синхронизационного канала. Данное значение зависит от EIRP базовой станции. До начала процесса передачи мобильное устройство с помощью схемы автоматической регулировки усиления (automatic gain control — AGC) приемника определяет мощность, переданную по прямому каналу. Предположим, что полученная мощность равна –83 дБмВт. Тогда в соответствии с алгоритмом управления мощность передаваемого сигнала будет равна (–73 дБмВт) – (–83 дБмВт), или 10 дБмВт.

Управление прямым и обратным каналами с использованием обратной связи. При передаче в прямом канале биты регулировки мощности замещают биты кодированного сигнала, в результате чего код получается "прореженным". В каждых шести сигналах Уолша два бита данных заменяются битами регулировки мощности. Сигналы Уолша передаются со скоростью 4800 сигналов/с; следовательно, скорость передачи битов регулировки мощности должна равняться 800 бит/с. Таким образом, в каждом кадре длительностью 20 мс содержится 16 регулирующих битов. Основная задача контура регулировки мощности — коррекция ожидаемых значений открытого цикла через каждые 1,25 мс с шагом 1 дБ. Последующие модификации этого метода позволяют уменьшить шаг до 0,5 или 0,25 дБ. Наиболее важным преимуществом скоростного и высокоточного регулирования мощности по обратной связи является значительное снижение средней мощности передачи в обратном канале. При использовании аналоговых радиосистем передаваемая мощность постоянна и достаточна для поддержания связи даже в случае замирания. Следовательно, в большинстве случаев аналоговые радиоустройства используют избыточную мощность сигнала. Системы CDMA позволяют установить мощность выходного сигнала мобильного устройства на уровне, достаточном для поддержания обратного канала. В среднем для работы мобильного устройства CDMA, соответствующего стандарту IS-95, требуется уровень мощности на 20–30 дБ ниже, чем в случае аналоговой системы AMPS [30].

---

[2] Логарифмическая единица измерения мощности сигнала по отношению к 1 милливатту (1 мВт = 0 дБмВт, 0,001 мВт = –30 дБмВт)

Прямое управление каналом. Базовая станция периодически снижает мощность сигнала, передаваемого мобильному устройству. Если мобильное устройство обнаруживает увеличение количества ошибок в кадрах, отправляется запрос на увеличение мощности базовой станцией. Изменения вносятся периодически, в зависимости от значения уровня ошибок в кадре.

Пример 12.6. Элементы передачи сигналов, используемые в стандарте IS-95

Существует большое количество элементов передачи сигналов, которые описаны в стандарте IS-95 и используются в системах связи CDMA: информационные биты, канальные биты, сигналы Уолша, элементарные сигналы Уолша, элементарные сигналы с расширенным спектром, сигналы BPSK. Рассмотрим обратный канал передачи данных, используемый для передачи оцифрованной речи со скоростью 9,6 Кбит/с, причем полученный сигнал характеризуется отношением $E_b/(N_0 + I_0) \approx E_b/I_0 = 7$ дБ (при $N_0 << I_0$). Требуется найти значения следующих параметров полученного сигнала, характеризующих спектральную плотность отношения энергии к шуму, а также мощности к шуму: P_r/I_0, E_c/I_0, E_w/I_0, E_{wch}/I_0, E_{ch}/I_0. Кроме того, нужно найти следующие параметры: R_c, R_w, R_{wch}, R_{ch}. Индексы c, w, wch и ch обозначают соответственно канальный бит, сигнал Уолша, элементарный сигнал Уолша и элементарный сигнал с расширенным спектром. Сколько элементарных сигналов расширенного спектра соответствует одному элементарному сигналу Уолша?

Решение

Ключ к решению данной задачи — фундаментальные соотношения между спектральной плотностью отношения мощности к шуму полученного сигнала и каждым из указанных параметров (см. раздел 9.7.7). Следовательно, можно записать следующее.

$$\frac{P_r}{I_0} = \frac{E_b}{I_0} R = \frac{E_c}{I_0} R_c = \frac{E_w}{I_0} R_w = \frac{E_{\text{wch}}}{I_0} R_{\text{wch}} = \frac{E_{\text{ch}}}{I_0} R_{\text{ch}} \tag{12.70}$$

Поскольку известно, что $E_b/N_0 = 7$ дБ (или 5), а скорость передачи данных $R = 9600$ бит/с, можно записать следующее.

$$\frac{P_r}{I_0} = \frac{E_b}{I_0} R = 48\,000 \text{ Гц или } 46{,}8 \text{ дБГц}.$$

Поскольку для обратного канала степень кодирования равна 1/3, можем записать

$$\frac{E_c}{I_0} = \left(\frac{1}{3}\right)\frac{E_b}{I_0} = \frac{5}{3} \text{ или } 2{,}2 \text{ дБ},$$

а также

$$R_c = 3 \times R = 3 \times 9600 = 28\,800 \text{ канальных бит/с.}$$

Каждый из 64 сигналов Уолша соответствует 6 канальным битам. Следовательно,

$$\frac{E_w}{I_0} = 6 \times \frac{E_c}{I_0} = 6 \times \left(\frac{5}{3}\right) = 10 \text{ или } 10 \text{ дБ},$$

а также

$$R_w = \left(\frac{1}{6}\right) R_c = \left(\frac{1}{6}\right) 28\,800 = 4800 \text{ сигналов Уолша/с}.$$

Сигнал Уолша состоит из 64 элементарных сигналов. Тогда

$$\frac{E_{wch}}{I_0} = \left(\frac{1}{64}\right)\frac{E_w}{I_0} = \left(\frac{1}{64}\right) \times 10 = \frac{10}{64} \text{ или } -8{,}1 \text{ дБ},$$

а также

$$R_{wch} = 64 \times R_w = 64 \times 4800 = 307\ 200 \text{ элементарных сигналов Уолша/с.}$$

В соответствии со стандартом IS-95 скорость передачи сигналов расширенного спектра равна 1,2288 миллионов элементарных сигналов в секунду. Тогда

$$\frac{E_{ch}}{I_0} = \frac{P_r}{I_0} \times \left(\frac{1}{R_{ch}}\right) = \left(\frac{48\,000}{1{,}2288 \times 10^6}\right) = 0{,}039 \text{ или } -14{,}1 \text{ дБ}$$

Количество элементарных сигналов расширенного спектра, содержащихся в элементарном сигнале Уолша, равно следующему.

$$\frac{R_{ch}}{R_{wch}} = \frac{1{,}2288 \times 10^6}{307200} = 4$$

12.8.4.5. Алгоритм типичного телефонного звонка

Включение и синхронизация. С момента включения питания мобильного устройства приемник начинает поиск контрольных сигналов. Эти сигналы поступают с разных базовых станций; следовательно, псевдослучайные коды этих сигналов имеют различные временные сдвиги (см. раздел 12.8.4.1). Сигналы одной из базовых станций отличаются от всех прочих сигналов сдвигом, равным длительности 64 элементарных сигналов. Поскольку короткий код имеет максимальную длину, его 15-уровневый регистр сдвига генерирует $2^{15} - 1 = 32\ 767$ бит. После заполнения последовательности битами, перед повторением всего процесса генерируется 32 768 бит. Следовательно, всего возможно 32 768/64 = 512 уникальных адресов. Поскольку базовые станции синхронизированы во времени с погрешностью в несколько микросекунд, 512 псевдослучайных кодов могут быть созданы с помощью сдвига во времени единичной псевдослучайной последовательности. При скорости передачи элементарных сигналов 1,2288 миллионов сигналов в секунду, 75 кадров короткого кода соответствуют интервалу в 2 секунды. Модификация короткого кода с нулевым сдвигом повторяется с наступлением каждой четной секунды. Рассмотрим базовую станцию, адрес которой задается сдвигом кода на 18. Цикл передачи такой станции начинается через ((18 × 64) элементарных сигналов × (1/1,2288 × 10^6) с/элементарный сигнал) 937,5 мкс после каждой четной секунды.

После того как мобильное устройство завершает поиск и настраивается на наиболее мощный контрольный сигнал, производится синхронизация с одним из 512 уникальных адресов базовых станций. Теперь мобильное устройство может выполнить сужение любого сигнала, поступающего от базовой станции. Однако для использования каналов передачи данных, доступа и поиска необходима синхронизация во времени с системой. При использовании контрольного сигнала в качестве эталона мобильное устройство когерентно демодулирует сигнал синхронизационного канала (32-ричный код Уолша), который станция передает постоянно. Сигналы синхронизационного канала содержат информацию о нескольких системных параметрах. Наиболее важной является информация о состоянии длинного кода в течение последующих 320 мс, что дает мобильному устройству время декодировать данные, заполнять регистры и синхронизироваться во времени с системой. Данный длинный код принадлежит группе кодов, используемых для каналов поиска и доступа. Мобильное устройство выбирает определенный заранее канал поиска, основываясь на его порядковом номере, после чего постоян-

но проверяет выбранный канал на предмет наличия входящих вызовов. После этого мобильное устройство может быть зарегистрировано базовой станцией, что в случае входящего звонка позволяет производить поиск местоположения мобильного устройства (что легче поиска по всей системе).

Переход в пассивное состояние. Мобильное устройство постоянно производит поиск альтернативных контрольных сигналов. Если обнаруживается контрольный сигнал с большей мощностью, мобильное устройство перенастраивается на соответствующую станцию. Поскольку звонок отсутствует, процесс перехода служит для обновления информации о местоположении устройства. Из синхронизационного канала мобильное устройство получает информацию о временном режиме работы системы. Если бы система включала в себя только одну базовую станцию, режим работы по времени был бы произвольным. Однако в случае нескольких станций используется процесс перехода (если использование времени в системе согласовывается). В стандарте IS-95 применяется всеобщее скоординированное время (Universally Coordinated Time — UTC) с отклонением ±3 мкс. На практике такая координация реализуется с помощью глобальной системы навигации и определения положения (Global Positioning System — GPS), которая устанавливается на каждой базовой станции.

Инициация соединения. Звонок инициируется после того, как пользователь набирает номер телефона и нажимает кнопку "*send*" (отправить). После этого выполняется проверочное соединение. Мобильное устройство использует регулятор мощности, устанавливая начальную мощность передачи в соответствии с контрольным сигналом (см. раздел 12.8.4.4). Все каналы доступа имеют разные модификации сдвига длинного кода. В начале проверочного соединения мобильное устройство псевдослучайно выбирает один из каналов доступа и ставит его в соответствие поисковому каналу. Проверочное соединение начинается в момент времени, соответствующий началу интервала канала доступа (что определяется псевдослучайным образом). Ключевым моментом процедуры предоставления доступа является проверка порядкового номера абонента. Такая проверка необходима, поскольку канал доступа может использоваться всеми абонентами без каких-либо ограничений.

Время начала передачи мобильным терминалом определяется первым компонентом многолучевого сигнала, который используется для демодуляции. Мобильное устройство не учитывает время задержки распространения и не вносит соответствующих поправок в параметры передаваемого сигнала. Вместо этого базовая станция постоянно выполняет поиск обратных каналов связи. Мобильное устройство "прослушивает" поисковый канал, ожидая отклика базовой станции. Если отклик не получен (во время использования канала доступа может возникнуть конфликтная ситуация), мобильное устройство повторяет попытку после паузы псевдослучайной длительности. Если же пробный доступ успешно получен, базовая станция предоставляет устройству канал данных (передает код Уолша).

В каналах передачи данных и поисковых каналах применяются различные сдвиги длинных кодов. Поэтому мобильное устройство переходит к использованию кода, который основывается на порядковом номере. После получения кода Уолша мобильное устройство передает последовательность нулей в канал данных, после чего ожидает положительного подтверждения приема от прямого канала данных. Если обмен сигналами прошел успешно, следующим шагом будет звонок вызываемого телефона. Телефонный разговор может начинаться.

Плавный переход. Во время телефонного разговора мобильное устройство может обнаружить альтернативный контрольный сигнал, более сильный по сравнению с используемым. В этом случае на базовую станцию отправляется контрольное сообщение, содержащее информацию о новой станции с более мощным сигналом, а также запрос на плавный переход. Исходная базовая станция передает запрос на контроллер, осуществляющий управление радиоресурсами (base station controller — BSC). В некоторых случаях BSC может быть совмещен с центром коммутации мобильных устройств (Mobile Switching Center — MSC), который управляет параметрами связи, не связанными с радиопередачей (в частности, переключением). Контроллер BSC связывается с "новой" базовой станцией и получает код Уолша. Этот код пересылается мобильному устройству через исходную базовую станцию. В процессе перехода мобильное устройство подключено к двум станциям одновременно. В это время также поддерживается связь между контроллером BSC и двумя базовыми станциями. Мобильное устройство совмещает голосовые сигналы, получаемые от двух станций, используя соответствующие контрольные сигналы в качестве когерентных фазовых эталонов. Прием одновременно двух сигналов, которые для мобильного устройства аналогичны двум многолучевым компонентам, обеспечивается RAKE-приемником. Сигналы мобильного устройства, поступающие на контроллер BSC, являются некогерентными. После сравнения двух полученных сигналов контроллером выбирается более качественный. Сигналы сравниваются с интервалом 20 мс (длительность одного кадра). Исходная базовая станция прекращает поддержку звонка только после того, как установлено соединение в новой ячейке. Подобная двойная поддержка связи снижает вероятность разрыва соединения и значительно улучшает качество связи на границе двух ячеек.

12.9. Резюме

Технология использования расширенного спектра (spread-spectrum — SS) была разработана в 1950-х годах. Расширенный спектр используется и сегодня в большинстве современных систем связи Национального аэрокосмического агентства (NASA), а также в армии США для обеспечения множественного доступа, устойчивости к интерференции и масштабирования. В данной главе перечислены основные методы расширения спектра, а также преимущества их использования. Кроме того, здесь приводится краткая историческая справка.

Поскольку изначально системы расширенного спектра разрабатывались для военных целей, в начале главы подробно рассмотрены методы повышения устойчивости к преднамеренным помехам. Применение псевдослучайных последовательностей является основой всех современных систем связи расширенного спектра. Поэтому здесь подробно описаны псевдослучайные последовательности. Кроме того, в этой главе подробно рассмотрены два основных метода связи расширенного спектра: использование прямой последовательности и скачкообразной перестройки частоты. Проанализирован также процесс синхронизации сигналов для систем связи расширенного спектра. Особое внимание уделено коммерческому использованию методов расширенного спектра. В частности, в главе рассматриваются системы связи CDMA, соответствующие стандарту IS-95.

Литература

1. Scholtz R. A. *The Origins of Spread Spectrum Communications.* IEEE Trans. Commun., vol. COM30, n. 5, May, 1982, pp. 822–854.
2. Shannon C. E. *Communication in the Presence of Noise.* Proc. IRE, January, 1949, pp. 10–21.
3. Dillard R. A. *Detectability of Spread Spectrum Signals.* IEEE Trans. Aerosp. Electron. Syst., July, 1979.
4. Simon M. K., Omura J. K., Scholtz R. A. and Levitt, B. K., *Spread Spectrum Communications.* Computer Science Press, Inc., Rockville, Md., 1985.
5. de Rosa L. A. and Rogoff M., Sec. I (Communications) of *Application of Statistical Methods to Secrecy Communication Systems.* Proposal 946, Fed. Telecommun. Lab., Nutley, N. J., August, 28, 1950.
6. Pickholtz R. L., Schilling D. L. and Milstein L. B. *Theory of Spread-Spectrum Communications — A Tutorial.* IEEE Trans. Commun., vol. COM30, n. 5, May, 1982, pp. 855–884.
7. Pickholtz R. L., Schilling D. L. and Milstein L. B. *Revisions to Theory of Spread-Spectrum Communications — A Tutorial.* IEEE Trans. Commun., vol. COM32, n. 2, February, 1984, pp. 211–212.
8. Simon M. K., Omura J. K., Scholtz R. A. and Levitt B. K. *Spead Spectrum Communications*, Vol. 2, Computer Science Press, Inc., Rockville, Md., 1985.
9. Simon M. K. and Polydoros A. *Coherent Detection of Frequency-Hopped Quadrature Modulations in the presence of Jamming*: Part I. QPSK and QASK; Part II QPR class I Modulation. IEEE Trans. Commun., vol. COM29, November, 1981, pp. 1644–1668.
10. Holmes J. K. and Chen C. C. *Acquisition Time Performance of PN Spread-Spectrum Systems.* IEEE Trans. Commun., COM-25, August, 1977, pp. 778–783.
11. Ward R. B. *Acquisition of Pseudonoise Signals by Sequential Estimation.* IEEE Trans. Commun., COM13, December, 1965, pp. 475–483.
12. Spilker J. J. and Magill, D. T. *The Delay-Lock Discriminator — An Optimum Tracking Device.* Proc. IRE, September, 1961.
13. Spiler J. J. *Delay-Lock Tracking of Binary Signals.* IEEE Trans. Space Electron. Telem., March, 1963.
14. Simon M. K. *Noncoherent Pseudonoise Code Tracking Perfomance of Spread Spectrum Receivers.* Commun., vol. COM25, March, 1977.
15. Ziemer R. E. and Peterson R. L. *Digital Communications and Spread Spectrum Systems.* Macmillan Publishing Company, New York, 1985.
16. Holmes J. K. *Coherent Spread Spectrum Systems.* John Wiley & Sons, Inc., New York, 1982.
17. Pursley M. B. *Performance Evaluation for Phase-Coded Spread-Spectrum Multiple-Access Communication*: Part I. System Analysis. IEEE Trans. Commun., vol. COM25, n. 8, August, 1977, pp. 795–799.
18. Geraniotis E. *Noncoherent Hybrid DS-SFH Spread-Spectrum Multiple-Access Communications.* IEEE Trans. Commun., vol. COM34, n. 9, September, 1986, pp. 862–872.
19. Geraniotis E. and Pursley M. B. *Error Probabilities for Direct-Sequence Spread-Spectrum Multiple-Access Communications*: Part I. Upper and Lower Bounds. IEEE Trans. Commun., vol. COM30, n. 5, May, 1982, pp. 985–995.
20. Geraniotis E., and Pursley M. B. *Error Probabilities for Direct-Sequence Spread-Spectrum Multiple-Access Communications*: Part II. Approximations. IEEE Trans. Commun., vol. COM30, n. 5, May, 1982, pp. 996–1009.
21. Schilling D. L., Milstein L. B., Pickholtz R. L. and Brown R. W. *Optimization of the Processing Gain of an M-ary Direct Sequence Spread Spectrum Communication System.* IEEE Trans. Commun., vol. COM28, n. 8, August, 1980, pp. 1389–1398.
22. Viterbi A. J. and Jacobs I. M. *Advances in Coding and Modulation for Noncoherent Channels Affected by Fading, Partial Band, and Multiple Access Interference*; in A. S. Viterbi, ed., *Advances in Communication Systems*, Vol. 4, Academic Press, Inc., New York, 1975.
23. Stark W. E. *Coding for Frequency-Hopped Spread-Spectrum Communication with Partial-Band Interference*: Part I. Capacity and Cutoff Rate. IEEE Trans. Commun., vol. COM33, n. 10, October, 1985, pp. 1036–1044.

24. Stark W. E. *Coding for Frequency-Hopped Spread-Spectrum Communication with Partial-Band Interference*: Part II. Coded Performance. IEEE Trans. Commun., vol. COM33, n. 10, October, 1985, pp. 1045–1057.
25. Milstein L. B., Davidovici S. and Schilling D. L. *The Effect of Multiple-Tone Interfering Signals on a Direct Sequence Spread Communication System.* IEEE Trans. Commun., vol. COM30, March, 1982, pp. 436–446.
26. Milstein L. B., Pickholtz R. L. and Schilling D. L. *Optimization of the Processing Gain of an FSK-FH System.* IEEE Trans. Commun., vol. COM28, July, 1980, pp. 1062–1079.
27. Huth G. K. *Optimization of Coded Spread Spectrum Systems Performance.* IEEE Trans. Commun., vol. COM25, August, 1977, pp. 763–770.
28. Viterbi A. J. *Spread-Spectrum Communications — Myths and Realities.* IEEE Commun. Mag., May, 1979, pp. 11–18.
29. Simon M. K., Omura J. K., Scholtz R. A. and Levitt B. *Spread Spectrum Communications Handbook.* Revised Edition, McGraw-Hill, Inc., New York, 1994.
30. Viterbi A. J. *The Orthogonal-Random Waveform Dichotomy for Digital Mobile Personal Communication.* IEEE Personal Communications, First Quarter 1994, pp. 18–24.
31. Kohno R., Meidan R. and Milstein L. B. *Spread Spectrum Access Methods for Wireless Communications.* IEEE Communications Magazine, January, 1995, pp. 58–67.
32. Pickholtz R. L., Milstein L. B. and Schilling D. L. *Spread Spectrum for Mobile Communications.* IEEE Trans. Vehicular Tech., vol. 40, n. 2, May, 1991, pp. 313–321.
33. Morrow R. K., Jr. and Lehnert J. S. *Bit-to-Bit Error Dependence on Slotted DS/SSMA Packet Systems with Random Signature Sequences.* IEEE Trans. Commun., vol. 37, n. 10, October, 1989, pp. 1052–1061.
34. Schilling D. L., et. al. *Spread Spectrum for Commercial Communications.* IEEE Communications Magazine, April, 1991, pp. 66–78.
35. Gilhousen K. S. *On the Capacity of a Cellular CDMA System.* IEEE Trans. Vehicular Tech., vol. 40, n. 2, May, 1991, pp. 303–312.
36. Viterbi A. M. and Viterbi A. J. *Erlang Capacity of a Power Controlled CDMA System.* IEEE JSAC, vol. 11, n. 6, pp. 892–899.
37. Padovani R. *Reverse Link Performance of IS-95 Based Cellular Systems.* IEEE Personal Communications, Third Quarter 1994, pp. 28–34.
38. *Wideband CDMA Special Issue.* IEEE Communications Magazine, vol. 36, n. 9, September, 1998.

Задачи

12.1. Объясните, почему линейный n-разрядный регистр сдвига с обратной связью максимальной длины способен генерировать последовательности с периодом не более $2^n - 1$.

12.2. Докажите, что для линейного n-разрядного регистра сдвига с обратной связью максимальной длины выходной разряд всегда должен подаваться на вход схемы обратной связи.

12.3. Рассмотрим передатчик расширенного спектра DS/BPSK, представленный на рис. 12.9, *а* (или 12.9, *б*). Последовательность $x(t)$ равна 1 0 0 1 1 0 0 0 1; скорость передачи данных 75 бит/с. Передача данных начинается с левого крайнего бита. Допустим, $g(t)$ генерируется регистром сдвига, который изображен на рис. 12.7. Начальное состояние регистра 1 1 1 1, а частота синхронизирующих импульсов равна 225 Гц.

а) Изобразите переданную последовательность $x(t)g(t)$.

б) Определите ширину полосы переданного (расширенного) сигнала.

в) Определите коэффициент расширения спектра сигнала.

г) Предположим, что ожидаемое время задержки $\hat{T}_d$ значительно превышает время передачи элементарного сигнала (см. рис. 12.9, *в*). Определите последовательность сужающих элементарных сигналов.

д) Найдите правило определения $\hat{x}(t)$ и ошибок.

12.4. В системе множественного доступа с кодовым разделением (CDMA) 24 терминала равной мощности одновременно используют полосу частот. Каждый терминал передает данные со скоростью 9,6 Кбит/с с помощью расширения спектра методом прямой последовательности, а также с использованием модуляции BPSK. Рассчитайте минимальную скорость передачи элементарных сигналов псевдослучайного кода, при которой вероятность битовой ошибки бита равна 10^{-3}. Предположим, что шумы приемника ничтожно малы по сравнению с интерференцией, вызванной другими пользователями.

12.5. Регистр сдвига с обратной связью, генерирующий псевдослучайные коды, создает последовательность размером 31 бит при частоте синхронизации 10 МГц. Найдите и отобразите графически автокорреляционную функцию и спектральную плотность последовательности. Допустим, что значения импульсов равны ±1.

12.6. Рассмотрим систему связи FH/MFSK, представленную на рис. 12.11. Будем считать, что генератор псевдослучайных кодов — это 20-разрядный линейный регистр сдвига с максимальной длиной последовательности. Каждое состояние регистра задает новый центр диапазона изменения частоты. Минимальный шаг между центрами полос (от скачка до скачка) равен 200 Гц. Частота тактового генератора регистра равна 2 кГц. Будем считать, что используется модуляция 8-FSK. Скорость передачи данных равна 1,2 Кбит/с.

а) Определите ширину полосы, в которой выполняются скачки частоты.

б) Найдите скорость передачи элементарных сигналов.

в) Сколько элементарных сигналов содержится в каждом информационном символе?

г) Найдите коэффициент расширения спектра сигнала.

12.7. На рис. 12.16 (раздел 12.4.5) приводится блок-схема демодулятора с быстрой перестройкой частоты (FFH). Изобразите блок-схему демодулятора с медленной перестройкой частоты (SFH) и объясните работу этой схемы.

12.8. Найдите среднее и среднеквадратическое отклонение времени, необходимого для обнаружения последовательности, модулированной BPSK с псевдослучайным кодом. Последовательность передается со скоростью 10 миллионов элементарных сигналов в секунду. Для обнаружения используется повторяющаяся процедура поиска с одновременной обработкой 100 элементарных сигналов. Последовательность считается обнаруженной, когда 100 полученных и сгенерированных элементарных сигналов совпадают. Отношение энергии полученного сигнала к спектральной плотности мощности шума составляет 9,6 дБ. Несоответствие во времени между полученным и сгенерированным кодами равно 1 мс. Будем считать вероятность ложного обнаружения последовательности пренебрежимо малой.

12.9. В системе связи CDMA 11 терминалов равной мощности передают сигналы на центральный узел. Каждый терминал передает информацию со скоростью 1 Кбит/с, используя сигнал расширенного спектра с использованием метода прямой последовательности, модулированный BPSK. Скорость передачи сигнала равна 100 000 элементарных сигналов в секунду.

а) Найдите отношение энергии, необходимой для передачи одного бита, к спектральной плотности мощности интерференции (E_b/I_0) с сигналами от других пользователей. Будем считать, что шумы, получаемые приемником, ничтожно малы по сравнению с интерференцией между пользователями.

б) Как изменится отношение (E_b/I_0), если все пользователи удвоят мощность выходного сигнала?

в) Необходимо увеличить количество пользователей до 101, при этом мощность выходных сигналов должна остаться равной. Как сохранить неизменным отношение E_b/I_0?

12.10. Система CDMA использует расширение спектра методом прямой последовательности. Ширина полосы передачи данных составляет 10 кГц, а полосы расширенного спектра — 10 МГц. При передаче единичного сигнала отношение E_b/N_0 для приемника равно 16 дБ.

а) Если необходимое значение $(E_b/N_0 + I_0)$ равно 10 дБ, сколько абонентов с одинаковой мощностью выходного сигнала могут одновременно использовать полосу? Учесть в решении шумы, поступающие на приемник.

б) Мощность передаваемого сигнала каждого пользователя снижена на 3 дБ. Сколько абонентов с равной выходной мощностью смогут использовать полосу?

в) Если значение полученного $E_b/N_0 \rightarrow \infty$ для каждого приемника, какое максимальное количество абонентов могут одновременно использовать полосу?

12.11. Для устранения эффектов многолучевого распространения используется система DS/SS. Разница пути распространения между прямым и побочным сигналами составляет 100 м. Какой должна быть скорость передачи элементарных сигналов для предотвращения многолучевой интерференции?

12.12. Необходимо установить связь между наземным передатчиком и синхронно работающим спутником при наличии умышленных помех. Скорость передачи данных равна 1 Кбит/с. Наземная станция использует 60-футовую антенну (60 футов = 18,288 метров). Для защиты от умышленных помех применяется сигнал, расширенный методом прямой последовательности, со скоростью передачи 10 Мбит/с. Станция умышленных помех использует 150-футовую антенну (150 футов = 45,72 метра); мощность ее передатчика равна 400 кВт. Будем считать потери, связанные с пространственными факторами и распространением сигналов, равными для обеих станций. Какой должна быть мощность передатчика наземной станции, чтобы отношение E_b/J_0 спутникового приемника было равно 16 дБ? Шумы приемника считать пренебрежимо малыми.

12.13. Данные, получаемые со скоростью 75 бит/с, закодированы. Степень кодирования равна 1/2. Кодированные биты модулируются с использованием 8-FSK. Символы FSK разделяются с помощью скачков частоты (2000 скачков/с).

а) Найдите скорость передачи элементарных сигналов.

б) Какова кратность разнесения (число независимых копий сигнала)?

в) Если в канале имеется два сигнала TDM с равным периодом скачков частот, как изменятся значения скорости передачи элементарных сигналов и символов? Как изменится кратность разнесения?

г) В канале имеется 80 сигналов TDM. Как изменятся значения скорости передачи элементарных сигналов и символов, а также кратность разнесения?

12.14. Некогерентная двоичная система FSK со скачкообразной перестройкой частоты характеризуется отношением $E_b/N_0 = 30$ дБ; ширина полосы равна 2 ГГц. Канальное кодирование не используется. Станция преднамеренных помех, работающая в том же широкополосном диапазоне, характеризуется полученным $J_0 = 100N_0$.

а) Найдите вероятность битовой ошибки P_B.

б) Станция преднамеренных помех использует лишь часть диапазона. Использование какой полосы позволит создавать помехи наиболее эффективно?

в) Найдите значение P_B для наиболее эффективного создания помех в определенной части диапазона.

г) Найдите значение P_B при отсутствии помех.

12.15. Некогерентная система связи с использованием скачкообразной перестройки частоты и модуляции 8-FSK совершает 1200 частотных скачков в секунду; ширина рабочей полосы системы 1 МГц. В течение одной секунды производится передача 3000 символов. Канальное кодирование не используется. Мощность сигнала на входе приемника составляет 10^{-12} Вт. Умышленные помехи создаются в части диапазона (50 кГц) передачи сигнала (станция помех использует часть своей рабочей полосы). Мощность полученных помех составляет 10^{-11} Вт. Температура системы равна 290 К. Найдите вероятность битовой ошибки.

12.16. Когерентная система DS/BPSK передает данные со скоростью 10 Кбит/с при наличии широкополосных умышленных помех. Канальное кодирование не используется. Потери мощности, связанные с распространением сигнала, равны для системы и станции умышленных помех.

а) Эффективная изотропно-излучаемая мощность (EIRP) передатчика и станции умышленных помех равна, соответственно, 20 и 60 кВт. Вычислите ширину полосы расширенного спектра, необходимую для достижения вероятности битовой ошибки $P_B = 10^{-5}$.

б) Станция умышленных помех работает в импульсном режиме. Найдите рабочий цикл, при котором помехи будут наносить максимальный ущерб. Найдите значение P_B для такого рабочего цикла.

12.17. Станция связи передает сигнал со скоростью скачкообразной перестройки частоты 10 000 скачков/с, чтобы избежать создания ретрансляционных помех.

а) Допустим, спутник, на который производится передача сигнала, находится на геосинхронной орбите (приблизительно 36 000 км) непосредственно над передатчиком. Кривизной поверхности Земли пренебрегаем. Вычислите радиус защищенности, в пределах которого передатчик ни при каких условиях не может подвергаться опасности создания ретрансляционных помех (станция умышленных помех находится на земле).

б) Станции умышленных помех необходимо 10 мкс для обнаружения частоты сигнала и настройки выходного канала генератора помех. Вычислите радиус защищенности, считая, что данная информация доступна передатчику.

12.18. Генератор ретрансляционных помех расположен на борту самолета (рис. 312.1). Для связи используется система FH/SS. Найдите минимальную скорость изменения частоты, при которой ретрансляционные помехи не будут ухудшать качество связи. Найдите необходимую минимальную скорость изменения частоты для случая, когда передатчик расположен на борту самолета, а генератор помех — на Земле.

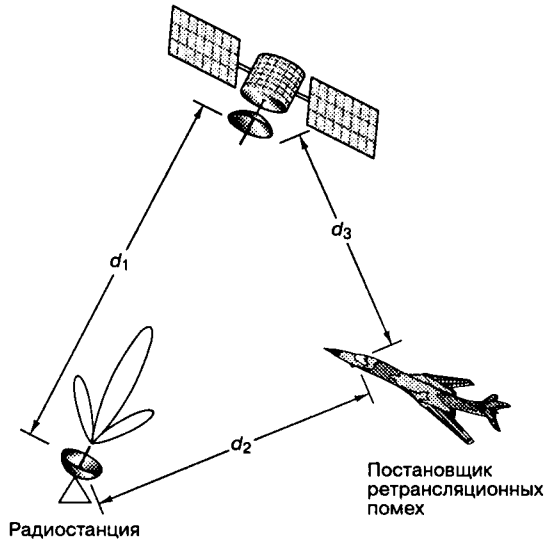


Рис. 312.1

12.19. Методы расширенного спектра могут применяться для выполнения требований государственных стандартов относительно плотности (мощности) потока излучения на поверхности Земли. Спутник связи, находящийся на высоте 36 000 км над уровнем моря, передает данные со скоростью 4 Кбит/с. Эффективная изотропно-излучаемая мощность равна 100 Вт. Найдите ширину полосы расширения, необходимую для того, чтобы плотность потока излучения на поверхности Земли не превышала –151 дБВт/м^2 для любой полосы шириной 4 кГц.

12.20. Для предотвращения негативного влияния умышленных помех на сигнал передатчик использует некогерентную модуляцию BFSK, а также скачкообразную перестройку частоты. Мощность сигнала на входе приемника равна 10 мкВт. При отсутствии умышленных помех отношение мощности сигнала к шуму очень велико. Мощность умышленных помех на входе приемника равна 1 Вт.

а) Станция умышленных помех генерирует гауссов шум, равный мощности во всем диапазоне изменения частоты (для данной полосы помехи можно считать белым шумом). Определите, во сколько раз должен быть увеличен диапазон полосы, чтобы позволить передатчику достичь вероятности битовой ошибки 10^{-4}.

б) Генератор помех снижает (относительно полной мощности) мощность шумов в половине диапазона на α ($0 \le \alpha \le 1$). Одновременно мощность шумов в другой части диапазона повышается на α (суммарная энергия шумов не изменяется). Передатчик не изменяет параметры перестройки частоты. Найдите выражение для вероятности битовой ошибки.

в) Определите оптимальное значение α для следующих случаев: эффективное отношение мощности сигнала к шуму велико; отношение мощности сигнала к шуму незначительно.

12.21. Применение методов расширенного спектра позволяет получить значительное преимущество при наличии преднамеренных помех. Объясните, почему использование расширенного спектра не дает преимуществ при шуме AWGN.

12.22. Мобильное радиоустройство расширения спектра методом прямой последовательности является частью сотовой системы CDMA. Характеристики системы: данные и коды SS модулируются BPSK; скорость передачи данных равна 8000 бит/с; частота несущей 1 ГГц; скорость передачи элементарных сигналов составляет 25 миллионов сигналов в секунду; максимальные потери сигнала при распространении 138,6 дБ; коэффициент усиления передающей антенны равен 5 дБ; добротность приемника $G/T = -18$ дБ/К; случайные потери, связанные с мелкомасштабным замиранием, составляют 30 дБ; прочие потери — 4 дБ; необходимое значение отношения $E_b/N_0 = 4$ дБ. Коэффициенты G_A, G_V, H_0 и γ равны, соответственно, 2,5; 2,5; 1,6; 1. *Подсказка*: описание параметров канала связи дано в главе 5.

а) Найдите значение мощности передатчика P_t в процессе мелкомасштабного замирания сигнала.

б) Насколько может быть снижено значение P_t при отсутствии мелкомасштабного замирания сигнала?

в) Найдите минимальное значение E_{ch}/N_0, соответствующее указанным параметрам.

г) Найдите коэффициент расширения спектра сигнала.

д) Найдите максимальное количество пользователей в ячейке.

12.23. В системе связи расширения спектра методом прямой последовательности при использовании модуляции BPSK (данных и кодов) необходимо поддерживать скорость передачи данных 9600 бит/с. Отношение (P_r/N_0) полученного сигнала до обнаружения равно 48 дБГц. Коэффициент усиления при расширении спектра равен 1000. Для исправления ошибок используется код БХЧ (63, 51). Определите, способна ли система с такими параметрами поддерживать уровень вероятности битовой ошибки 10^{-4}. Используйте уравнение (6.46) для вычисления вероятности ошибки в декодированном бите.

12.24. а) Каждому пользователю сотовой системы телефонной связи CDMA с использованием метода прямой последовательности необходимо, чтобы отношение E_b/I_0 было равно 6 дБ для приемлемого качества передачи голоса. Скорость передачи элементарных сигналов равна 3,68 миллионов сигналов в секунду; скорость передачи данных — 14,4 Кбит/с. Коэффициенты G_V, H_0 и γ равны, соответственно, 2,5; 1,5; 1,5. Во время речевых пауз передача сигнала не производится. Найдите максимальное количество пользователей в ячейке.

б) Отношение E_b/I_0 было снижено на 1 дБ за счет использования эффективного кода коррекции ошибок. Найдите максимальное количество пользователей в ячейке.

12.25. Система связи расширенного спектра с использованием метода прямой последовательности использует для передачи данных модуляцию QPSK. Необходимо, чтобы значение вероятности битовой ошибки было равно 10^{-5}, а отношение E_{ch}/I_0 не превышало –30,4 дБ. Считая синхронизацию идеальной, найдите минимально необходимое количество элементарных сигналов в 1 бите.

12.26. Система связи расширенного спектра с использованием метода прямой последовательности использует для передачи данных модуляцию QPSK. Коэффициент расширения спектра сигнала равен 20 дБ. Используется код исправления ошибок со степенью кодирования 1/2. Необходимое значение вероятности битовой ошибки равно 10^{-5}. Считая синхронизацию идеальной, найдите минимальные значения E_{ch}/I_0 и E_c/I_0, достаточные для удовлетворения указанного требования.

12.27. а) Система расширенного спектра с быстрой перестройкой частоты (FFH/SS) для передачи данных использует модуляцию 8-FSK и код коррекции ошибок со степенью кодирования 1/2. Коэффициент повторной передачи элементарных сигналов $N = 4$. Другими словами, каждый символ пересылается четыре раза во время разных частотных скачков. Необходимое значение E_b/I_0 равно 13 дБ. Элементарные сигналы передаются со скоростью 32 000 сигналов в секунду; ширина полосы частотных скачков — 1,2 МГц. Найдите скорость передачи данных R, коэффициент расширения спектра сигнала G_p, а также отношения (P_r/I_0), E_{ch}/I_0, E_s/I_0 и E_c/I_0.

б) Соответствуют ли ширина полосы и коэффициент расширения спектра сигнала системы требованиям Part-15 для полосы частот ISM?

12.28. Сотовая система телефонной связи CDMA соответствует стандарту IS-95 с некоторыми модификациями: скорость передачи элементарных сигналов расширенного спектра равна 10,24 сигналов/с; скорость передачи данных — 20 Кбит/с; для обратной связи используется 256-ричный код Уолша. Данные, закодированные кодом со степенью кодирования 1/2, модулируются сигналом Уолша, для чего отношение E_b/I_0 должно быть равно 6 дБ. Найдите значения следующих параметров: P_r/I_0, E_c/I_0, E_w/I_0, E_{wch}/I_0 и E_{ch}/I_0. Найдите также значения R_c, R_w и R_{wch}. Индексы *c*, *w*, *wch* и *ch* обозначают, соответственно, канальный бит, сигнал Уолша, элементарный сигнал Уолша и элементарный сигнал расширенного спектра. Найдите коэффициент расширения спектра сигнала. Определите, сколько элементарных сигналов расширенного спектра соответствуют одному элементарному сигналу Уолша.

Вопросы

12.1. Импульсно-кодовая модуляция (PCM) и частотная модуляция (FM) позволяют расширить спектр сигнала данных. Почему сигналы PCM и FM не считают сигналами расширенного спектра (см. раздел 12.1)?

12.2. Назовите четыре основных преимущества систем связи расширенного спектра (см. раздел 12.1.1).

12.3. Укажите три критерия, в соответствии с которыми псевдослучайный сигнал будет казаться случайным (см. раздел 12.2.1).

12.4. Дайте определение элементарного сигнала для систем, использующих метод прямой последовательности, а также для систем со скачкообразной перестройкой частоты (см. разделы 12.3.2 и 12.4.4).

12.5. Что подразумевается под устойчивым сигналом (см. раздел 12.4.2)?

12.6. Объясните разницу между быстрой и медленной скачкообразной перестройкой частоты (см. раздел 12.4.4).

12.1. В чем отличие коэффициента расширения спектра сигнала для системы, использующей метод прямой последовательности, и системы со скачкообразной перестройкой частоты (см. разделы 12.3.2 и 12.4.6)?

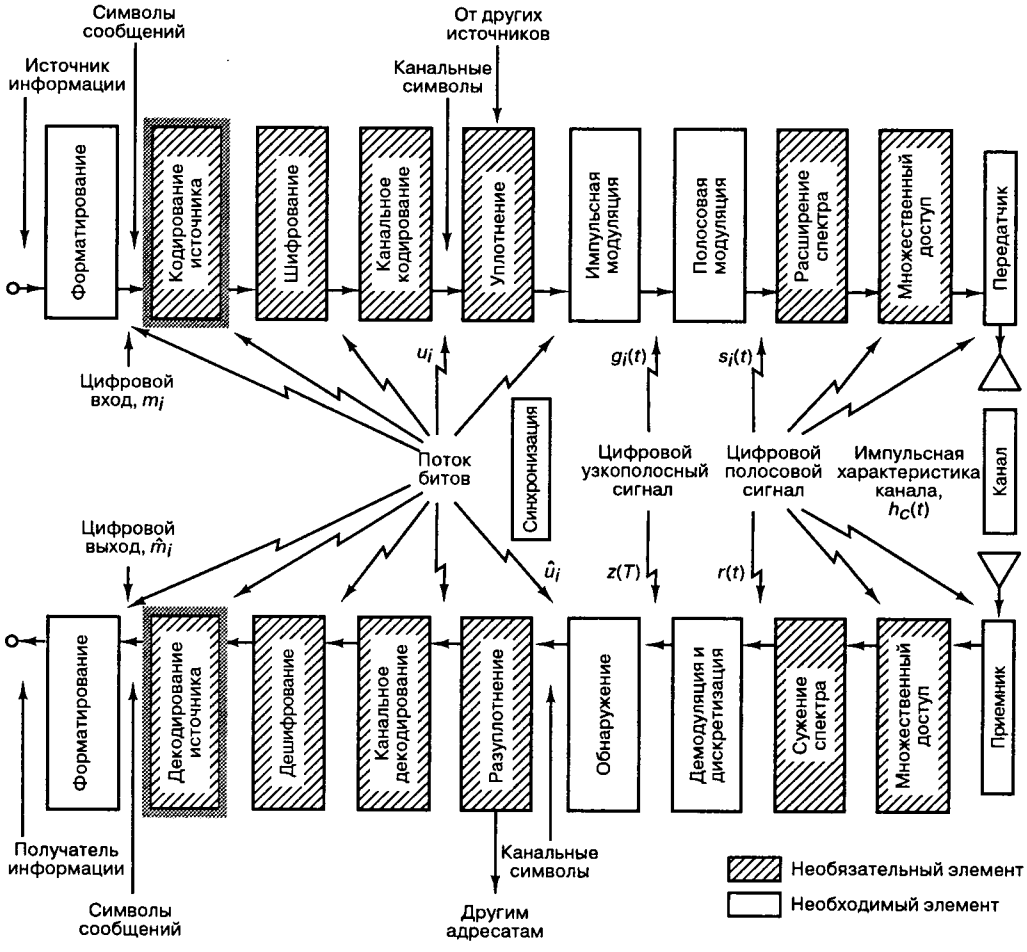
12.2. Объясните, каким образом система расширенного спектра расшифровывает сигналы, "скрытые" в шумах (см. раздел 12.5).

12.3. Системы, соответствующие стандарту IS-95, используют коды Уолша для совершенно разных задач при передаче в прямом и обратном каналах. Объясните использование кодов Уолша в обоих случаях (см. разделы 12.8.4.1 и 12.8.4.2).

ГЛАВА 13

Кодирование источника

Фредрик Дж. Харрис (Fredric J. Harris)
Университет Сан-Диего
Сан-Диего, Калифорния



13.1. Источники

Кодирование источника связано с задачей создания эффективного описания исходной информации. Эффективное описание допускает снижение требований к памяти или полосе частот, связанных с хранением или передачей дискретных реализаций исходных данных. Для дискретных источников способность к созданию описаний данных со сниженной скоростью передачи зависит от информационного содержимого и статистической корреляции исходных символов. Для аналоговых источников способность к созданию описаний данных со сниженной скоростью передачи (согласно принятому критерию точности) зависит от распределения амплитуд и временной корреляции волнового сигнала источника. Целью кодирования источника является получение описания исходной информации с хорошей точностью при данной номинальной скорости передачи битов или допуск низкой скорости передачи битов, чтобы получить описание источника с заданной точностью. Чтобы понять, где эффективны методы и средства кодирования источника, важно иметь общие меры исходных параметров. По этой причине в данном разделе изучаются простые модели дискретных и аналоговых источников, а затем дается описание того, как кодирование источника может быть применено к этим моделям.

13.1.1. Дискретные источники

Дискретные источники генерируют (или выдают) последовательность символов $X(k)$, выбранную из исходного алфавита в дискретные промежутки времени kT, где $k = 1, 2, \ldots$ — счетные индексы. Если алфавит содержит конечное число символов, скажем N, говорят, что источник является *конечным дискретным* (finite discrete source). Примером такого источника является выход 12-битового цифро-аналогового преобразователя (один из 4096 дискретных уровней) или выход 10-битового аналого-цифрового преобразователя (один из 1024 двоичных 10-кортежей). Еще одним примером дискретного источника может послужить последовательность 8-битовых ASCII-символов, введенных с клавиатуры компьютера.

Конечный дискретный источник определяется последовательностью символов (иногда называемых алфавитом) и вероятностью, присвоенной этим символам (или буквам). Будем предполагать, что источник кратковременно стационарный, т.е. присвоенные вероятности являются фиксированными в течение периода наблюдения. Пример, в котором алфавит фиксирован, а присвоенные вероятности изменяются, — это последовательность символов, генерируемая клавиатурой, когда кто-то печатает английский текст, за которым следует печать испанского и наконец французского текстов.

Если известно, что вероятность каждого символа X_j есть $P(X_j)$, можно определить *самоинформацию* (self-information) $I(X_j)$ для каждого символа алфавита.

$$I(X_j) = -\log_2(p_j) \tag{13.1}$$

Средней самоинформацией для символов алфавита, называемой также *энтропией источника* (source entropy), является следующее.

$$H(X) = \mathbf{E}\{I(X_j)\} = -\sum_{j=1}^{N} p_j \log_2(p_j), \tag{13.2}$$

где $\mathbf{E}\{X\}$ — математическое ожидание X. Энтропия источника определяется как среднее количество информации на выход источника. Энтропия источника — это средний объем неопределенности, которая может быть разрешена с использованием алфавита. Таким образом, это среднее количество информации, которое должно быть отправлено через канал связи для разрешения этой неопределенности. Можно показать, что это количество информации в битах на символ ограничено снизу нулем, если не существует неопределенности, и сверху $\log_2(N)$, если неопределенность максимальна.

$$0 \le H(X) \le \log_2(N) \tag{13.3}$$

Пример 13.1. Энтропия двоичного источника

Рассмотрим двоичный источник, который генерирует независимые символы 0 и 1 с вероятностями p и $(1-p)$. Этот источник описан в разделе 7.4.2, а его функция энтропии представлена на рис. 7.5. Если $p = 0{,}1$ и $(1-p) = 0{,}9$, энтропия источника равна следующему.

$$H(X) = -[p \log_2(p) + (1-p) \log_2(1-p)] = \\ = 0{,}47 \text{ бит/символ} \tag{13.4}$$

Таким образом, этот источник может быть описан (при использовании соответствующего кодирования) с помощью менее половины бита на символ, а не одного бита на символ, как в текущей форме.

Отметим, что первая причина, по которой кодирование источника работает, — это то, что информационное содержание N-символьного алфавита, используемое в действительных системах связи, обычно меньше верхнего предела соотношения (13.3). Известно, что, как отмечено в примере 7.1, символы английского текста не являются равновероятными. Например, высокая вероятность конкретных букв в тексте используется как часть стратегии игры Хенгмана (Hangman). (В этой игре игрок должен угадывать буквы, но не их позиции в скрытом слове известной длины. За неверные предположения назначаются штрафы, а буквы всего слова должны быть определены до того, как произойдет шесть неверных предположений.)

Дискретный источник называется источником *без памяти* (memoryless), если символы, генерируемые источником, являются статистически независимыми. В частности, это означает, что их совместная вероятность двух символов является просто произведением вероятностей соответствующих символов.

$$P(X_j, X_k) = P(X_j | X_k) P(X_k) = P(X_j) P(X_k) \tag{13.5}$$

Следствием статистической независимости есть то, что информация, требуемая для передачи последовательности M символов (называемой M-кортежем) данного алфавита, точно в M раз превышает среднюю информацию, необходимую для передачи отдельного символа. Это объясняется тем, что вероятность статистически независимого M-кортежа задается следующим образом.

$$P(X_1, X_2, \ldots, X_M) = \prod_{n=1}^{M} P(X_n) \tag{13.6}$$

Поэтому средняя на символ энтропия статистически независимого M-кортежа равна следующему.

$$H_M(X) = \frac{1}{M}\mathbf{E}\{-\log_2 P(X_1 X_2, \ldots, X_M)\} =$$
$$= \frac{1}{M}\sum_{X_m}[-P(X_m)\log_2 P(X_m)] = \tag{13.7}$$
$$= H(X)$$

Говорят, что дискретный источник имеет память, если элементы источника, образующие последовательность, не являются независимыми. Зависимость символов означает, что для последовательности M символов неопределенность относительно M-го символа уменьшается, если известны предыдущие $(M-1)$ символов. Например, большая ли неопределенность существует для следующего символа последовательности CALIFORNI_? M-кортеж с зависимыми символами содержит меньше информации или разрешает меньше неопределенности, чем кортеж с независимыми символами. Энтропией источника с памятью является следующий предел.

$$H(X) = \lim_{M \to \infty} H_M(X) \tag{13.8}$$

Видим, что энтропия M-кортежа из источника с памятью всегда меньше, чем энтропия источника с тем же алфавитом и вероятностью символов, но без памяти.

$$H_M(M)_{\text{с памятью}} < H_M(M)_{\text{без памяти}} \tag{13.9}$$

Например, известно, что при данном символе (или букве) "q" в английском тексте следующим символом, вероятно, будет "u". Следовательно, в контексте системы связи, если сказать, что буква "u" следует за буквой "q", то это дает незначительную дополнительную информацию о значении слова, которое было передано. Можно привести и другой пример. Наиболее вероятным символом, следующим за буквами "th", может быть один из таких символов: a, e, i, o, u, r и пробел. Таким образом, дополнение следующим символом данного множества разрешает некоторую неопределенность, но не очень сильно. Формальная формулировка сказанного выше: средняя энтропия на символ M-кортежа из источника с памятью *убывает при увеличении длины M*. Следствие: более эффективным является групповое кодирование символов из источника с памятью, а не кодирование их по одному. При кодировании источника размер последовательности символов, рассматриваемой как группа, ограничивается сложностью кодера, ограничениями памяти и допустимой задержкой времени.

Чтобы помочь понять цели, преследуемые при кодировании источников с памятью, построим простые модели этих источников. Одна из таких моделей называется *Марковским источником первого порядка* (first-order Markov source) [1]. Эта модель устанавливает соответствие между множеством состояний (или символов в контексте теории информации) и условными вероятностями перехода к каждому последующему состоянию. В модели первого порядка переходные вероятности зависят только от настоящего состояния. Иными словами, $P(X_{i+1}|X_i, X_{i-1}, \ldots) = P(X_{i+1}|X_i)$. Память модели не распространяется дальше настоящего состояния. В контексте двоичной последовательности это выражение описывает вероятность следующего бита при данном значении текущего бита.

Пример 13.2. Энтропия двоичного источника с памятью

Рассмотрим двоичный (т.е. двухсимвольный) Марковский источник второго порядка, описанный диаграммой состояний, изображенной на рис. 13.1. Источник определен вероятностями переходов состояний $P(0|1)$ и $P(1|0)$, равными 0,45 и 0,05. Энтропия источника X — это взвешенная сумма условных энтропий, соответствующих вероятностям переходов модели.

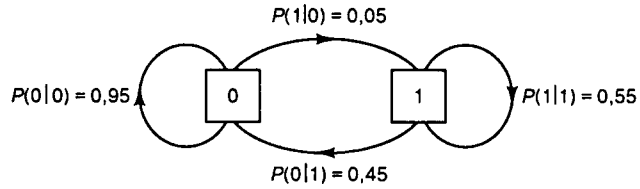


Рис. 13.1. Диаграмма переходов от состояния к состоянию для Марковской модели первого порядка

$$H(X) = P(0)H(X|0) + P(1)H(X|1), \tag{13.10}$$

где

$$H(X|0) = -[P(0|0)\log_2 P(0|0) + P(1|0)\log_2 P(1|0)]$$

и

$$H(X|1) = -[P(0|1)\log_2 P(0|1) + P(1|1)\log_2 P(1|1)].$$

Априорная вероятность каждого состояния находится с помощью формулы полной вероятности.

$$P(0) = P(0|0)P(0) + P(0|1)P(1)$$

$$P(1) = P(1|0)P(0) + P(1|1)P(1)$$

$$P(0) + P(1) = 1$$

Вычисляя априорные вероятности с использованием переходных вероятностей, получим следующее.

$$P(0) = 0{,}9 \text{ и } P(1) = 0{,}1$$

При вычислении энтропии источника с использованием равенства (13.10) получим следующее.

$$H(X) = [P(0)\,H(X|0) + P(1)\,H(X|1)] =$$

$$=(0{,}9)(0{,}286) + (0{,}1)(0{,}993) = 0{,}357 \text{ бит/символ}$$

Сравнивая этот результат с результатом примера 13.1, видим, что источник с памятью имеет энтропию ниже, чем источник без памяти, даже несмотря на то что априорные вероятности символов те же.

Пример 13.3. Коды расширения

Алфавит двоичного Марковского источника (пример 13.2) состоит из 0 и 1, появляющихся с вероятностями 0,9 и 0,1, соответственно. Последовательные символы не являются независимыми, и для использования преимуществ этой зависимости можно определить новое множество кодовых символов — двоичные 2-кортежи (коды расширения).

Двоичные 2-кортежи	Символ расширения	Вероятность символа расширения
00	a	$P(a) = P(0\|0)P(0) = (0{,}95)(0{,}9) = 0{,}855$
11	b	$P(b) = P(1\|1)P(1) = (0{,}55)(0{,}1) = 0{,}055$
01	c	$P(c) = P(0\|1)P(1) = (0{,}45)(0{,}1) = 0{,}045$
11	d	$P(d) = P(1\|0)P(0) = (0{,}05)(0{,}9) = 0{,}045$

Здесь крайняя правая цифра 2-кортежа является самой ранней. Энтропия для этого алфавита кодов расширения находится посредством обобщения равенства (13.10).

$$H(\mathbf{X}_2) = P(a)\,H(\mathbf{X}_2|a) + P(b)\,H(\mathbf{X}_2|b) + P(c)\,H(\mathbf{X}_2|c) + P(d)\,H(\mathbf{X}_2|d)$$

$$H(\mathbf{X}_2) = 0{,}825 \text{ бит/выходной символ}$$

$$H(\mathbf{X}_2) = 0{,}412 \text{ бит/входной символ,}$$

где $\mathbf{X}_k$ — расширение k-го порядка источника $\mathbf{X}$. Более длинный код расширения, использующий преимущества зависимости соседствующих символов, имеет следующий вид.

Двоичный 3-кортеж	Символ расширения	Вероятность символа расширения
000	a	$P(a) = P(0\|00)P(00) = (0{,}95)(0{,}855) = 0{,}8123$
100	b	$P(b) = P(1\|00)P(00) = (0{,}05)(0{,}855) = 0{,}0428$
001	c	$P(c) = P(0\|01)P(01) = (0{,}95)(0{,}045) = 0{,}0428$
111	d	$P(d) = P(1\|11)P(11) = (0{,}55)(0{,}055) = 0{,}0303$
110	e	$P(e) = P(1\|10)P(10) = (0{,}55)(0{,}045) = 0{,}0248$
011	f	$P(f) = P(0\|11)P(11) = (0{,}45)(0{,}055) = 0{,}0248$
010	g	$P(g) = P(0\|10)P(10) = (0{,}45)(0{,}045) = 0{,}0203$
101	h	$P(h) = P(1\|01)P(01) = (0{,}05)(0{,}045) = 0{,}0023$

Используя снова обобщение уравнения (13.10), энтропию для этого кода расширения можно найти как

$$H(\mathbf{X}_3) = 1{,}223 \text{ бит/выходной символ}$$

$$H(\mathbf{X}_3) = 0{,}408 \text{ бит/входной символ.}$$

Отметим, что энтропия односимвольного, двухсимвольного и трехсимвольного описаний источника (0,470, 0,412 и 0,408 бит, соответственно) асимптотически убывает к энтропии источника, равной 0,357 бит/входной символ. Напомним, что энтропия источника — это нижний предел в битах на входной символ для этого алфавита (память бесконечна), и этот предел не может быть достигнут с помощью кодирования конечной длины.

13.1.2. Источники волновых сигналов

Источник волнового сигнала — это случайный процесс некоторой случайной переменной. Считается, что эта случайная переменная — время, так что рассматриваемый волновой сигнал — это изменяющийся во времени волновой сигнал. Важными примерами изменяющихся во времени волновых сигналов являются выходы датчиков, используемых для контроля процессов и описывающих такие физические величины, как температура, давление, скорость и сила ветра. Значительный интерес представляют такие примеры, как речь и музыка. Волновой сигнал может также быть функцией одной или более пространственных величин (т.е. расположение на плоскости с координатами x и y). Важными примерами пространственных волновых сигналов являются единичные зрительные образы, такие как фотография, или движущиеся зрительные образы, такие как последовательные кадры художественного фильма (24 кадра/с). Пространственные волновые сигналы часто преобразуются в изменяющиеся во времени волновые сигналы посредством сканирования. Например, это делается для систем факсимильной связи и передач в формате JPEG, а также для стандартных телевизионных передач.

13.1.2.1. Функции плотности амплитуд

Дискретные источники описываются путем перечисления их возможных элементов (называемых буквами алфавита) и с помощью их многомерных функций плотности вероятности (probability density function — pdf) всех порядков. По аналогии источники волновых сигналов подобным образом описываются в терминах их функций плотности вероятности, а также параметрами и функциями, определенными с помощью этих функций плотности вероятности. Многие волновые сигналы моделируются как случайные процессы с классическими функциями плотности вероятности и простыми корреляционными свойствами. В процессе моделирования различаются краткосрочные, или локальные (временные), характеристики и долгосрочные, или глобальные. Это деление необходимо, так как многие волновые сигналы являются нестационарными.

Функция плотности вероятности реального процесса может быть не известна разработчику системы. Конечно, в реальном времени для короткого предшествующего интервала можно быстро построить выборочные плотности и использовать их как разумные оценки в течение последующего интервала. Менее претенциозная задача — это создание краткосрочных средних параметров, связанных с волновыми сигналами. Эти параметры — выборочное среднее (или среднее по времени), выборочная дисперсия (или среднеквадратическое значение процесса с нулевым средним) и выборочные коэффициенты корреляции, построенные на предшествующем выборочном интервале. При анализе волновых сигналов входной волновой сигнал преобразуется в процесс с нулевым средним путем вычитания его среднего значения. Например, это происходит в устройствах сравнения сигналов, используемых в аналого-цифровых преобразователях, для которых вспомогательная схема измеряет внутренние смещения от уровня постоянного напряжения канала передачи данных и вычитает их в процессе, известном как *автонуль* (autozero). Далее оценка дисперсии часто используется для масштабирования входного волнового сигнала, чтобы сопоставить динамику размаха амплитуды последующего волнового сигнала, обусловленную схемой. Этот процесс, выполняемый при сборе данных, называется *автоматической регулировкой усиления* (automatic gain control — AGC, АРУ). Функцией этих операций, связанных с предварительным формированием сигналов, — вычитание среднего, контроль дисперсии или выравнивание усиления (показанных на рис. 13.2) — является нормирование функций плотности вероятности входного волнового сигнала. Это нормирование обеспечивает оптимальное использование ограниченного динамического диапазона последующих записывающих, передающих или обрабатывающих подсистем.

Многие источники волновых сигналов демонстрируют значительную корреляцию амплитуды на последовательных временны*х* интервалах. Эта корреляция означает, что уровни сигнала на последовательных временны*х* интервалах не являются независимыми. Если временны*е* сигналы независимы на последовательных интервалах, автокорреляционная функция будет импульсной. Многие сигналы, представляющие инженерный интерес, имеют корреляционные функции конечной ширины. Эффективная ширина корреляционной функции (в секундах) называется временем корреляции процесса и подобна временно*й* константе фильтра нижних частот. Этот временно*й* интервал является показателем того, насколько большой сдвиг вдоль оси времени требуется для потери корреляции между данными. Если время корреляции большое, то это значит, что амплитуда волнового сигнала меняется медленно. Наоборот, если время корреляции мало, делаем вывод, что амплитуда волнового сигнала значительно меняется за очень малый промежуток времени.

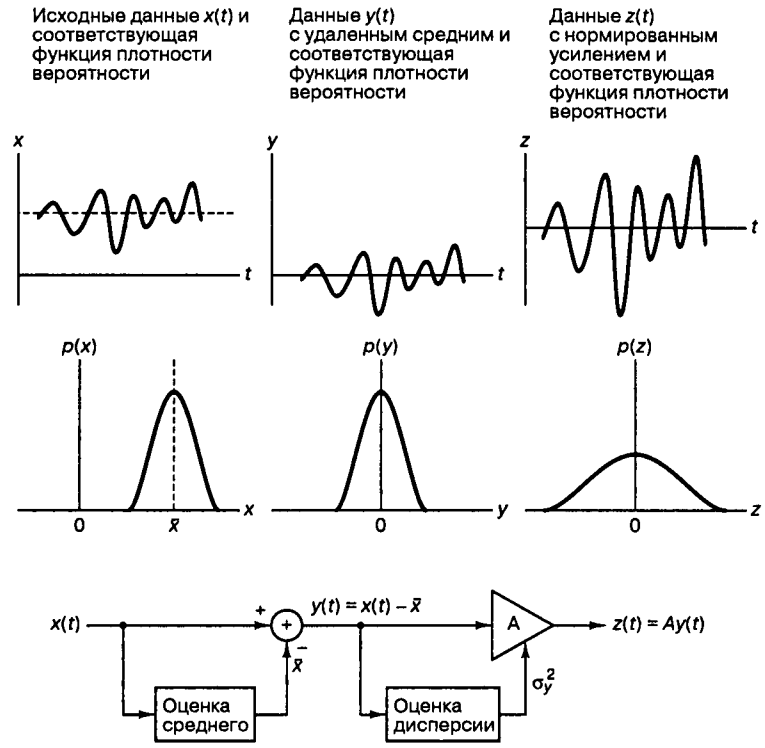


Рис. 13.2. Удаление среднего и нормирование дисперсии (регулировка усиления) для зависимых от данных систем предварительного формирования сигнала

13.2. Квантование амплитуды

Квантование амплитуды — это задача отображения выборок волновых сигналов непрерывной амплитуды в конечное множество амплитуд. Аппаратное обеспечение, которое выполняет отображение, — это аналого-цифровой преобразователь (analog-to-digital converter — ADC, АЦП). Квантование амплитуды происходит после операции выборки-запоминания. Простейшее устройство квантования, которое можно изобразить, выполняет мгновенное отображение с каждого непрерывного входного выборочного уровня в один из предопределенных, равномерно расположенных выходных уровней. Квантующие устройства, которые характеризуются равномерно расположенными приращениями между возможными выходными уровнями, называются *равномерными устройствами квантования*, или *линейными квантующими устройствами*. Возможные мгновенные характеристики входа/выхода легко изображаются с помощью простого ступенчатого графика, подобного изображенному на рис. 13.3. На рис. 13.3, *а*, *б* и *г* представлены устройства с равномерными шагами квантования, а на рис. 13.3, *в* — устройство с неравномерным шагом квантования. На рис. 13.3, *а* характеристика устройства имеет нуль в центре шага квантования, а на рис. 13.3, *б* и *г* — на границе шага квантования. Отличительная особенность устройств, имеющих характеристики с нулем в центре шага квантования и характеристики с нулем на границе шага квантования, связана, соответственно, с наличием или отсутствием выходных

изменений уровня, если входом квантующего устройства является шум низкого уровня. На рис. 13.3, *г* представлено смещенное (т.е. усекающее) устройство квантования, а другие устройства, изображенные на рисунке, являются несмещенными и называются *округляющими*. Такие несмещенные устройства квантования представляют собой идеальные модели, но в аналого-цифровых преобразователях округление не реализуется никогда. Как правило, устройства квантования реализуются как усекающие преобразователи. Термины "характеристика с нулем в центре шага квантования" (midtread) или "характеристика с нулем на границе шага квантования" (midriser) относятся к ступенчатым функциям и используются для описания того, имеются ли в начале координат горизонтальная или вертикальная составляющая ступенчатой функции. Пунктирная линия единичного наклона, проходящая через начало координат, представляет собой неквантованную характеристику входа/выхода, которую пытаются аппроксимировать ступенчатой функцией. Разность между ступенчатой функцией и отрезком линии единичного наклона представляет собой ошибку аппроксимации, допускаемую устройством квантования на каждом входном уровне. На рис. 13.4 показана ошибка аппроксимации амплитуды в сравнении с входной амплитудой функции для каждой из характеристик квантующего устройства, изображенных на рис. 13.3. Рис. 13.4 соответствует рис. 13.3. Часто эта ошибка моделируется как шум квантования, поскольку последовательность ошибок, полученная при преобразовании широкополосного случайного процесса, напоминает аддитивную последовательность шума. Однако, в отличие от действительно аддитивных источников шума, ошибки преобразования являются сигнально зависимыми и высоко структурированными. Желательно было бы нарушить эту структуру, что можно сделать путем введения независимых шумовых преобразований, известных как *псевдослучайный шум*, предшествующих шагу преобразования. (Эта тема обсуждается в разделе 13.2.4.)

Линейное устройство квантования легко реализовать и очень легко понять. Оно представляет собой универсальную форму квантующего устройства, поскольку не предполагает никаких знаний о статистике амплитуд и корреляционных свойствах входного волнового сигнала, а также не использует преимуществ требований к точности, предоставляемых пользователями. Устройства квантования, которые используют указанные преимущества, являются более эффективными как кодеры источника и предназначены для более специфических задач, чем общие линейные устройства квантования. Эти квантующие устройства являются более сложными и более дорогими, но они оправдывают себя с точки зрения улучшения производительности системы. Существуют приложения, для которых равномерные устройства квантования являются наиболее желаемыми преобразователями амплитуды. Это — приложения обработки сигналов, графические приложения, приложения отображения изображений и контроля процессов. Для некоторых иных приложений более приемлемыми преобразователями амплитуды являются неравномерные адаптивные квантующие устройства. Эти устройства включают в себя кодеры волнового сигнала для эффективного запоминания и эффективной связи, контурные кодеры для изображений, векторные кодеры для речи и аналитические/синтетические кодеры (такие, как вокодер) для речи.

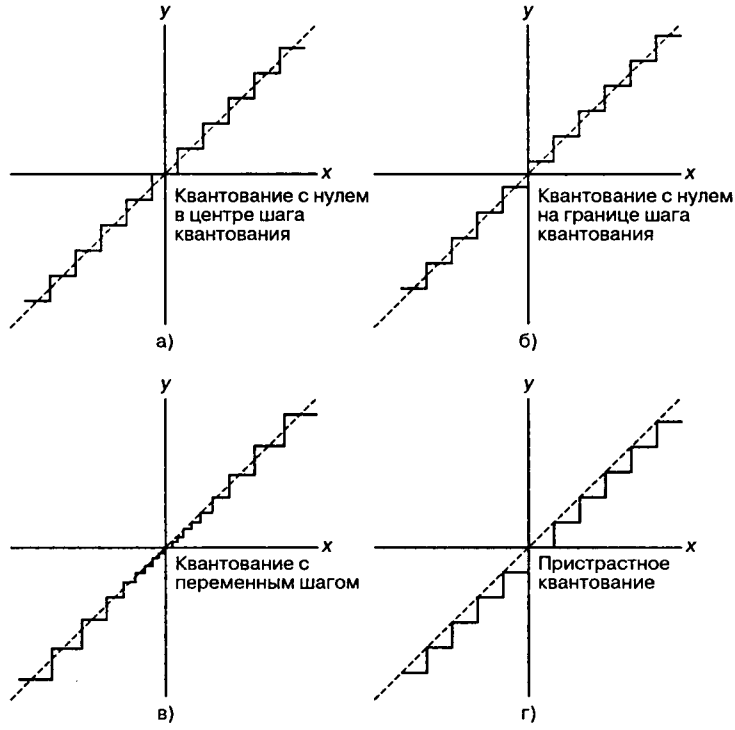


Рис. 13.3. Различные передаточные функции устройства квантования

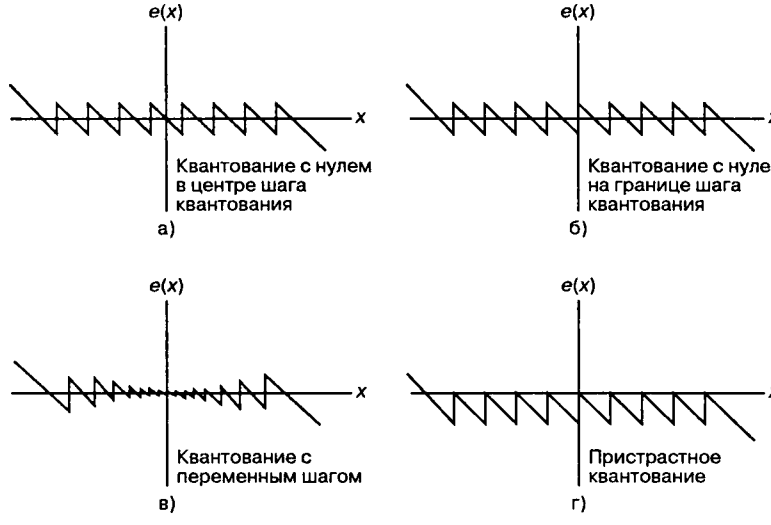


Рис. 13.4. Мгновенная ошибка для различных передаточных функций устройства квантования

13.2.1. Шум квантования

Разность между входом и выходом преобразователя называется *ошибкой квантования* (quantizing error). На рис. 13.5 изображен процесс отображения входной последовательности $x(t)$ в квантованную выходную последовательность $\hat{x}(t)$. Получение $\hat{x}(t)$ можно представить как сложение каждого $x(t)$ с ошибочной последовательностью $e(t)$.

$$\hat{x}(t) = x(t) + e(t)$$

Ошибочная последовательность $e(t)$ детерминированно определяется входной амплитудой через зависимость мгновенной ошибки от амплитудной характеристики, изображенной на рис. 13.4. Отметим, что ошибочная последовательность демонстрирует две различные характеристики в различных входных рабочих областях.

Первым рабочим интервалом является гранулированная область ошибок, соответствующая подаче на вход пилообразной характеристики ошибки. Внутри этого интервала квантующие устройства ограничены размерами соседних ступенчатых подъемов. Ошибки, которые случаются в этой области, называются *гранулированными* (granular errors), или иногда *ошибками квантования* (quantizing error). Входной интервал, для которого ошибки преобразования являются гранулированными, определяет динамическую область преобразователя. Этот интервал иногда называется *областью линейного режима* (region of linear operation). Соответствующее использование квантующего устройства требует, чтобы условия, порожденные входным сигналом, приводили динамическую область входного сигнала в соответствие с динамической областью устройства квантования. Этот процесс является функцией сигнально зависимой системы регулировки усиления, называемой *автоматической регулировкой усиления* (automatic gain control — AGC, АРУ), которая показана на пути прохождения сигнала на рис. 13.5.

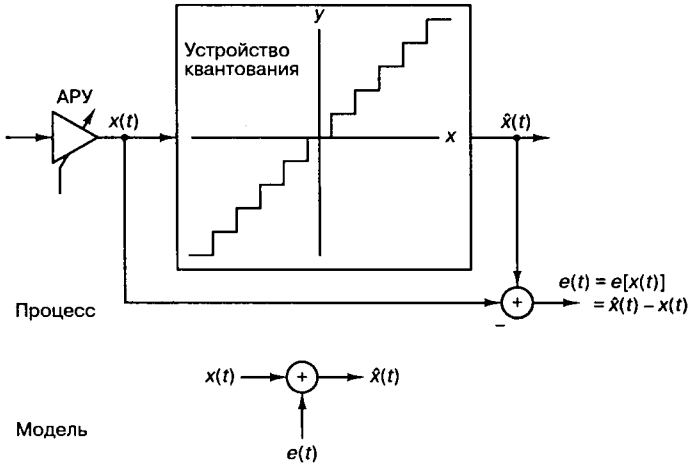


Рис. 13.5. Процесс и модель повреждения входного сигнала шумом квантования

Вторым рабочим интервалом является негранулированная область ошибок, соответствующая линейно возрастающей (или убывающей) характеристике ошибки. Ошибки, которые происходят в этом интервале, называются ошибками *насыщения* (saturation error) или *перегрузки* (overload error). Когда квантующее устройство работает в этой области, говорят, что преобразователь *насыщен*. Ошибки насыщения больше,

чем гранулированные ошибки, и могут оказывать большее нежелательное влияние на точность воспроизведения информации.

Ошибка квантования, соответствующая каждому значению входной амплитуды, представляет слагаемое ошибки или шума, связанное с данной входной амплитудой. Если интервал квантования мал в сравнении с динамической областью входного сигнала и входной сигнал имеет гладкую функцию плотности вероятности в интервале квантования, можно предположить, что ошибки квантования равномерно распределены в этом интервале, как изображено на рис. 13.6. Функция плотности вероятности с нулевым средним соответствует округляющему квантующему устройству, в то время как функция плотности вероятности со средним $-q/2$ соответствует усекающему квантующему устройству.

Квантующее устройство, или аналого-цифровой преобразователь (analog-to-digital converter — ADC, АЦП), определяется числом, размером и расположением своих уровней квантования (или границами шагов и соответствующими размерами шагов). В равномерном квантующем устройстве размеры шагов равны и расположены на одинаковом расстоянии. Число уровней N обычно является степенью 2 вида $N = 2^b$, где b — число бит, используемых в процессе преобразования.

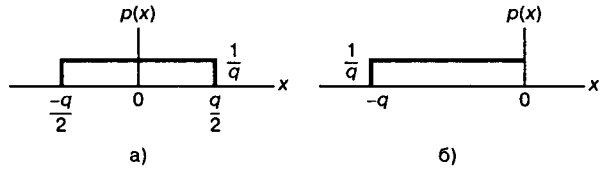


Рис. 13.6. Функции плотности вероятности для ошибки квантования, равномерно распределенной в интервале квантили, q: а) функция плотности вероятности для округляющего преобразователя; б) функция плотности вероятности для усекающего преобразователя

Это число уровней равномерно распределено в динамической области возможных входных уровней. Обычно этот интервал определяется как $\pm E_{max}$, подобно ±1,0 В или ±5,0 В. Таким образом, для полного интервала $2E_{max}$ величину шага преобразования получим в следующем виде.

$$q = \frac{2E_{max}}{2^b} \tag{13.11}$$

В качестве примера использования равенства (13.11) шаг квантования (в дальнейшем называемый *квантилью*) для 10-битового преобразователя, работающего в области ±1,0V, равен 1,953 мВ. Иногда рабочая область преобразователя изменяется так, что квантиль является "целым" числом. Например, изменение рабочей области преобразователя до ±1,024 В приводит к шагу квантования, равному 2,0 мВ. Полезным параметром равномерного квантующего устройства является его выходная дисперсия. Если предположить, что ошибка квантования равномерно распределена в отдельном интервале ширины q, дисперсия квантующего устройства (которая представляет собой шум квантующего устройства или мощность ошибки) для ошибки с нулевым средним находится следующим образом.

$$\sigma^2 = \int_{-q/2}^{q/2} e^2 p(e)de = \int_{-q/2}^{q/2} e^2 \frac{1}{q} de = \frac{q^2}{12}, \tag{13.12}$$

где $p(e) = 1/q$ в интервале q — это функция плотности вероятности (probability density function — pdf) ошибки квантования e. Таким образом, среднеквадратическое значение шума квантования в интервале квантили ширины q равно $q\sqrt{12}$ или $0{,}29q$. Уравнение (13.12) определяет мощность шума квантования в интервале размером в одну квантиль в предположении, что ошибки равновероятны в пределах интервала квантования. Если включить в рассмотрение работу в интервале насыщения квантующего устройства или рассмотреть неравномерные устройства квантования, то получим, что интервалы квантования не имеют равной ширины внутри области изменения входной переменной и плотность амплитуды не является равномерной внутри интервала квантования. Можно вычислить эту зависящую от амплитуды энергию ошибки σ_q^2, усредняя квадраты ошибок по амплитуде переменной, взвешенной вероятностью этой амплитуды. Это можно выразить следующим образом.

$$\sigma_q^2 = \mathbf{E}\{[X - q(x)]^2\} = \int_{-\infty}^{\infty} e^2(x)p(x)dx\,, \tag{13.13}$$

где x — входная переменная, $q(x)$ — ее квантованная версия, $e(x) = x - q(x)$ — ошибка, а $p(x)$ — функция плотности вероятности амплитуды x. Интервал интегрирования в формуле (13.13) можно разделить на два основных интервала: один отвечает за ошибки в ступенчатой или линейной области квантующего устройства, а второй — за ошибки в области насыщения. Определим амплитуду насыщения квантующего устройства как E_{max}. Предположим также, что передаточная функция квантующего устройства есть четно-симметричной и такой же является функция плотности вероятности для входного сигнала. Мощность ошибки σ_q^2, определенная равенством (13.13), является полной мощностью ошибки, которая может быть разделена следующим образом.

$$\sigma_q^2 = 2\int_0^{\infty} e^2(x)p(x)dx = \tag{13.14,а}$$

$$= 2\int_0^{E_{max}} e^2(x)p(x)dx + 2\int_{E_{max}}^{\infty} e^2(x)p(x)dx = \tag{13.14,б}$$

$$= \sigma_{Lin}{}^2 + \sigma_{Sat}{}^2$$

Здесь $\sigma_{Lin}{}^2$ — мощность ошибки в линейной области, а $\sigma_{Sat}{}^2$ — мощность ошибки в области насыщения. Мощность ошибки $\sigma_{Lin}{}^2$ может быть далее разделена на подынтервалы, соответствующие последовательным дискретным входным уровням квантующего устройства (т.е. квантилям). Если предположить, что существует N таких уровней квантили, интеграл превращается в следующую сумму.

$$\sigma_{Lin}{}^2 = 2\sum_{n=0}^{N/2-1}\int_{x_n}^{x_{n+1}} e^2(x)p(x)dx\,, \tag{13.15}$$

где x_n — уровень квантующего устройства, а интервал или шаг между двумя такими уровнями называется *интервалом квантили* (quantile interval). Напомним, что N, как пра-

вило, является степенью 2. Таким образом, существует $N/2-1$ положительных уровней, $N/2-1$ отрицательных уровней и нулевой уровень — всего $N-1$ уровень и $N-2$ интервала. Теперь, если аппроксимировать плотность на каждом интервале квантили константами $q_n = (x_{n+1} - x_n)$, выражение (13.15) упростится до следующего вида.

$$\sigma_{\text{Lin}}{}^2 = 2\sum_{n=0}^{N/2-1} \frac{x^3}{3}\Bigg|_{x=-q_n/2}^{x=+q_n/2} p(x_n) = \\ = 2\sum_{n=0}^{N/2-1} \frac{q_n{}^2}{12} p(x_n) q_n \tag{13.16}$$

где $e(x)$ в равенстве (13.15) было заменено x из (13.16), поскольку $e(x)$ — линейная функция от x, имеющая единичный наклон и проходящая через нуль в центре каждого интервала. Кроме того, пределы интегрирования в равенстве (13.15) были заменены в соответствии с изменениями x внутри интервала квантили. Поскольку область изменения была обозначена через q_n, нижний и верхний пределы могут быть обозначены как $x = -q_n/2$ и $x = +q_n/2$. Равенство (13.16) описывает мощность ошибки в линейной области в виде суммы мощности ошибки $q_n^2/12$ в каждом интервале квантили, взвешенной вероятностью $p(x_n)q_n$ этой энергии ошибки.

13.2.2. Равномерное квантование

Если устройство квантования имеет равномерно расположенные квантили, равные q, и все интервалы равновероятны, выражение (13.16) упрощается далее.

$$\sigma_{\text{Lin}}{}^2 = \frac{2}{12}\sum_{n=0}^{N/2-1} q_n{}^2 p(x_n) q_n = \frac{2}{12}\sum_{n=0}^{N/2-1} q^2 \frac{1}{q(N-2)} q = \frac{q^2}{12} \tag{13.17}$$

Если квантующее устройство работает не в области насыщения (мощности шума квантования), тогда $\sigma_q^2 = \sigma_{\text{Lin}}^2$, и эти величины часто используются как взаимозаменяемые. Отметим, что мощность шума сама по себе не будет полно описывать поведение шума устройства квантования. Более полной мерой качества является отношение второго центрального момента (дисперсии) шума квантования к входному сигналу. Если предположить, что входной сигнал имеет нулевое среднее, дисперсия сигнала равна следующему.

$$\sigma_x^2 = \int_{-\infty}^{\infty} x^2 p(x) dx \tag{13.18}$$

Дальнейшее изучение среднего шума квантующего устройства требует конкретизации функции плотности и устройства.

Пример 13.4. Равномерное квантующее устройство

Определим дисперсию устройства квантования и отношение мощности шума к мощности сигнала (noise-to-signal power ratio — NSR) для равномерно распределенного в полной динамической области сигнала, созданного устройством квантования с 2^b расположенными на одинаковых расстояниях уровнями квантили. В этом случае шума насыщения не существует и должна быть вычислена только величина линейного шума. Каждый интервал квантили равен следующему.

$$q = (2E_{\max})2^{-b} \tag{13.19}$$

Здесь $2E_{\max}$ — это входной интервал между положительной и отрицательной границами линейной области квантования.

Решение

Подставляя выражение (13.19) в формулу (13.12) или (13.17), получим следующую мощность шума квантования (в линейной области).

$$\sigma_q^2 = \frac{1}{12}(2E_{\max}2^{-b})^2 = \frac{1}{12}(2E_{\max})^2 2^{-2b} \tag{13.20}$$

Мощность входного сигнала находится посредством интегрирования выражения (13.18) для равномерной плотности вероятности в интервале длины $2E_{\max}$ с центром в точке 0, так что $p(x) = 1/(2E_{\max})$, и дисперсия сигнала находится следующим образом.

$$\sigma_x^2 = \int_{-E_{\max}}^{+E_{\max}} \frac{1}{2E_{\max}} x^2 dx = \frac{1}{12}(2E_{\max})^2 \tag{13.21}$$

Рассматривая отношение мощности шума к мощности сигнала (NSR), получим следующее.

$$\text{NSR} = \frac{\sigma_q^2}{\sigma_x^2} = 2^{-2b} \tag{13.22}$$

Теперь, превращая NSR в децибелы, получим следующее.

$$\text{NSR}_{\text{дБ}} = 10\lg(\text{NSR}) = 10\lg(2^{-2b}) = \tag{13.23,а}$$

$$= -20b\lg(2) = -6{,}02b(\text{дБ}) \tag{13.23,б}$$

Выражение (13.23, б) свидетельствует о том, что за каждый бит, который используется в процессе преобразования, мы платим –6,02 дБ отношения шума к сигналу. Действительно, NSR для любого равномерного квантующего устройства, не работающего в области насыщения, имеет следующий вид.

$$\text{NSR}_{\text{дБ}} = -6{,}02b + C \tag{13.24}$$

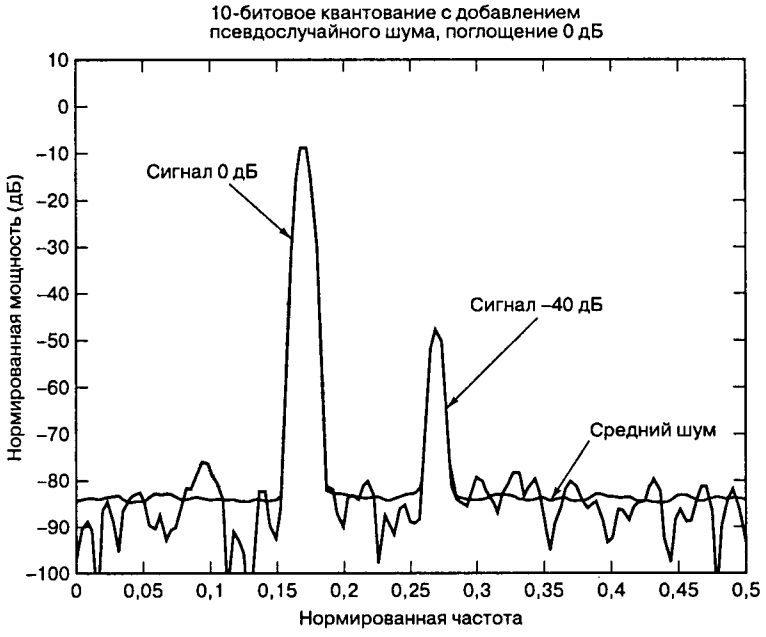
Здесь член C зависит от функции плотности вероятности сигнала (probability density function — pdf); он положителен для функций плотности, являющихся узкими по отношению к уровню насыщения преобразователя.

13.2.2.1. Сигнал и шум квантования в частотной области

До настоящего момента шум квантования обсуждался с точки зрения его влияния на выборку временного ряда, представляющую дискретный сигнал. Шум квантования может быть также описан в частотной области; это позволяет взглянуть на влияние условий работы, что и будет сделано ниже. В процессе этого изучения предполагается также рассмотрение насыщения (раздел 13.2.3), возмущения (раздел 13.2.4) и квантующих устройств с обратной связью по шуму (раздел 13.2.6).

На рис. 13.7 представлено дискретное преобразование Фурье двух синусоид, которые были образованы линейным 10-битовым АЦП. Сравнительные амплитуды двух синусоид равны 1,0 и 0,01 (т.е. одна на 40 дБ ниже другой). На рис. 13.7, *а* сигнал низкой частоты (обозначенный 0 дБ) масштабируется на 1 дБ ниже полной динамической области 10-битового квантующего устройства, которую для удобства будем считать единичной. Отметим, что на рис. 13.7, *а* полномасштабный сигнал 0 дБ находит-

ся на 6 дБ ниже входного уровня поглощения 1 дБ. Это объясняется наличием множителя 1/2 в спектральном разложении действительного сигнала по всем ненулевым частотам. Среднее отношение сигнала к шуму квантования (SNR) для 10-битового квантующего устройства равно 60 + C дБ. Для полномасштабной синусоиды константа C равна 1,76 дБ, что делает суммарное отношение SNR примерно равным 62 дБ. При дискретном преобразовании Фурье (discrete Fourier transform — DFT, ДПФ), которое выполнялось для получения графика на рис. 13.7, длина равнялась 256. Поскольку отношение SNR преобразования увеличивается пропорционально длине преобразования (или времени интегрирования), то благодаря преобразованию SNR улучшается на 24 дБ [2] с потерей 3,0 дБ вследствие усечения. Таким образом, на выходе преобразования вершина SNR вследствие квантования равна 62 + 24 – 3 = 83 дБ. Шумовой сигнал на каждой частоте ДПФ может быть представлен как квадратный корень из суммы квадратов гауссовых случайных величин, которая описывается как случайная величина, имеющая распределение хи-квадрат с двумя степенями свободы. Дисперсия (мощность шума) равна квадрату среднего. Таким образом, имеем значительные колебания вокруг математического ожидания уровня мощности шума. Для получения устойчивой оценки нижнего уровня шума нам потребуется среднее по ансамблю. Видно, что нижний уровень шума (получен с помощью 400 средних) равен –83 дБ. К сигналу перед квантованием был добавлен псевдослучайный шум (описанный в разделе 13.2.4), чтобы рандомизировать ошибки квантования. На рис. 13.7, *б* и *в* входные сигналы ослабляются относительно полномасштабного входа на 20 и 40 дБ. Это ослабление увеличивает константу C в формуле (13.24) на 20 и 40 дБ, что проявляется как уменьшение спектральных уровней входных синусоид на эти же величины. Отметим, что входной сигнал наивысшей частоты (рис. 13.7, *в*), который теперь уменьшился на 80 дБ относительно полной шкалы, располагается на 3 дБ ниже среднего уровня шума преобразователя. Синусоида самой низкой частоты на рис. 13.7, *в* теперь ослаблена на 40 дБ относительно полной шкалы, поэтому характеризуется SNR на 40 дБ меньшим, чем для сигнала на рис. 13.7, *а*.



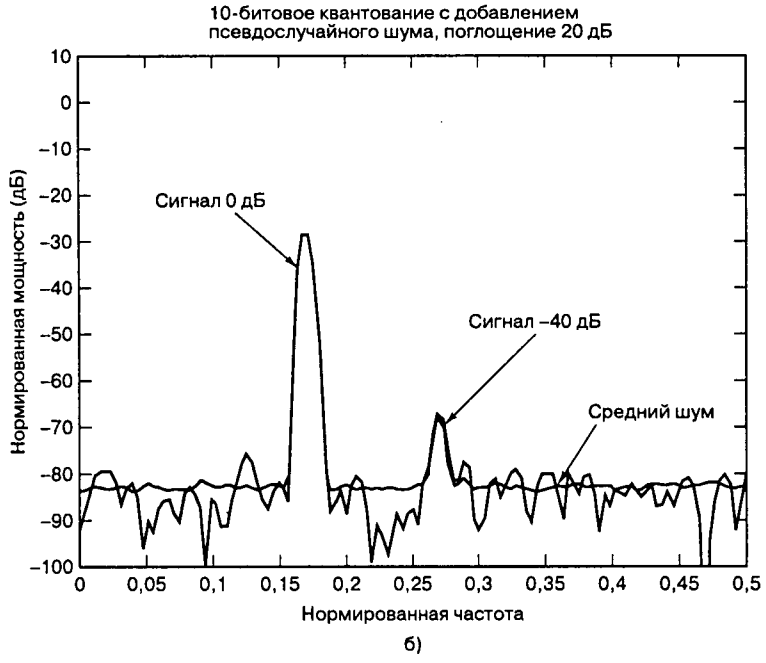


Рис. 13.7. Энергетический спектр сигналов, квантованных равномерным АЦП

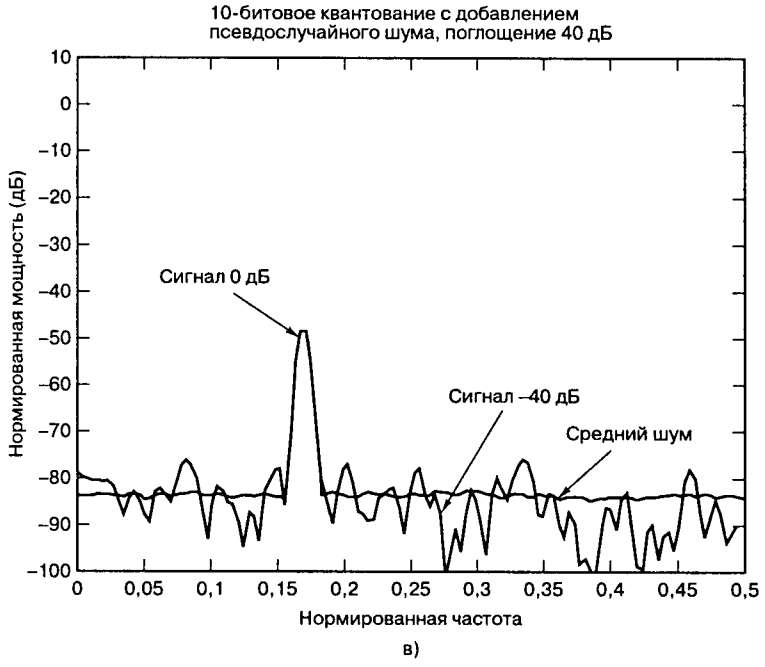


Рис. 13.7. Энергетический спектр сигналов, квантованных равномерным АЦП (окончание)

При минимизации среднего отношения шума к сигналу квантования мы сталкиваемся с противоречием в требованиях. С одной стороны, желательно удерживать сигналы большими по отношению к интервалу квантования q с целью получения большого SNR. С другой стороны, необходимо удерживать сигнал малым, чтобы избежать насыщения квантующего устройства. Противоречивые требования разрешаются путем масштабирования входного сигнала; в результате его среднеквадратическое значение представляет собой заданную долю полномасштабной области значений квантующего устройства. Указанная доля выбирается так, чтобы согласовать ошибки насыщения (взвешенные вероятностями их появления) с ошибками квантования (взвешиваются аналогично) и таким образом достигнуть минимального отношения шума к сигналу. Положение этой желательной рабочей точки преобразователя обсуждается в следующем разделе.

13.2.3. Насыщение

На рис. 13.8 представлено среднее NSR равномерного квантующего устройства как функция отношения уровня насыщения квантующего устройства к среднеквадратическому значению сигнала. На рисунке изображены отношения NSR сигналов с тремя различными функциями плотности вероятности: арксинус (синусообразная плотность сигнала), равномерная и гауссова.

По оси абсцисс (рис. 13.8) отложено отношение уровня насыщения квантующего устройства к среднеквадратическому уровню входного сигнала. При каждой из трех плотностей для фиксированного числа бит существует значение абсциссы, соответствующее минимуму NSR. Другими словами, для данной входной плотности можно оп-

ределить уровень входного сигнала (связанный с насыщением), при котором достигается минимум NSR.

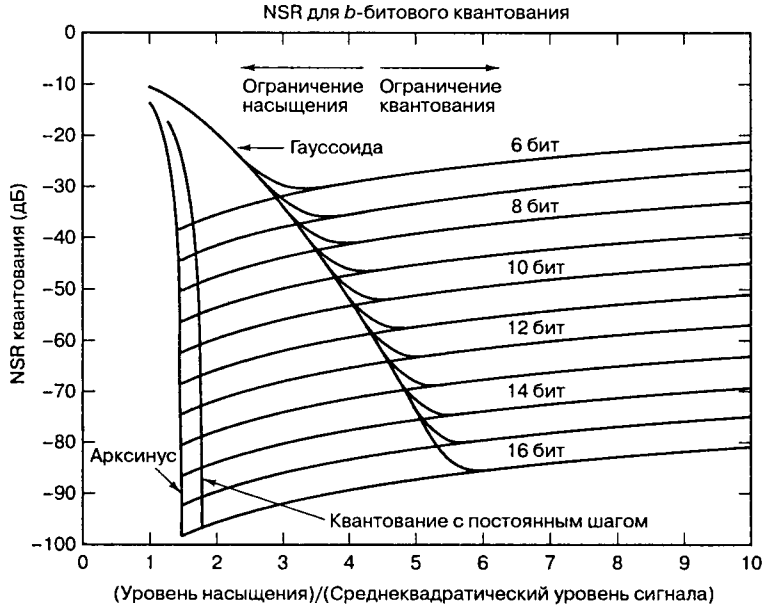


Рис. 13.8. Отношение NSR аналого-цифрового преобразователя в сравнении с отношением уровня насыщения АЦП к среднеквадратическому уровню сигнала

Уменьшенные уровни входных сигналов соответствуют большим значениям NSR на оси абсцисс и представляют собой движение вправо. Увеличенные уровни входных сигналов также соответствуют большим значениям NSR на оси абсцисс и представляют собой движение влево. Это увеличение происходит вследствие работы в области насыщения устройства квантования. Отметим, что скорость изменения отношения NSR при движении влево от оптимальной рабочей точки выше, чем при движении вправо. Например, это, в частности, верно для равномерной плотности и плотности типа арксинуса. Это свидетельствует о том, что шум насыщения более нежелателен, чем линейный шум квантования. Как следствие, если допустить ошибку в определении рабочей точки, называемой *точкой атаки* квантующего устройства, то будет лучше иметь ошибку на стороне превышения поглощения, чем на стороне недостаточного поглощения входного сигнала. Начало насыщения происходит в точках с различными значениями абсциссы. Для синусообразного сигнала (плотность типа арксинуса) это происходит примерно в точке $\sqrt{2}$. Для треугольных сигналов (равномерная плотность) это случается примерно в точке $\sqrt{3}$. Для шумоподобных сигналов (гауссова плотность), когда уровень сигнала сокращается относительно насыщения, насыщение происходит непрерывно, с убывающей вероятностью. Рассмотрим в качестве примера 10-битовый АЦП, имеющий отношение NSR –60 дБ для равномерной плотности при работе на вершине насыщения и NSR –62 дБ для плотности типа арксинуса при работе на вершине насыщения. С другой стороны, тот же 10-битовый преобразователь имеет минимум NSR приблизительно в точке –52 дБ для

всех плотностей, когда среднеквадратический уровень равен 1/4 уровня насыщения (точка 4 на оси абсцисс). Данный рисунок иллюстрирует, что шум насыщения более опасен, чем шум квантования. Этому можно дать достаточно простое объяснение, изучив мгновенную характеристику ошибки (как показано на рис. 13.4) и отметив, что ошибки насыщения очень велики в сравнении с ошибками квантования. Таким образом, малое насыщение, даже если оно случается нечасто, будет вносить большой вклад в средний уровень шума квантующего устройства.

Шум насыщения и шум квантования отличаются несколько по-иному. Шум квантования приближается к белому шуму. По этой причине к аналоговому сигналу до квантования могут намеренно добавляться сигналы псевдослучайного шума. Отметим, что шум насыщения подобен белому шуму только тогда, когда входной сигнал имеет широкую полосу частот и может быть гармонически связанным с входным сигналом, если тот имеет узкую полосу частот. Таким образом, влияние шума квантования может быть отфильтровано или усреднено, так как по характеристикам — это белый шум. С другой стороны, шум насыщения неотличим от содержимого полезного сигнала и в общем случае не может быть устранен с помощью последовательного усреднения или фильтрующих технологий.

На рис. 13.9 представлены дискретные преобразования Фурье того же сигнального множества, что и на рис. 13.8, квантованного 10-битовым АЦП. Кроме того, на рис. 13.9 пиковая амплитуда сигнала выбрана так, чтобы на 10% (0,83 дБ) превышать уровень насыщения АЦП. Отметим, что некоторые спектральные составляющие больше, чем сигнал в –40 дБ. Эти составляющие (шум насыщения) будут возрастать еще больше, когда отклонения сигнала будут идти глубже в режим насыщения. Чтобы увидеть существенную разницу во влиянии слишком слабого поглощения сигнала (следовательно, имеем насыщение) на выход шума АЦП, сравните этот рисунок с рис. 13.7.

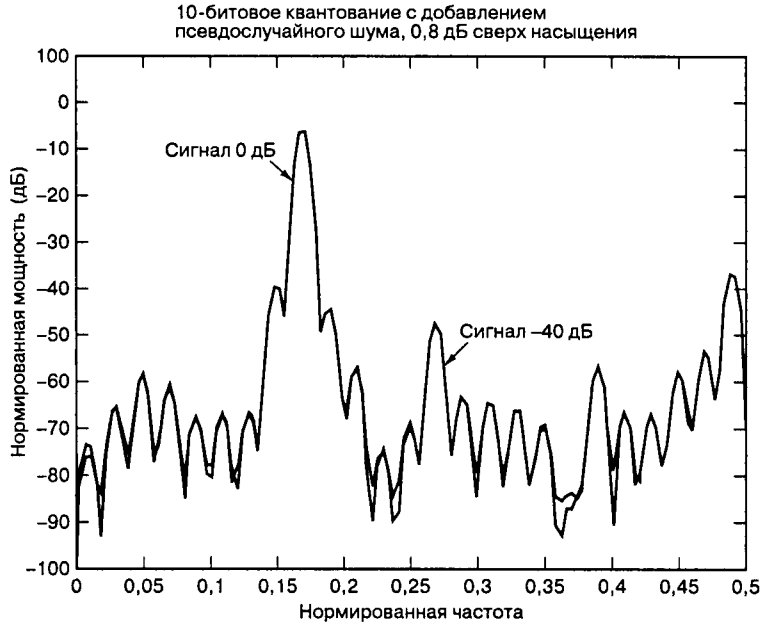


Рис. 13.9. Энергетический спектр равномерно квантованных сигналов с насыщением квантующего устройства на пиках сигнала в 0,8 дБ вне полномасштабного входного уровня

13.2.4. Добавление псевдослучайного шума

Добавление псевдослучайного шума представляет собой одно из самых разумных применений шума как полезного инженерного инструмента. Псевдослучайный шумовой сигнал — это небольшое возмущение или помеха, добавленные к измеряемому процессу, чтобы ограничить влияние малых локальных нелинейностей. Наиболее знакомой формой псевдослучайного шума является встряхивание компаса перед собственно его использованием. В данном случае имеем последовательность малых импульсов, применяемую для вывода движения стрелки из локальной области, которая имеет нелинейный коэффициент трения при малых скоростях. Более сложным примером того же эффекта является механическое псевдослучайное возмущение, применяемое к вращающимся лазерным лучам лазерного лучевого гироскопа с целью вывода гироскопа из ловушки низкоуровневой частоты, известной как *мертвая полоса* [3].

В случае аналого-цифрового преобразователя цель псевдослучайного шума — ограничить (или избежать) локальные разрывы (т.е. подъемы и ступени) мгновенной передаточной функции входа/выхода. Чтобы лучше представить себе влияние этих разрывов, можно перечислить ожидаемые свойства ошибочной последовательности, образованной процессом квантования, с последующим изучением действительных свойств той же последовательности. Ошибочная последовательность квантующего устройства моделируется как аддитивный шум. Давайте рассмотрим ожидаемые свойства такой последовательности шума.

1. Нулевое среднее $\mathbf{E}\{e(n)\} = 0$
2. Белый шум $\mathbf{E}\{e(n)e(n+m)\} = \sigma^2\delta(m)$
3. Отсутствие корреляции с данными x(n) $\mathbf{E}\{e(n)x(n+m)\} = 0$

В данном случае m и n — выборочные индексы, $\delta(m)$ — дельта-функция Дирака. Изучение рис. 13.10, на котором представлена последовательность выборок, образованная усекающим АЦП, позволяет сделать следующие наблюдения.

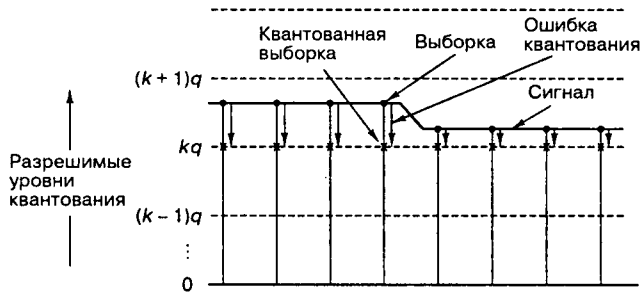


Рис. 13.10. Последовательность дискретных данных квантуется в ближайшие наименьшие уровни квантили посредством присвоенной ошибочной последовательности

1. Вся ошибочная последовательность имеет одну и ту же полярность; следовательно, ее среднее не равно нулю.
2. Последовательность не является независимой при переходе от выборки к выборке; следовательно, она не является белым шумом.
3. Последовательность ошибки коррелирует с входом; следовательно, она не является независимой.

Повторяющиеся измерения того же сигнала будут давать в результате тот же шум, и, таким образом, усреднение ни по какому числу измерений не уменьшит отклонение от истинного входного сигнала. Парадоксально, но мы хотели бы видеть этот шум "более шумным". Если шум является независимым на последовательных измерениях, усреднение будет сокращать отклонение от истинных значений. Таким образом, столкнувшись с проблемой, что получаемый шум не является тем шумом, который нам необходим, выбираем возможность изменить этот шум, добавляя к нему наш собственный. Измерения дополняются возмущением, чтобы превзойти нежелательный низкоуровневый шум устройства квантования. Дополненное возмущение в известном смысле преобразует *плохой шум в хороший* [4].

Пример 13.5. Линеаризация с помощью псевдослучайного шума

Предположим, рассматриваются квантующие устройства, которые могут измерять только целые величины и превращать входные данные в наименьшие ближайшие целые — процесс, называемый *усечением*. Сделано 10 измерений сигнала, скажем, амплитуды 3,7. При отсутствии добавочного сигнала все замеры равны 3,0. Теперь перед измерениями добавим к входной последовательности равномерно распределенную (на интервале от 0 до 1) случайную числовую последовательность. Последовательность данных имеет следующий вид.

Измерение	Необработанный сигнал	Квантованный необработанный сигнал	Псевдослучайный шум	Суммарный сигнал	Квантованный суммарный сигнал
1	3,7	3,0	0,3485	4,0485	4,0
2	3,7	3,0	0,8685	4,5685	4,0
3	3,7	3,0	0,2789	3,9789	3,0
4	3,7	3,0	0,3615	4,0615	4,0
5	3,7	3,0	0,1074	3,8074	3,0
6	3,7	3,0	0,2629	3,9629	3,0
7	3,7	3,0	0,9252	4,6252	4,0
8	3,7	3,0	0,5599	4,2599	4,0
9	3,7	3,0	0,3408	4,0408	4,0
10	3,7	3,0	0,5228	4,2228	4,0
Средние =		3,0	0,4576	4,1576	3,7
Среднее псевдослучайного шума				0,4576	
Среднее суммарного сигнала – среднее псевдослучайного шума				3,7	

В этом примере для удаления смещения квантующего устройства был использован смещенный псевдослучайный шум. Среднее суммированных и преобразованных измерений (при наличии корректного измерения) в общем случае будет ближе к истинному сигналу, чем несуммированные с псевдослучайным шумом и преобразованные измерения [5, 6].

Чтобы проиллюстрировать влияние процесса добавления псевдослучайного шума на процесс квантования изменяющегося во времени сигнала, рассмотрим следующий эксперимент. Пусть синусоидальный сигнал, имеющий амплитуду 1,0, подавляется на 60 дБ. Тогда ослабляемый сигнал имеет полную амплитуду 0,001, что составляет примерно половину интервала квантования, равного 0,001957, для десятибитового равномерного устройства квантования (получается делением удвоенной амплитуды сигнала 2 на $2^{10} - 2$). Когда на округляющее квантующее устройство подается ослабленная синусоида, на выходе будут получаться в основном все нули, за исключением отдельных единиц в ±1 квантиль, что происходит в том случае, когда вход пересекает уровень $\pm q/2$, равный 0,000979

(соответствующий наименее значимому биту АЦП). Если входной сигнал ослаблен еще на 0,23 дБ, пороговые уровни самого младшего бита никогда не будут пересекаться и выходная последовательность будет представлять собой все нули. Теперь добавим псевдослучайный шум со среднеквадратической амплитудой, равной 0,001, к ослабленной синусоиде амплитуды 0,001 так, чтобы сумма сигнала с псевдослучайным шумом регулярно пересекала уровни $\pm q/2$ АЦП. На рис. 13.11 изображен спектр мощности, полученный путем преобразования и усреднения 400 реализаций этого суммарного сигнала. В результате ослабленный на 60 дБ сигнал на пределе разрешающей способности АЦП все еще присутствовал и, будучи точно измеренным, составил –63 дБ (–3 дБ вследствие округления). Псевдослучайный шум давал эффект расширения динамической области АЦП (как правило, с 9 до 12 дБ или с 1,5 до 2,0 бит) и повысил эффективность ступенчатой аппроксимации АЦП.

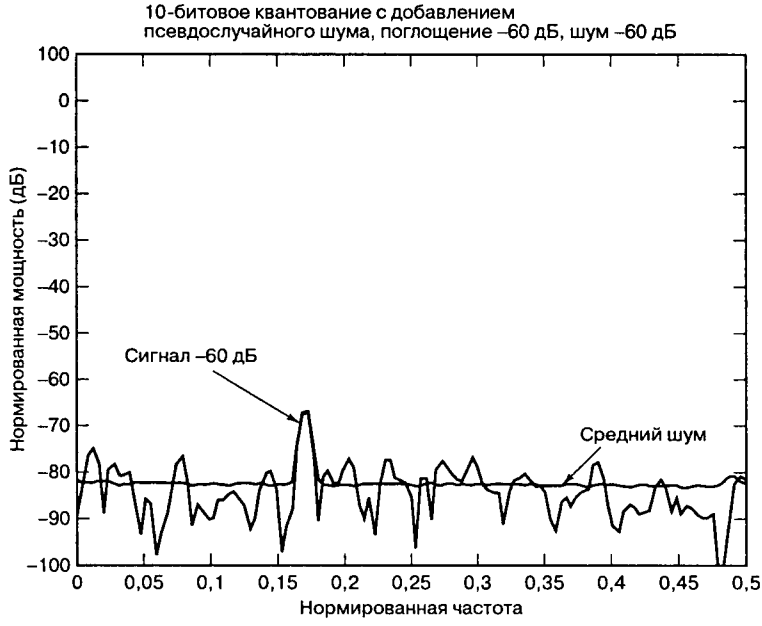


Рис. 13.11. Спектр мощности равномерного АЦП с добавлением псевдослучайного низкоуровневого сигнала

13.2.5. Неравномерное квантование

Равномерные квантующие устройства представляют собой наиболее распространенный тип аналого-цифровых преобразователей, так как они наиболее устойчивы. Под "устойчивостью" подразумевается, что они относительно нечувствительны к незначительным изменениям входных статистик. Эта устойчивость достигается в результате того, что преобразователи не настраиваются окончательно на одно конкретное множество входных параметров. Это позволяет им работать хорошо даже при наличии неопределенных входных параметров; даже незначительные изменения входных статистик приводят к несущественным изменениям выходных статистик.

Когда существует малая неопределенность в статистиках входного сигнала, можно создать неравномерное устройство квантования, которое дает меньшее отношение NSR, чем равномерное устройство квантования, использующее то же количество бит. Это реализует-

ся с помощью деления входной динамической области на неравномерные интервалы так, что мощность шума, взвешенная вероятностью появления на каждом интервале, является одинаковой. Для оптимального квантующего устройства могут быть найдены итерационные решения для границ принятия решения и размеров шагов для конкретных плотностей и малого количества бит. Эта задача упрощается путем моделирования неравномерного устройства квантования как последовательности операторов, как изображено на рис. 13.12. Сначала входной сигнал отображается с помощью нелинейной функции, называемой *компрессором* (compressor), в альтернативную область уровней. Эти уровни равномерно квантуются, и квантованные уровни сигнала затем отображаются с помощью дополняющей нелинейной функции, называемой *экспандером* (expander), в выходную область уровней. Объединяя части наименований каждой из операций COMpress и exPAND, получим название процесса: *компандирование* (companding).

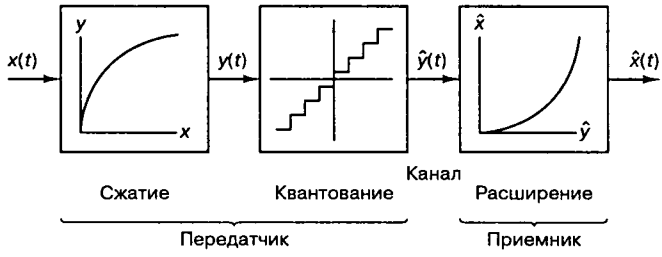


Рис. 13.12. Неравномерное устройство квантования как последовательность операторов: сжатие, равномерное квантование и расширение

13.2.5.1. Субоптимальное неравномерное квантование

Изучая характеристику компрессора $y = C(x)$ на рис. 13.13, видим, что размеры шага квантования для выходной переменной y связаны с размерами шага квантования входной переменной x через наклон $\dot{C}(x)$ (например, $\Delta y = \Delta x \dot{C}(x)$). Для произвольной функции плотности вероятности и произвольной характеристики компрессора можно достичь выходной дисперсии шума квантования [7].

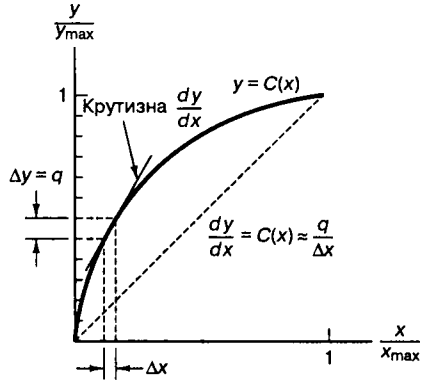


Рис. 13.13. Характеристика компрессора C(x) и оценка локального наклона $\dot{C}(x)$

$$\sigma_q^2 = \frac{q^2}{12} \int_{-x_{max}}^{x_{max}} \frac{p(x)}{\left|\dot{C}(x)\right|^2} dx \tag{13.25}$$

Для определенной функции плотности вероятности может быть найдена характеристика компрессора $C(x)$, которая минимизирует σ_q^2. Оптимальный закон сжатия для данной функции плотности вероятности выражается следующим образом [8].

$$C(x) = \int_0^x \sqrt[3]{Kp(z)}dz \tag{13.26}$$

Находим, что оптимальная характеристика сжатия пропорциональна интегралу от кубического корня от входной функции плотности вероятности. Это называется *точной настройкой* (fine tuning). Если компрессор настроен на работу с одной функцией плотности, а используется с другой (например, отличающейся только масштабом), говорят, что устройство квантования рассогласовано, и вследствие этого может существенно снижаться эффективность функционирования [6].

13.2.5.2. Логарифмическое сжатие

В предыдущем разделе был представлен закон сжатия для случая, когда входная функция плотности вероятности сигнала хорошо определена. Сейчас обратимся к случаю, в котором об этой функции известно мало. Это, например, происходит, когда средняя энтропия входного сигнала является случайной величиной. Например, уровень голоса случайно выбранного телефонного пользователя может варьироваться от одного экстремального значения (доверительный шепот) до другого (крик).

При неизвестной функции плотности вероятности характеристика компрессора неравномерного устройства квантования должна быть выбрана так, чтобы результирующий шум не зависел от конкретной плотности. Хотя это и представляется идеальным, достижение такой независимости может оказаться невозможным. Однако мы хотим компромисса и будем пытаться установить возможную независимость среди большого числа входных дисперсий и плотностей. Пример квантующего устройства, которое показывает отношение SNR, независимое от функции плотности вероятности входного сигнала, можно представить с помощью рис. 2.18. На этом рисунке можно наблюдать значительное отличие в отношениях NSR для входных сигналов с различными амплитудами, квантованных с помощью равномерного квантующего устройства. Для сравнения можно видеть, что неравномерное устройство квантования допускает только большие ошибки для больших сигналов. Преимущество такого подхода понятно интуитивно. Если SNR должно быть независимо от распределения амплитуды, шум квантования должен быть пропорционален входному уровню. В формуле (13.25) представлена дисперсия шума квантующего устройства для произвольной функции плотности вероятности и произвольной характеристики компрессора. Дисперсия сигнала для любой функции плотности вероятности равна следующему.

$$\sigma_x^2 = \int_{-\infty}^{\infty} x^2 p(x)dx \tag{13.27}$$

При отсутствии насыщения SNR квантующего устройства имеет следующий вид.

$$\frac{\sigma_x^2}{\sigma_q^2} = \frac{\int_{-x_{max}}^{x_{max}} x^2 p(x)dx}{(q^2/12)\int_{-x_{max}}^{x_{max}} [p(x)/\dot{C}^2(x)]dx} \tag{13.28}$$

Чтобы SNR не зависело от конкретной плотности, необходимо, чтобы числитель был масштабированной версией знаменателя. Это требование равносильно следующему.

$$[\dot{C}(x)]^2 = \left(\frac{K}{x}\right)^2 \tag{13.29}$$

или

$$\dot{C}(x) = \frac{K}{x} \tag{13.30}$$

Отсюда с помощью интегрирования находим следующее.

$$C(x) = \int_0^x \frac{K}{z} dz \tag{13.31}$$

или

$$C(x) = \ln x + \text{const} \tag{13.32}$$

Этот результат является интуитивно привлекательным. *Логарифмический компрессор* допускает *постоянное* SNR на выходе, поскольку с использованием логарифмической шкалы одинаковые расстояния (или ошибки) являются в действительности одинаковыми отношениями, а это и требуется для того, чтобы SNR оставалось фиксированным в области входного сигнала. Константа в равенстве (13.32) нужна для согласования граничных условий по x_{max} и y_{max}. Учитывая эти граничные условия, получим логарифмический преобразователь следующего вида.

$$\frac{y}{y_{max}} = \frac{C(x)}{y_{max}} = \ln\left(\frac{x}{x_{max}}\right) \tag{13.33}$$

Вид сжатия, предложенный логарифмической функцией, изображен на рис. 13.14, *а*. Сложность, связанная с этой функцией, состоит в том, что она не отображает отрицательные входные сигналы. Отрицательные сигналы учитываются путем добавления отраженной версии логарифма на отрицательную полуось. Эта модификация изображается на рис. 13.14 и влечет за собой следующее.

$$\frac{y}{y_{max}} = \ln\left(\frac{|x|}{x_{max}}\right)\text{sgn}(x), \tag{13.34}$$

где

$$\text{sgn}\, x = \begin{cases} +1 & \text{для } x \geq 0 \\ -1 & \text{для } x < 0 \end{cases}$$

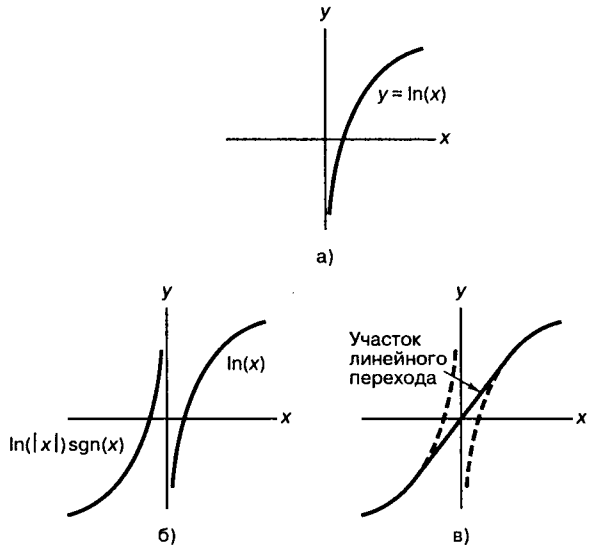


Рис. 13.14. Логарифмическое сжатие: а) прототип логарифмической функции для закона сжатия; б) прототип функции ln|x| sgn *x для закона сжатия; в) функция* ln|x| sgn *x с плавным соединением между сегментами*

Еще одна возникающая в этой ситуации сложность состоит в том, что сжатие, описанное равенством (13.34), не является непрерывным в начале координат; в действительности оно не имеет смысла в начале координат. Необходимо выполнить плавное соединение между логарифмической функцией и линейным отрезком, проходящим через начало координат. Существует две стандартные функции сжатия, выполняющие это соединение, — μ-закон компандера и *A*-закон компандера.

Компандер, использующий μ-закон. Компандер, использующий μ-закон, введенный компанией Bell System для использования в Северной Америке, имеет следующий вид.

$$y = C(x) = y_{\max} \frac{\ln[1 + \mu(|x|/x_{\max})]}{\ln(1+\mu)} \operatorname{sgn} x \tag{13.35}$$

Приблизительное поведение этого компрессора в областях, соответствующих малым и большим значениям аргумента, является следующим.

$$y = C(x) = \begin{cases} y_{\max} \dfrac{\mu(|x|/x_{\max})}{\ln(\mu)} & \text{для } \mu \dfrac{|x|}{x_{\max}} << 1 \\ y_{\max} \dfrac{\ln[\mu(|x|/x_{\max})]}{\ln(\mu)} & \text{для } \mu \dfrac{|x|}{x_{\max}} >> 1 \end{cases} \tag{13.36}$$

Параметр μ в компандере, использующем μ-закон, обычно устанавливался равным 100 для 7-битового преобразователя. Позже он изменился до 255 для 8-битового преобразователя. В настоящее время стандартным североамериканским конвертером является 8-битовый АЦП с μ = 255.

Пример 13.6. Среднее SNR для компрессора, использующего μ-закон

Среднее SNR для компрессора, использующего μ-закон, можно оценить, подставляя выражение для μ-закона в формулу (13.28). Для положительных значений входной переменной x закон сжатия имеет следующий вид.

$$y = C(x) = y_{\max} \frac{\ln[1 + \mu(|x|/x_{\max})]}{\ln(1+\mu)} \tag{13.37}$$

Затем производная равна следующему.

$$y = C(x) = y_{\max} \frac{1}{\ln(1+\mu)} \frac{\mu(1/x_{\max})}{1+\mu(|x|/x_{\max})} \tag{13.38}$$

Для значений входной переменной, для которых $\mu(x/x_{\max})$ является большим в сравнении с единицей, производная переходит в следующее выражение.

$$y = C(x) \approx \frac{1}{x} \frac{y_{\max}}{\ln(\mu)} \tag{13.39}$$

Подставляя $1/C(x)$ в формулу (13.28), получаем следующее.

$$\text{SNR} = \frac{\sigma_s^2}{\sigma_q^2} = \frac{1}{(q^2/12)[\ln(\mu)/y_{\max}]^2} = \tag{13.40}$$

$$= 3\left(\frac{2y_{\max}}{q}\right)^2 \left(\frac{1}{\ln(\mu)}\right)^2 \tag{13.41}$$

Отношение $2y_{\max}/q$ приблизительно равно числу уровней квантования (2^b) для b-битового сжимающего устройства квантования. Для 8-битового преобразователя с $\mu = 255$ имеем следующее.

$$\text{SNR} = 3\left[\frac{2^8}{\ln(255)}\right]^2 = 3(46{,}166)^2 = 38{,}1\,\text{дБ} \tag{13.42}$$

Для сравнения на рис. 13.15 представлено отношение SNR АЦП, использующего μ-закон. Здесь SNR изображено для входных синусоид различной амплитуды. Там же изображен уровень 38,1 дБ, вычисленный в формуле 13.42, и SNR для линейного квантующего устройства с той же областью входных амплитуд. Как и предсказывалось, квантующее устройство, использующее μ-закон, поддерживает постоянное SNR для значительного диапазона входных уровней. Зубчатость кривой производительности (гранулярность квантующего устройства) вызвана логарифмической функцией сжатия. Реальные преобразователи, помимо этого, показывают дополнительную зубчатость вследствие кусочно-линейной аппроксимации непрерывной кривой μ-закона.

На рис. 13.16 представлено дискретное преобразование Фурье пары входных синусоид относительных амплитуд 1,0 (0 дБ) и 0,01 (–40 дБ). Входной сигнал квантуется с помощью 10-битового преобразователя, использующего μ-закон ($\mu = 500$), и на рис. 13.16, *а–в* уровни сигнала ослабляются на 1,20 и 40 дБ относительно полномасштабного входа. Отметим, что уровни шума квантования для полномасштабного сигнала на рис. 13.16, *а* выше, чем у равномерного АЦП (–72 дБ против –83 дБ, как видно из рис. 13.7). Для ослабленных сигналов отмечаем улучшенное отношение SNR логарифмически сжимающего АЦП по сравнению с равномерным АЦП. Видно, что поскольку уровни входного сигнала уменьшились, шум квантования также снизился, и при ослаблении в 40 дБ уровень шума упал до –108 дБ. Таким образом, логарифмически сжимающие АЦП не имеют проблемы "видения" входного сигнала низкого уровня даже при ослаблении на 40 дБ,

как на рис. 13.16, *в*, в то время как тот же сигнал теряется среди шума равномерного преобразователя, как показано на рис. 13.7, *в*.

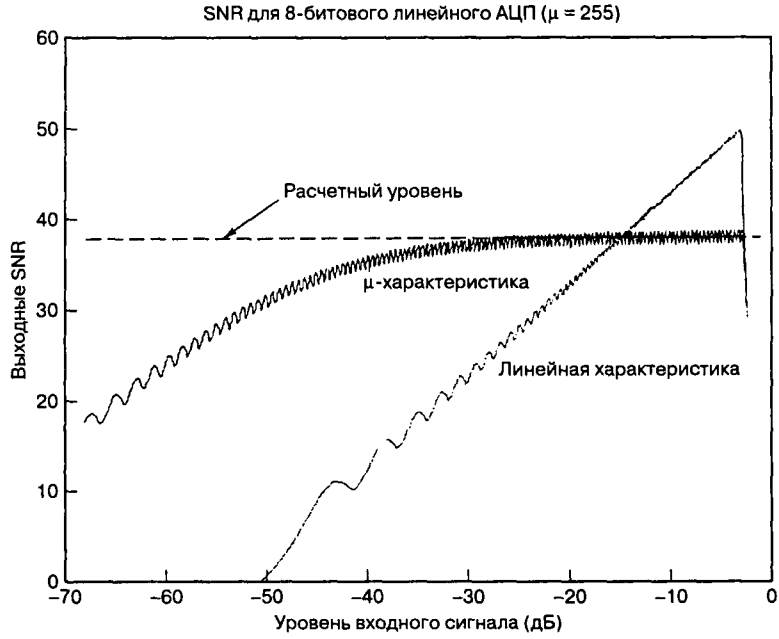


Рис. 13.15. Предсказанное и измеренное отношение SNR для АЦП, использующего μ-закон

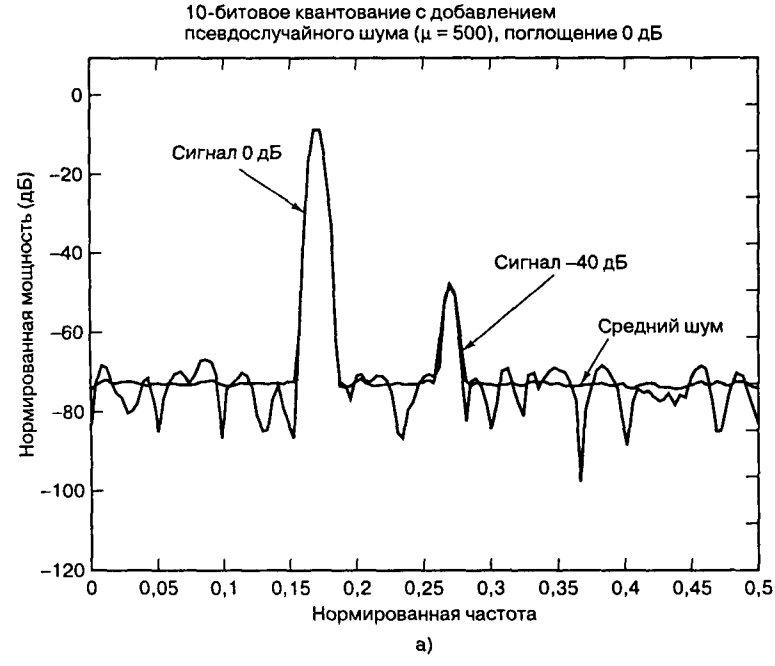
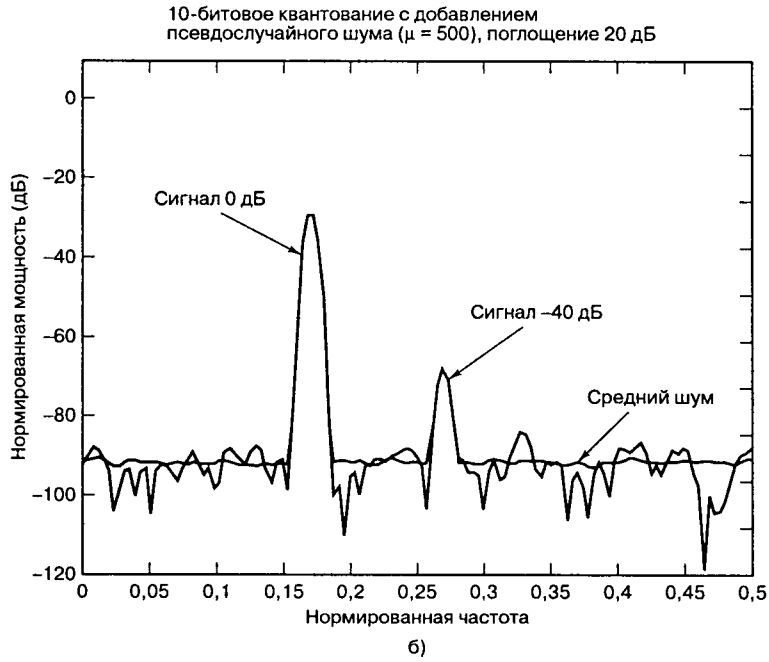
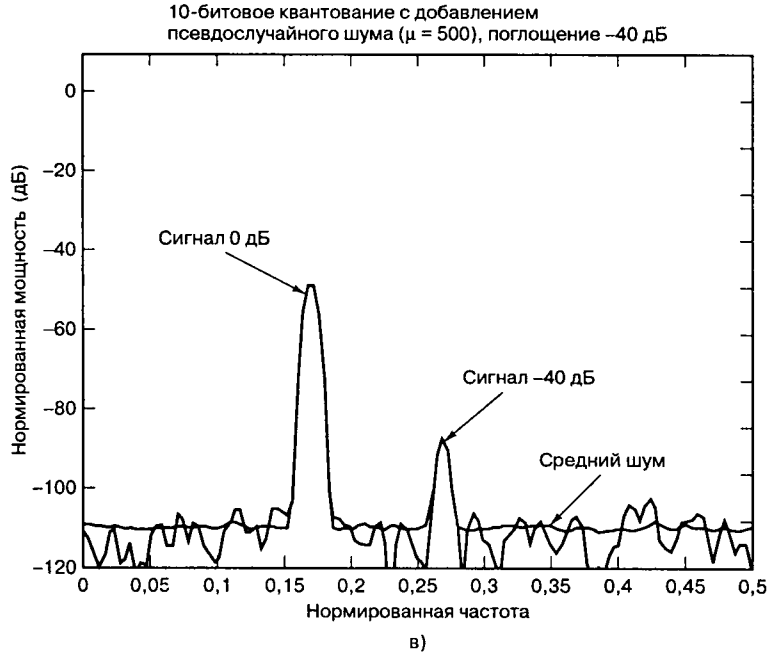


Рис. 13.16. Спектр мощности сигналов АЦП, использующего μ-закон



б)



в)

Рис. 13.16. Спектр мощности сигналов АЦП, использующего μ-закон (окончание)

Реальная характеристика компрессора, использующего μ-закон, описана формулой (13.35). Как показано на рис. 13.17, 16 сегментов линейных хорд аппроксимируют функциональное выражение на 256 возможных выходных уровнях. Восемь из этих сегментов расположены в первом квадранте, восемь — в третьем квадранте и сегмент "0" имеет один и тот же наклон в обоих квадрантах. Вдоль каждого сегмента хорды квантование является равномерным по четырем битам преобразования низшего порядка. Таким образом, 8-битовый сжимающий формат преобразования имеет следующий вид.

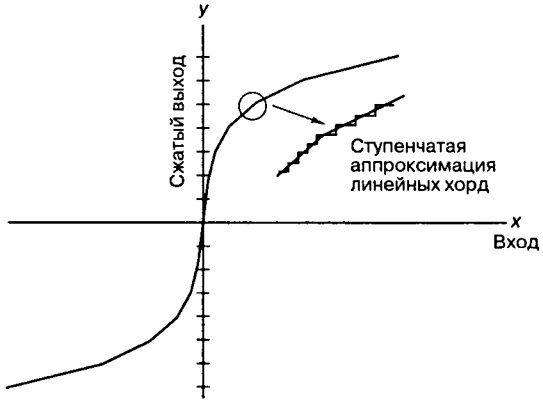


Рис. 13.17. Семибитовое сжатое квантование для 16-сегментной аппроксимации μ-закона

$$\underbrace{b_7}_{\text{бит знака}} \qquad \underbrace{b_6b_5b_4}_{\text{сегмент}} \qquad \underbrace{b_3b_2b_1b_0}_{\text{положение в сегменте}}$$

Он представляет собой кусочную аппроксимацию хордами до плавной функции и ступенчатую аппроксимацию каждой хорды, учитывающую дополнительную зубчатость в кривой SNR, которая представлена на рис. 13.15.

Компандер, использующий A-закон. Этот компандер является стандартом CCITT (Consultative Committee for International Telephone and Telegraphy — Международный консультативный комитет по телеграфии и телефонии, МККТТ), а следовательно европейским стандартом аппроксимации логарифмического сжатия. Характеристика компрессора имеет следующий вид.

$$y = C(x) = \begin{cases} y_{\max} \dfrac{A(|x|/x_{\max})}{1+\ln(A)} \operatorname{sgn} x & \text{для } 0 < \dfrac{|x|}{x_{\max}} < \dfrac{1}{A} \\ y_{\max} \dfrac{1+\ln[A(|x|/x_{\max}]}{1+\ln A} \operatorname{sgn} x & \text{для } \dfrac{1}{A} < \dfrac{|x|}{x_{\max}} < 1 \end{cases} \tag{13.43}$$

Стандартным значением параметра A является 87,56, и (при использовании 8-битового преобразователя) SNR для этого значения равно 38,0 дБ. Сжимающая характеристика A-закона аппроксимируется подобно тому, как это делалось для компрессора, использующего μ-закон, — с помощью последовательности 16 линейных хорд, охватывающих выходную область. Нижние две хорды в каждом квадранте являются в действительности хордами сигнала, соответствующими линейному сегменту компрессора, использующего A-закон. Одним важным отличием между характеристиками сжатия A- и μ-законов является то, что стандарт A-закона имеет характеристику с ну-

лем на границе шага квантования, в то время как стандарт μ-закона — характеристику с нулем в центре шага квантования. Таким образом, компрессор с *A*-законом не имеет нулевого значения, и следовательно, для него не существует интервала, на котором бы при нулевом входе не передавались данные.

Существует прямое отображение из формата АЦП, использующего 8-битовое сжатие с *A*-законом, в 12-битовый линейный двоичный код и из формата 8-битового сжатия с μ-законом в 13-битовый линейный код [8]. Эта операция позволяет преобразование аналоговой информации в цифровую с помощью равномерного устройства квантования с последующим отображением в меньшее число бит в кодовом преобразователе. Кроме того, это позволяет обратное отображение в приемнике (т.е. расширение) производить на числовой выборке.

Импульсно-кодовая модуляция. Одной из задач, выполняемых в ходе импульсно-кодовой модуляции (pulse-code modulation — PCM), является преобразование исходных волновых сигналов в дискретные двоичные последовательности. Эта задача производится с помощью трехэтапного процесса — дискретизации, квантования и кодирования. Процесс дискретизации изучался в главе 2, а процесс квантования — в данной главе и в главе 2. Отметим, что процесс кодирования, следующий за квантованием (см. рис. 2.2), часто воплощается на аппаратном уровне и выполняется тем же устройством, что и квантование. Вообще, процесс может быть описан следующим образом: последовательная аппроксимация аналого-цифровых преобразователей образует последовательные биты декодированных данных с помощью обратной связи, сравнения и процесса принятия решения. В процессе обратной связи постоянно задается вопрос, входной сигнал находится выше или ниже средней точки остаточного интервала неопределенности. С помощью этой технологии интервал неопределенности сокращается до половинного на каждом шаге сравнения и принятия решения до тех пор, пока интервал неопределенности не совпадет с допустимым интервалом квантования.

При последовательной аппроксимации результат каждого предыдущего решения снижает неопределенность, которая должна быть разрешена во время следующего преобразования. Аналогично результаты предшествующих преобразований аналоговой информации в цифровую могут использовать для уменьшения неопределенности, которая должна быть разрешена во время следующего преобразования. Эта редукция неопределенности достигается путем передачи каждой последующей выборке вспомогательной информации из более ранних выборок. Эта информация называется избыточной частью сигнала, и с помощью ее передачи сокращается интервал неопределенности, в котором квантующее устройство и кодер должны вести поиск следующей выборки сигнала. Передача данных — это один из методов, с помощью которых достигается *снижение избыточности.*

13.3. Дифференциальная импульсно-кодовая модуляция

Используя прошлые данные для измерения (т.е. квантования) новых переходим от обычной импульсно-кодовой модуляции (pulse-code modulation — PCM) к дифференциальной (differential PCM — DPCM). В DPCM предсказание следующего выборочного значения формируется на основании предыдущих значений. Для квантующего устройства это предсказание можно рассматривать в качестве инструкции по руководству при поиске следующего выборочного значения в конкретном интервале. Если для предсказания используется избыточность сигнала, область неопределенности сокращается и квантование можно проводить с уменьшенным числом решений (или

бит) для данного уровня квантования или с уменьшенным числом уровней квантования для данного числа решений (или бит). Сокращение избыточности реализуется путем вычитания предсказания из следующего выборочного значения. Эта разность называется *ошибкой предсказания* (prediction error).

Устройства квантования, описанные в разделе 13.2, называются *мгновенными* устройствами квантования или устройствами квантования *без памяти*, так как цифровые преобразования основаны на единичной (текущей) входной выборке. В разделе 13.1 были определены свойства источников, которые допускают сокращение интенсивности источника. Этими свойствами были неравновероятные уровни источника и зависимые выборочные значения. Мгновенные квантующие устройства кодируют источник, принимая во внимание плотность вероятности, сопоставленную с каждой выборкой. Методы квантования, которые принимают во внимание корреляцию между выборками, являются квантующими устройствами *с памятью*. Эти квантующие устройства уменьшают избыточность источника сначала посредством превращения коррелированной входной последовательности в связанную последовательность с уменьшенной корреляцией, уменьшенной дисперсией и уменьшенной полосой частот. Затем эта новая последовательность квантуется с использованием меньшего количества бит.

Корреляционные характеристики источника можно представить во временной области с помощью выборки его автокорреляционной функции и в частотной области — его спектром мощности. Если изучается спектр мощности $G_x(f)$ кратковременного речевого сигнала, как изображено на рис. 13.18, то видим, что спектр имеет глобальный максимум в окрестности от 300 до 800 Гц и убывает со скоростью от 6 до 12 дБ/октаву. Изучая этот спектр, можно взглянуть на определенные свойства временной функции, из которой он получен. Видим, что большие изменения сигнала происходят медленно (низкая частота), а быстрые (высокая частота) должны иметь малую амплитуду. Эквивалентная интерпретация может быть дана в терминах автокорреляционной функции сигнала $R_x(T)$, как изображено на рис. 13.19. Здесь широкая, медленно меняющаяся автокорреляционная функция свидетельствует о том, что при переходе от выборки к выборке будет только слабое изменение и что для полного изменения амплитуды требуется временной интервал, превышающий интервал корреляции. Интервал (или радиус) корреляции, рассмотренный на рис. 13.19, является временной разностью между максимальной и первой нулевой корреляцией. В частности, значение корреляции для типичного единичного выборочного запаздывания лежит в диапазоне примерно от 0,79 до 0,87, а радиус корреляции имеет порядок от 4 до 6 выборочных интервалов, равных T секунд на интервал.

Поскольку разность между соседними временными выборками для речи мала, используемый метод кодирования базируется на передаче от выборки к выборке разностей, а не действительных выборочных значений. В действительности, последовательные разности представляют собой частный случай класса преобразователей с памятью, называемых N-отводными линейными кодерами с предсказанием. Эти кодеры, иногда именуемые кодерами с предсказаниями и поправками, предсказывают следующее входное выборочное значение на основании предыдущих входных выборочных значений. Эта структура показана на рис. 13.20. В этом типе преобразователя передатчик и приемник имеют одинаковую модель предсказания, которая получена из корреляционных характеристик сигнала. Кодер дает ошибку предсказания (или остаток) как разность между следующим измеренным и предсказанным выборочными значениями. Математически контур предсказания описывается следующим образом.

Рис. 13.18. Типичный спектр мощности для речевых сигналов

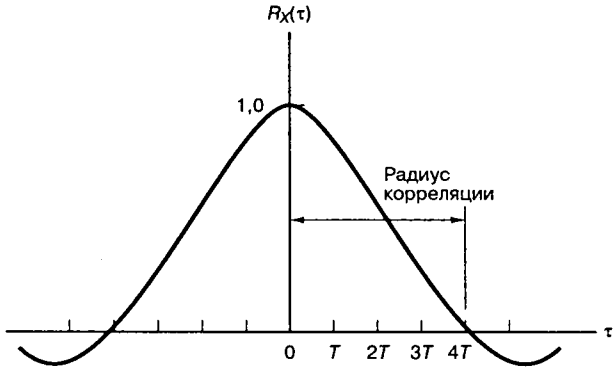


Рис. 13.19. Автокорреляционная функция для типичных речевых сигналов

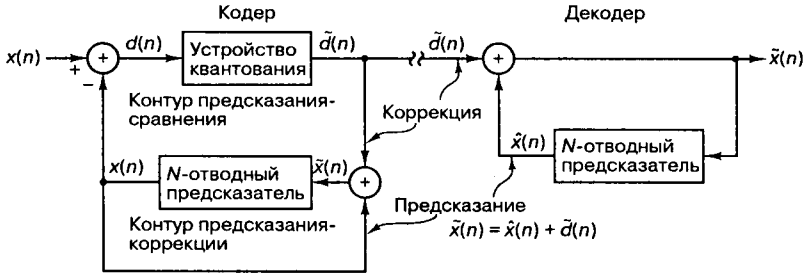


Рис. 13.20. N-отводный дифференциальный импульсно-кодовый модулятор с предсказанием

$$d(n) = x(n) - \hat{x}(n),$$

где $x(n)$ — n-я входная выборка, $\hat{x}(n)$ — предсказанное значение выборки, а $d(n)$ — соответствующая ошибка предсказания. Эта операция производится в контуре пред-

сказания и сравнения, верхний контур кодера изображен на рис. 13.20. Кодер корректирует свои предсказания, составляя сумму предсказанного значения и ошибки предсказания. Математически контур коррекции описывается следующим образом.

$$\tilde{d}(n) = \text{quant}[d(n)]$$
$$\tilde{x}(n) = \hat{x}(n) + \tilde{d}(n)$$

Здесь quant(·) представляет операцию квантования, $\tilde{d}(n)$ — квантованная версия ошибки предсказания, а $\tilde{x}(n)$ — скорректированная и квантованная версия входной выборки. Это делается в контуре предсказания и поправок, в нижнем цикле кодера и в единственном контуре декодера на рис. 13.20. Декодер должен быть также проинформирован об ошибках предсказания, чтобы использовать свой контур коррекции для поправки своего предсказания. Декодер "повторяет" обратный цикл кодера. Задача связи состоит в передаче разности (ошибки сигнала) между предсказанными и действительными выборочными данными. По этой причине описанный класс кодеров часто называется дифференциальным импульсно-кодовым модулятором (differential pulse code modulator — DPCM). Если модель предсказания дает предсказания, близкие к действительным выборочным значениям, для остатков будет характерна уменьшающаяся дисперсия (по отношению к исходному сигналу). Из раздела 13.2 известно, что число бит, которое требуется для перемещения данных через канал с заданной точностью, связано с дисперсией сигнала. Следовательно, уменьшенная последовательность остатков может быть передана через канал с уменьшенной скоростью.

Преобразователи с предсказанием должны иметь кратковременную память, которая поддерживает проводимые в реальном времени операции, требуемые для алгоритма предсказания. Кроме того, они часто будут иметь долгосрочную память, которая поддерживает медленные, зависимые от данных операции, такие как автоматическая регулировка усиления, коррекция коэффициентов фильтра. Предсказатели, которые включают медленные, зависимые от данных регулирующие алгоритмы, называются *адаптивными*.

13.3.1. Одноотводное предсказание

Одноотводный линейный кодирующий фильтр с предсказанием (linear prediction coding filter — фильтр LPC) в процессе модуляции DPCM предсказывает последующее входное выборочное значение, основываясь на предшествующем входном выборочном значении. Уравнение предсказания имеет следующий вид.

$$x(n|n-1) = ax(n-1|n-1) \tag{13.44}$$

Здесь $x(n|m)$ — оценка x в момент n при данных всех выборках, собранных за время m и a — параметр, используемый для минимизации ошибки предсказания. Полученная после измерений ошибка предсказания имеет следующий вид.

$$d(n) = [x(n) - x(n|n-1)] = \tag{13.45,а}$$
$$= [x(n) - ax(n-1|n-1)] \tag{13.45,б}$$

Среднеквадратическая ошибка имеет следующий вид.

$$\mathbf{E}\{d^2(n)\} = \mathbf{E}\{x(n)x(n) - 2ax(n)x(n-1|n-1) + a^2x(n-1|n-1)x(n-1|n-1)\} \tag{13.46}$$

Если $x(n-1|n-1)$ является несмещенной оценкой $x(n-1)$, равенство (13.46) может быть записано следующим образом.

$$R_d(0) = R_x(0) - 2aR_x(1) + a^2R_x(0) = \tag{13.47,а}$$

$$= R_x(0)[1 + a^2 - 2aC_x(1)] \tag{13.47,б}$$

В данном случае $R_d(n)$ и $R_x(n)$ являются автокорреляционными функциями ошибки предсказания и входного сигнала. $R_d(0)$ — мощность ошибки, $R_x(0)$ — мощность сигнала, а $C_x(n) = R_x(n)/R_x(0)$ — нормированная автокорреляционная функция. Параметр a можно выбрать так, чтоб он минимизировал мощность ошибки предсказания, указанную в формуле (13.47). Для этого нужно частную производную по a от $R_d(0)$ положить равной нулю.

$$\frac{\partial R_d(0)}{\partial a} = R_x(0)\left[2a - 2C_x(1)\right] \tag{13.48}$$

Решая данное уравнение, получим оптимальное значение a^{opt}.

$$a^{opt} = C_x(1) \tag{13.49}$$

Подставляя a^{opt} в уравнение (13.47), получим следующее.

$$R_d^{opt}(0) = R_x(0)[1 + a^{opt}C_x(1) - 2a^{opt}C_x(1) = \tag{13.50,а}$$

$$= R_x(0)[1 - a^{opt}C_x(1)] = \tag{13.50,б}$$

$$= R_x(0)[1 - C_x^2(1)] \tag{13.50,в}$$

Усиление предсказания (prediction gain) кодера можно определить как отношение входной и выходной дисперсий, $R_x(0)/R_d(0)$. Для фиксированной частоты передачи бит этот коэффициент представляет собой увеличение в выходном SNR, а для фиксированного выходного SNR — сокращение описания скорости передачи бит. Отметим, что, как использовалось в равенстве (13.50,б), усиление предсказания для оптимального предсказателя всегда больше единицы для любого значения корреляции сигнала $R_x(0)$. С другой стороны, как использовалось в равенстве (13.47,б), оно больше единицы для неоптимального одноотводного единичного предсказателя, только если корреляция сигнала превышает 0,5.

Пример 13.7. Усиление предсказания для одноотводного фильтра LPC

Сигнал с коэффициентом корреляции $C_x(1)$, равным 0,8, должен квантоваться одноотводным фильтром LPC. Определите усиление предсказания, если коэффициент предсказания 1) оптимизирован по отношению к минимальной ошибке предсказания; 2) положен равным единице.

Решение

а) Из уравнения (13.50,в) имеем следующее.

$$R_d^{opt}(0) = R_x(0)(1 - 0{,}64) = 0{,}36R_x(0) \tag{13.51,а}$$

$$\text{Усиление предсказания} = 1/(0{,}36) = 2{,}78 \text{ или } 4{,}44 \text{ дБ} \tag{13.51,б}$$

б) Из уравнения (13.47,б) имеем

$$R_d(0) = 2R_x(0)(1 - 0{,}8) = 0{,}40R_x(0). \tag{13.51,в}$$

$$\text{Усиление предсказания} = 1/(0{,}40) = 2{,}50 \text{ или } 3{,}98 \text{ дБ} \tag{13.51,г}$$

13.3.2. *N*-отводное предсказание

N-отводный фильтр LPC предсказывает последующее выборочное значение на основании линейной комбинации предшествующих N выборочных значений. Будем предполагать, что квантованные оценки, которые используются предсказывающими фильтрами, являются несмещенными и безошибочными. Приняв это предположение, можно опустить двойные индексы (использованные в разделе 13.3.1) для данных в фильтре, но использовать их для предсказания. Тогда уравнение *N*-отводного предсказания принимает следующий вид.

$$x(n|n-1) = a_1 x(n-1) + a_2 x(n-2) + \ldots + a_N x(n-N) \tag{13.52}$$

Ошибка предсказания принимает следующий вид.

$$d(n) = x(n) - x(n|n-1) = \tag{13.53,а}$$

$$= x(n) - a_1 x(n-1) - a_2 x(n-2) - \ldots - a_N x(n-N) \tag{13.53,б}$$

Среднеквадратическая ошибка предсказания имеет вид

$$\mathbf{E}\{d(n)d(n)\} = \mathbf{E}\{[x(n) - x(n|n-1)]^2\}. \tag{13.54}$$

Ясно, что среднеквадратическая ошибка предсказания выражается через квадрат коэффициентов фильтра a_j. Можно образовать частные производные от среднеквадратических ошибок по каждому коэффициенту, как это делалось в разделе 13.3.1, и найти коэффициенты, которые обращают частные производные в нуль. Формально, вычисляя частные производные по j-му коэффициенту (до раскрытия $x(n|n-1)$), получим следующее.

$$\frac{\partial R_d(0)}{\partial a_j} = \mathbf{E}\left\{2\left[x(n) - x(n|n-1)\right]\frac{\partial x(n|n-1)}{\partial a_j} x(n|n-1)\right\} = \tag{13.55,а}$$

$$= \mathbf{E}\left\{2\left[x(n) - x(n|n-1)\right]\left[-x(n-j)\right]\right\} = \tag{13.55,б}$$

$$= 2\mathbf{E}\left\{\left[x(n) - a_1 x(n-1) - a_2 x(n-2) - \ldots - a_N x(n\_N)\right]\left[-x(n-j)\right]\right\} = \tag{13.55,в}$$

$$= 2\left[R_x(j) - a_1 R_x(j-1) - a_2 R_x(j-2) - \ldots - a_N R_x(j-N)\right] \tag{13.55,г}$$

Эта система уравнений (по одному для каждого j) может быть записана в матричной форме, и тогда она будет называться нормальными уравнениями.

$$\begin{bmatrix} R_x(1) \\ R_x(2) \\ R_x(3) \\ \cdot \\ \cdot \\ \cdot \\ R_x(N) \end{bmatrix} = \begin{bmatrix} R_x(0) & R_x(-1) & R_x(-2) & \ldots & R_x(-N+1) \\ R_x(1) & R_x(0) & R_x(-1) & \ldots & R_x(-N+2) \\ R_x(2) & R_x(1) & R_x(0) & \ldots & R_x(-N+3) \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ R_x(N-1) & R_x(N-2) & R_x(N-3) & \ldots & R_x(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdot \\ \cdot \\ \cdot \\ a_N \end{bmatrix}^{\text{opt}} \tag{13.56,а}$$

Нормальные уравнения могут быть записаны более компактно.

$$\mathbf{r}_x(1, N) = \mathbf{R}_{xx}\mathbf{a}^{\text{opt}}, \tag{13.56,б}$$

где $\mathbf{r}_x(1, N)$ — это корреляционный вектор задержек от 1 до N, $\mathbf{R}_{xx}$ — корреляционная матрица (предполагается процесс с нулевым средним), а $\mathbf{a}^{\text{opt}}$ — вектор оптимальных весовых коэффициентов фильтра.

Чтобы изучить решения нормальных уравнений, запишем уравнение (13.54) для среднеквадратической ошибки в матричной форме.

$$R_d(0) = \mathbf{E}\{[x(n) - \mathbf{a}^T\mathbf{x}(n-1)][x(n) - \mathbf{x}^T(n-1)\mathbf{a}]\} = \tag{13.57,а}$$

$$= R_x(0) - \mathbf{r}_x^T(1, N)\mathbf{a} - \mathbf{a}^T\mathbf{r}_x(-1, -N) + \mathbf{a}^T\mathbf{R}_{xx}\mathbf{a}, \tag{13.57,б}$$

где $\mathbf{r}^T$ — транспонированная матрица для матрицы $\mathbf{r}$. Замена $\mathbf{a}^{\text{opt}}$ на $\mathbf{a}$ в равенстве (13.57,б) с последующей заменой $\mathbf{r}_x(1, N)$ на $\mathbf{R}_{xx}\mathbf{a}^{\text{opt}}$ дает следующее.

$$R_d(0) = R_x(0) - \mathbf{r}_x^T(1, N)\mathbf{a}^{\text{opt}} - \mathbf{a}^{\text{opt}\,T}\mathbf{r}_x(-1, -N) + \mathbf{a}^{\text{opt}\,T}\mathbf{r}_x(1, N) = \tag{13.58,а}$$

$$= R_x(0) - \mathbf{r}_x^T(-1, -N)\mathbf{a}^{\text{opt}} \tag{13.58,б}$$

Теперь можем перенести правую часть уравнения (13.5б) в левую и использовать уравнение (13.58,б) для дополнения верхней строки матрицы, чтобы получить "чистый" вид оптимального предсказателя.

$$\begin{bmatrix} R_x(0) & R_x(-1) & R_x(-2) & R_x(-3) & \dots & R_x(-N) \\ R_x(1) & R_x(0) & R_x(-1) & R_x(-2) & \dots & R_x(-N+1) \\ R_x(2) & R_x(1) & R_x(0) & R_x(-1) & \dots & R_x(-N+2) \\ R_x(3) & R_x(2) & R_x(1) & R_x(0) & \dots & R_x(-N+3) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ R_x(N) & R_x(N-1) & R_x(N-2) & R_x(N-3) & \dots & R_x(0) \end{bmatrix} \begin{bmatrix} 1 \\ -a_1 \\ -a_2 \\ -a_3 \\ \vdots \\ -a_N \end{bmatrix}^{\text{opt}} = \begin{bmatrix} R_d(0) \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{13.59}$$

В этой форме ненулевой выход матричного произведения имеет место только в момент нуль, что подобно выходному импульсу.

Верхняя строка уравнения (13.59) свидетельствует о том, что мощность ошибки предсказания имеет следующий вид.

$$R_d(0) = R_x(0)[1 - a_1C_x(1) - a_2C_x(2) - \dots - a_NC_x(N)] \tag{13.60}$$

Сравните это равенство с (13.50,б). Интересное свойство оптимального N-отводного фильтра с предсказанием состоит в том, что множество коэффициентов, которое задает минимальную среднеквадратическую ошибку предсказания, с нулевой ошибкой предсказывает также последующие $N-1$ корреляционных выборок на основании предшествующих $N-1$ корреляционных выборок. Для фиксированных коэффициентов фильтра кодер DPCM может давать усиление предсказания относительно линейного квантования от 6 до 8 дБ [9, 10]. Это усиление, по сути, независимо от длины фильтра, если длина превосходит три или четыре отвода. Дополнительное усиление имеет место, если кодер обладает медленными адаптивными свойствами. Адаптивные кодеры вводятся в разделе 13.3.3 и подробнее обсуждаются в разделе 13.3.4.

13.3.3. Дельта-модуляция

Дельта-модуляция, часто обозначаемая как Δ-модуляция, представляет собой процесс внедрения низкой разрешающей способности аналого-цифрового преобразователя в контур обратной связи дискретных данных, работающий со скоростью, значительно превышающей частоту Найквиста. Причиной возникновения этой технологии стало то, что в процессе преобразования скорость — это менее дорогой ресурс, чем точность, и разумнее будет использовать более быстрые процессы обработки сигналов для получения более высокой точности.

Из равенства (13.50,в) следует, что усиление предсказания для одноотводного предсказателя могло бы быть *большим*, если бы нормированный коэффициент корреляции $C_x(1)$ был близок к единице. Для того чтобы увеличить корреляцию выборок, фильтр с предсказанием обычно работает со скоростью, которая далеко превосходит частоту Найквиста. Например, частота произведения выборок может быть выбрана в 64 раза большей, чем частота Найквиста. Тогда для полосы частот в 20 кГц с номинальной частотой выборки 48 кГц фильтр с сильно корреляционным предсказанием будет работать с частотой 3 072 МГц. Причина выбора такой высокой частоты дискретизации заключается в следующем: необходимо убедиться, что выборочные данные имеют высокую корреляцию, так что простой одноотводный предсказатель будет давать малую ошибку предсказания, которая, в свою очередь, допускает работу устройства квантования с очень малым количеством бит в контуре коррекции ошибок. Простейшей формой устройства квантования является однобитовый преобразователь; по сути, это просто компаратор, который обнаруживает и сообщает знак разности сигнала. Как следствие, ошибкой предсказания сигнала является 1-битовое слово, которое имеет интересное преимущество — оно не требует следить за порядком слов при последовательной обработке.

Блок-схема одноотводного линейного предсказателя, изображенного на рис. 13.20, с небольшой модификацией показана на рис. 13.21. Отметим, что одноотводный контур предсказания-коррекции является сейчас просто интегратором и в декодере за контуром предсказания-коррекции следует восстанавливающий фильтр нижних частот. Этот фильтр устраняет выходящий за полосу частот шум квантования, который генерируется двухуровневым кодированием и распространяется за пределы информационной полосы частот этого кодирующего процесса. Кодер полностью описывается частотой дискретизации, размером шага квантования (для разрешения ошибки предсказания или *допустимой ошибки* контура) и восстанавливающим фильтром. Уравнения для предсказания и остаточной ошибки модулятора имеют следующий вид.

$$x(n|n-1) = x(n-1|n-1) \qquad (13.61,а)$$

$$d(n) = x(n) - x(n|n-1)\,, \qquad (13.61,б)$$

где n — выборочный индекс. Эта структура, иногда называемая дельта-модулятором, представляет собой процесс DPCM, при котором контур предсказания-коррекции состоит из цифрового аккумулятора.

13.3.4. Сигма-дельта-модуляция

Структура Σ-Δ-модулятора может быть изучена с помощью различных средств; наиболее привлекательными являются модифицированный одноотводный преобразователь DPCM, а также преобразователь с обратной связью по ошибке. Начнем с модифицированного одноотводного преобразователя DPCM. Как указывалось ранее,

контур зависит от высокой корреляции последовательных выборок, чего можно достичь за счет передискретизации.

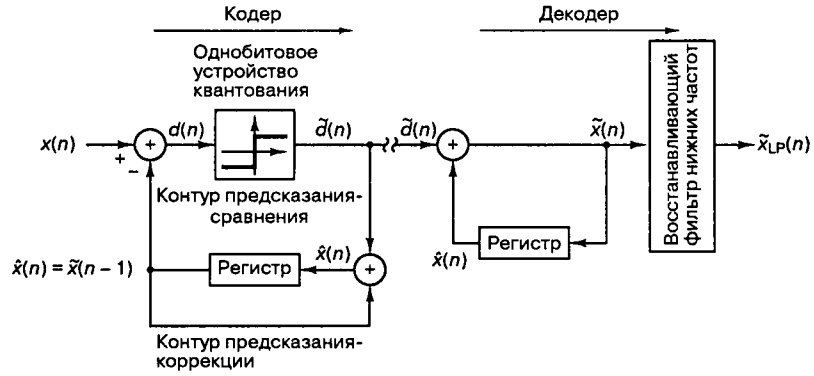


Рис. 13.21. Одноотводный, однобитовый кодер DPCM (дельта-модулятор)

Корреляцию поступающих на модулятор выборочных данных можно усилить посредством предварительной фильтрации данных интегратором и компенсации этой фильтрации с помощью выходного фильтра-дифференциатора. Эта структура изображена на рис. 13.22, где интеграторы, дифференциатор и задержка выражены в терминах z-преобразования (см. приложение Д). Затем для получения выигрыша от реализации можно перегруппировать блоки прохождения сигнала. На вход кодера поступают сигналы с выходов двух цифровых интеграторов, которые затем суммируются и вводятся в контур квантования. Первая модификация состоит в том, чтобы использовать один цифровой интегратор, сдвигая два интегратора через суммирующее устройство в кодер. Вторая модификация состоит в том, что выходной фильтр-дифференциатор может быть сдвинут в декодер, что делает ненужным цифровой интегратор на входе в декодер. Все, что остается от декодера, — это восстанавливающий фильтр нижних частот. Полученная упрощенная форма модифицированной системы DPCM изображена на рис. 13.23. Эта форма, названная *сигма-дельта*-модулятором, содержит интегратор (*сигма*) и модулятор DPCM (*дельта*) [11].

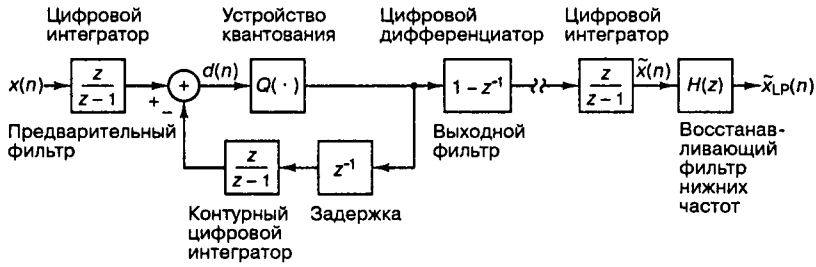


Рис. 13.22. Однобитовый дельта-модулятор

Понять Σ-Δ-модулятор можно путем рассмотрения контура обратной связи по шуму. Понятно, что устройство квантования для получения выходного сигнала добавляет ошибку к своему входному сигналу. Когда выборки образовываются со значительным запасом, то высоко коррелируют не только выборки, но и ошибки. Когда ошибки высоко коррелируют, они предсказуемы, и, таким образом, они могут быть вычтены из сигнала, отправленного на устройство квантования прежде, чем произойдет процесс квантования. Когда сиг-

нал и ошибка представляются передискретизованными выборками, предшествующая ошибка квантования может быть использована как хорошая оценка текущей ошибки.

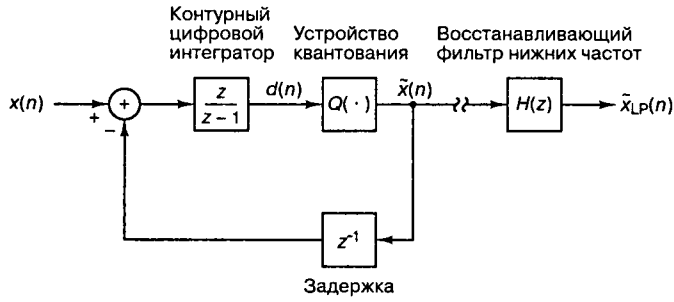


Рис. 13.23. Σ-Δ-модулятор как перегруппированный Δ-модулятор

Предшествующая ошибка, образованная как разность между входом и выходом устройства квантования, помещается в регистр запаздывания для использования в качестве оценки следующей ошибки квантования. Эта структура изображена на рис. 13.24. Схему прохождения сигнала на рис. 13.24 можно перерисовать так, чтобы акцентировать внимание на двух входах (сигнал и шум квантования) и на двух контурах (включающий устройство квантования и не включающий его). Эта форма изображена на рис. 13.25 и является общепринятой для точного изображения участка обратной связи цифрового интегратора. Эта схема имеет ту же структуру, что и представленная на рис. 13.23. Из рис. 13.25 видно, что выход Σ-Δ-модулятора и его z-преобразование (см. приложение Д) могут быть записаны в следующем виде.

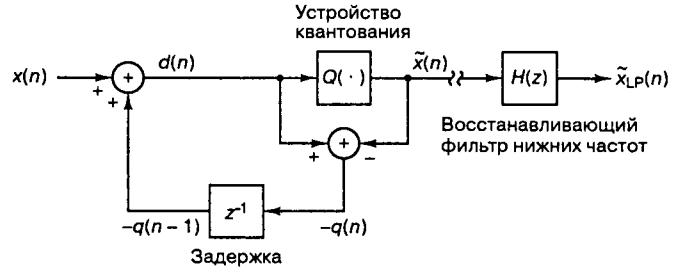


Рис. 13.24. Σ-Δ-модулятор как процесс обратной связи по шуму

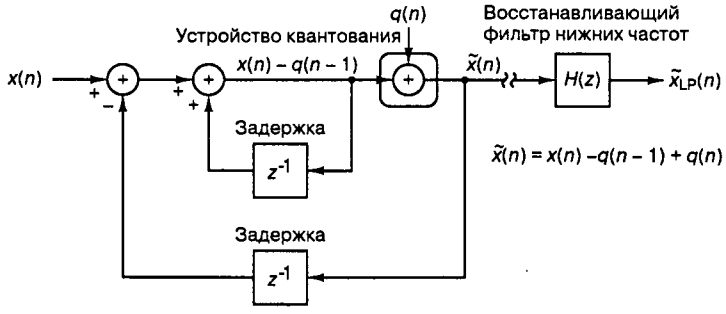


Рис. 13.25. Устройство квантования с обратной связью по шуму, изображенное как Σ-Δ-модулятор

$$y(n) = \tilde{x}(n) - q(n-1) + q(n) = \\ = x(n) + [q(n) - q(n-1)] \tag{13.62}$$

$$Y(Z) = X(Z) - Z^{-1}Q(Z) + Q(Z) = \\ = X(Z) + Q(Z)[1 - Z^{-1}] = \\ = X(Z) + Q(Z)\frac{Z-1}{Z} \tag{13.63}$$

Равенство (13.63) свидетельствует о том, что контур не влияет на входной сигнал, поскольку в контуре циркулирует только шум, и только шум испытывает влияние контура. Интегратор в обратной связи по шумовому сигналу превращается (с помощью контура обратной связи единичного усиления) в дифференциатор источника шума.

Удобный механизм отображения частотной передаточной функции предлагает z-плоскость (подобно своему эквиваленту, s-плоскости) (см. приложение Д). Такая функция обычно описывается как дробь, числитель и знаменатель которой имеют форму полиномов, причем корни последних считаются, соответственно, *нулями* и *полюсами* передаточной функции. Эти нули и полюсы могут рассматриваться как поверхность над плоскостью, представляющей модуль передаточной функции. Эту поверхность можно представить в виде резинового полотна, натянутого относительно земли на столбики, расположенные в полюсах, и притянутого к земле в нулевых положениях. Модуль частотной характеристики представляет собой уровень этой поверхности при обходе единичной окружности в z-плоскости (или ось $i\omega$ в s-плоскости). Отметим, что *передаточная функция шума* (noise transfer function — NTF), которая представляет собой функцию преобразования частоты контура, примененную к шуму, имеет полюс в начале координат и переходит через нуль в точке постоянной составляющей ($z = e^{i\theta}$, $\theta = 0$, так что $z = 1$). График, изображающий полюс и нуль функции NTF, спектральную характеристику NTF, а также типичный спектр входного сигнала представлены на рис. 13.26. Отметим, что нуль функции NTF расположен на постоянной составляющей, в окрестности которой шум квантования подавляется NTF. Таким образом, благодаря NTF возле постоянной составляющей нет значительного шума, и при этом спектр сигнала ограничен значительной передискретизацией, выполненной для того, чтобы спектр принадлежал малой окрестности вокруг постоянной составляющей с шириной примерно в 1,5% частоты дискретизации. Функцией восстанавливающего фильтра является подавление шума квантования вне полосы частот сигнала. Частота дискретизации на выходе фильтра теперь снижена для согласования с сокращенной полосой частот сигнала, практически свободного от шума. Дополнительное подавление шума может быть получено с помощью повышения порядка нуля функции NTF. Многие Σ-Δ-модуляторы созданы с функциями NTF, которые имеют нули второго или третьего порядка. Поскольку нули NTF обращают мощность выходного шума в нуль, вряд ли имеет значение, какой уровень мощности шума подан в контур обратной связи. Следовательно, большинство Σ-Δ-модуляторов создается для работы в системах, состоящих из 1-битовых преобразователей плюс несколько высокоточных модуляторов, каждый из которых работает с 4-битовыми преобразователями.

13.3.4.1. Шум Σ-Δ-модулятора

В предыдущем разделе упоминалось, что с помощью Σ-Δ-модулятора можно добиться улучшения SNR в квантованных данных за счет передискретизации. Рассмотрим, как это происходит при передискретизованных фильтрованных данных с шумом AWGN, а затем изучим тот же процесс со сформированным шумом.

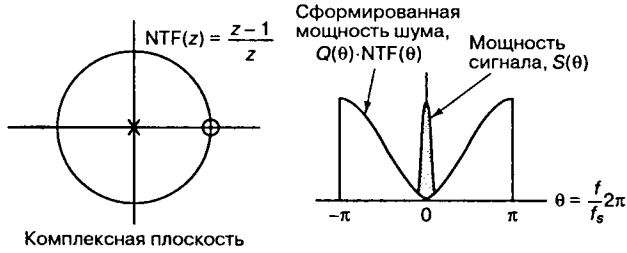


Рис. 13.26. Передаточная функция шума в z-плоскости, спектр мощности сигнала и сформированный шум Σ-Δ-модулятора

Если шум квантования белый, а сигнал дискретизуется с частотой, превосходящей частоту Найквиста, белый шум равномерно распределен в спектральном интервале, равном частоте дискретизации. Этот интервал называется *первой зоной Найквиста*, или *основной полосой*. Поскольку энергия шума квантования зафиксирована на величине $q^2/12$ (см. формулу (13.12)), спектральная плотность мощности шума квантования для сигнала, дискретизованного с частотой f_x, должна быть $q^2/(12f_x)$ Вт/Гц. Работа устройства квантования с повышенной частотой дискретизации уменьшает спектральную плотность мощности шума квантующего устройства в полосе частот сигнала. Передискретизованные данные могут численно фильтроваться с целью отсечения выходящего за полосу шума квантования, после чего можно снизить частоту дискретизации до частоты Найквиста. Если сигнал выбирается с частотой, вдвое превышающей частоту Найквиста, фильтрация отбросит половину мощности шума. Отсечение половины мощности шума сокращает среднеквадратическое значение амплитуды квантованного шума в $\sqrt{2}$ раз или мощности на 3 дБ. Чтобы уменьшить мощность шума на 6 дБ и таким образом улучшить шум квантования на 1 бит (см. формулу (13.24)), необходимо осуществить выборку с четырехкратной частотой и отсечь фильтром три четверти шума квантования. Итак, каждое удвоение частоты произведения выборки относительно частоты Найквиста приводит к улучшению SNR преобразователя белого шума на 3 дБ (или половину бита).

Рассмотрим частоту, на которой можно улучшить SNR уже сформированного шума преобразователя, производящего выборку с повышенной частотой. Передаточная функция шума формирующего Σ-Δ-фильтра имеет нуль на постоянной составляющей, что приводит к нулю второго порядка в спектральной характеристике мощности фильтра. Если разложить спектральную характеристику фильтра в ряд Тейлора и отбросить все члены после первого ненулевого слагаемого, получим следующую простую аппроксимацию зависимости фильтра, справедливую в окрестности спектра сигнала.

$$
\begin{aligned}
H^2(\omega) &= \left[2\sin\left(\frac{\omega}{2\omega_s}\right)\right]^2 = \\
&= 2\left[1-\cos\left(\frac{\omega}{\omega_s}\right)\right] = \\
&= 2\left\{1-\left[1-\frac{1}{2!}\left(\frac{\omega}{\omega_s}\right)^2+\ldots+\right]\right\} \approx \left(\frac{\omega}{\omega_s}\right)^2 = \left(\frac{f}{f_s}\right)^2
\end{aligned}
\qquad (13.64)
$$

Здесь f_s — частота дискретизации модулятора.

Мощность сформированного шума, "выжившая" после прохождения фильтра нижних частот, который следует за Σ-Δ-модулятором, имеет следующий вид.

$$N(\omega) = \frac{N_0}{2}\int_{-f_{BW}}^{f_{BW}} \left(\frac{f}{f_s}\right)^2 df = \frac{N_0}{2}\frac{1}{3}\left(\frac{f}{f_s}\right)^3\Bigg|_{f=-f_{BW}}^{f=f_{BW}} = \frac{N_0}{3}\left(\frac{f_{BW}}{f_s}\right)^3 \qquad (13.65)$$

Отношение слагаемого шума для сигнала, выбранного с частотой f_s, к слагаемому шума для сигнала, выбранного с частотой $2f_s$, с последующей фильтрацией до той же выходной полосы частот сигнала f_{BW} равно порядка 8–9 дБ. Таким образом, Σ-Δ-модулятор с единственным нулем в функции NTF улучшает SNR на 9 дБ или на 1,5 бит при удвоении частоты дискретизации. Сигма-дельта-модуляторы, созданные с множественными цифровыми интеграторами и контурами обратной связи, имеют большее число переходов через нуль в NTF. Выполнив аналогичные выкладки, можно найти, что NTF Σ-Δ-модулятора с 2 и 3 нулями улучшает SNR на 15 и 21 дБ (или 2,5 и 3,5 бит). Таким образом, двухнулевой Σ-Δ-модулятор, работающий с частотой, в 64 раза (или удвоенной шесть раз) превышающей частоту Найквиста, дает улучшение SNR на 90 дБ. Спектр, изображенный на рис. 13.27, был образован двухнулевым Σ-Δ-модулятором, и если учесть (1) потери в 6 дБ вследствие спектрального разложения реального сигнала, (2) уменьшение на 2 дБ амплитуды относительно полномасштабного сигнала, (3) потери в 3 дБ вследствие отбрасывания членов дискретного преобразования Фурье, то уровень шума на 79 дБ находится ниже спектрального максимума.

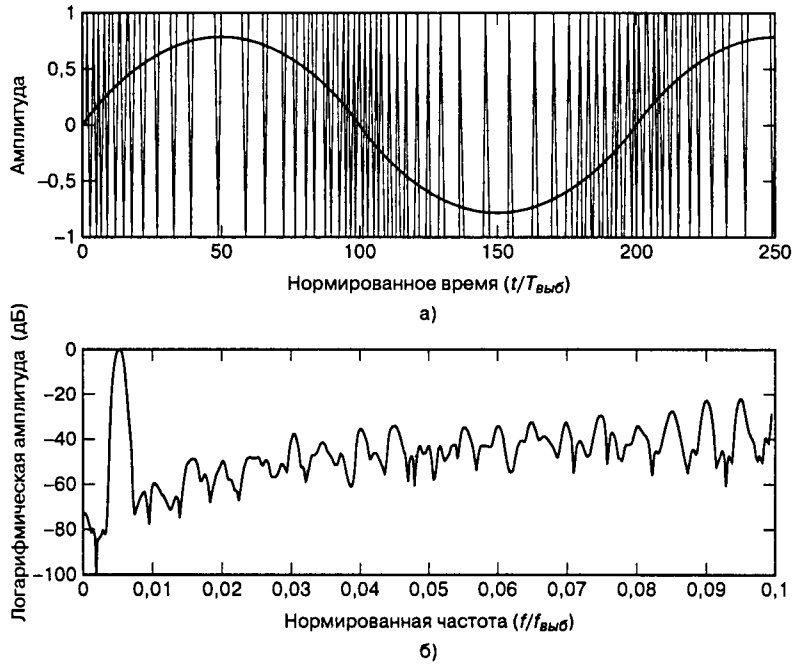


Рис. 13.27. Однобитовый Σ-Δ-модулятор: а) входной и выходной временные ряды; б) спектральная характеристика

13.3.5. Сигма-дельта-аналого-цифровой преобразователь

Σ-Δ-аналого-цифровой преобразователь (analog-to-digital converter — ADC, АЦП) обычно реализуется как интегральная схема, построенная на основе Σ-Δ-модулятора. Для образования полной системы схема должна содержать вспомогательные подсистемы: аналоговый фильтр защиты от наложения спектров (anti-alias filter), схему выборки-запоминания (sample-and-hold circuit), интегратор на переключаемых конденсаторах для модулятора (switched-capacitor integrator), цифро-аналоговый преобразователь (digital-to-analog converter — DAC, ЦАП) с обратной связью и цифровой фильтр повторной выборки (resampling filter). Вследствие высокой передискретизации, аналоговый фильтр защиты от наложения спектров может представлять собой просто *RC*-цепь с широкой полосой перехода, захватывающей многие октавы. Цифро-аналоговый преобразователь необходим для формирования аналогового сигнала обратной связи. Поскольку ЦАП включен в контур обратной связи, он не выигрывает от изменения коэффициента обратной связи, и, следовательно, его линейность и точность должны соответствовать уровню производительности всей системы. Σ-Δ-модулятор сохраняет точность сигнала в ограниченном сегменте дискретного спектра. Для доступа к этому сегменту спектра высокой точности выход модулятора должен быть отфильтрован и дискретизован с пониженной частотой. Фильтр последующей обработки, расположенный за схемой модуляции, отбрасывает внешний шум, расположенный в полосе частот, существующей вследствие передискретизации. Обычно это фильтр повторной выборки с линейным изменением фазы и конечной импульсной характеристикой.

На рис. 13.27, *а* изображен входной синусоидальный сигнал, выбираемый с повышенной частотой, и соответствующий выходной сигнал однобитового Σ-Δ-модулятора с двумя нулями. На рис. 13.27, *б* представлена спектральная характеристика выходного ряда. Отметим, что спектр сформированного шума в окрестности сигнала находится приблизительно на 80 дБ ниже максимума спектра входной синусоиды. Отметим также, что амплитуды выходного сигнала ограничены диапазоном ±1 и контур, по сути, выполняет модуляцию квадратного сигнала пропорционально амплитуде входного сигнала. На рис. 13.28 представлены временной ряд и спектр, полученный на выходе фильтра с дискретизацией на пониженной частоте, следующего за модулятором.

13.3.6. Сигма-дельта-цифро-аналоговый преобразователь

Σ-Δ-модулятор, изначально разрабатываемый как блок в АЦП, выполняет основную часть цифро-аналогового преобразования. Практически все высококачественное аудиооборудование и большинство цифро-аналоговых преобразователей систем связи снабжены Σ-Δ-конвертерами. Процесс использует Σ-Δ-модулятор как цифро-цифровое преобразование, которое преобразует высокоточное (скажем, 16-битовое) представление передискретизованных цифровых данных в представление низкой точности (скажем, 1-битовое). Передискретизованный однобитовый поток данных затем доставляется в 1-битовый ЦАП с двумя аналоговыми выходными уровнями, определенными с той же точностью, что и 16-битовый преобразователь. Преимущество использования однобитового ЦАП с высокой скоростью, но только с двумя уровнями, состоит в том, что скорость — это менее дорогой ресурс, чем точность. 2-уровневый высокоскоростной ЦАП заменяет ЦАП низкой скорости, который мог бы разрешить 65 536 различных уровней.

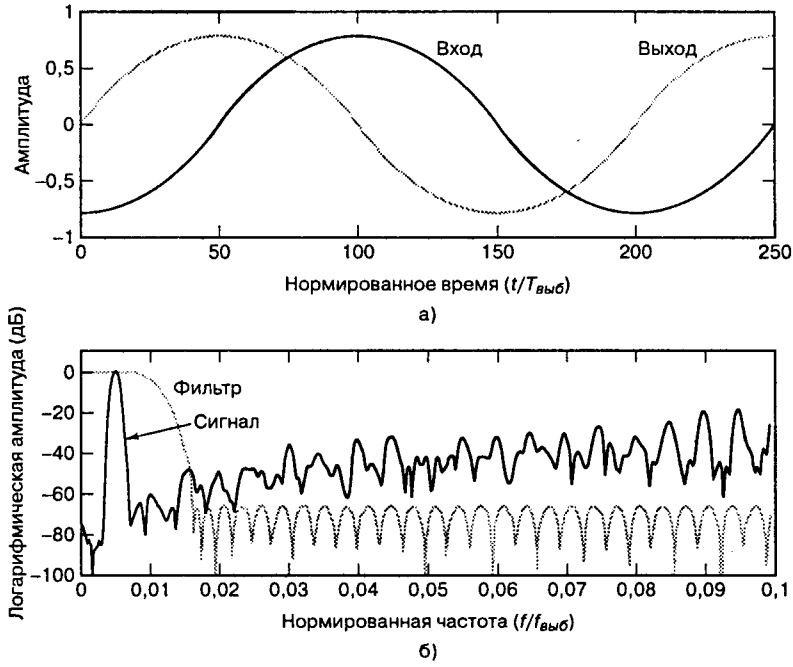


Рис. 13.28. Фильтр последующей обработки, следующий за Σ-Δ-модулятором: а) входной и выходной временные ряды; б) спектральная характеристика

Очень простая аналоговая фильтрация низкого уровня, следующая за 1-битовым ЦАП, подавляет спектр внеполосного шума и выдает исходные цифровые данные с высокой точностью и в сокращенной полосе частот. Повторное квантование перевыбранных данных представляет собой обработку сигнала с использованием цифрового Σ-Δ-модулятора. Единственная дополнительная задача, которую требуется выполнить при использовании Σ-Δ-ЦАП, состоит в необходимости увеличения частоты произведения выборки в 64 раза, по сравнению с частотой Найквиста. Это выполняется с помощью интерполирующего фильтра, работающего на основе методов цифровой обработки сигналов; этот фильтр представляет собой стандартный блок, который имеется в большинстве систем, использующих ЦАП для перехода между источником цифрового сигнала и аналоговым выходом [12].

В качестве стандартной иллюстрации процесса рассмотрим проигрыватель компакт-дисков, использующий интерполирующий фильтр для реализации преобразования с четырехкратным повышением частоты, приводящего к отделению периодического спектра, который связан с дискретными данными. Это позволяет сглаживающему фильтру, который следует за ЦАП, иметь более широкую полосу частот и, следовательно, меньшее число компонентов и меньшую стоимость реализации. Спецификация компакт-диска содержит такие термины, как, например, "4-to-1 oversampled" ("перевыбран с четырехкратной частотой"), чтобы отразить наличие интерполирующих фильтров. После того как с помощью интерполятора 1:4 будет выполнено четырехкратное увеличение частоты дискретизации, дальнейшее преобразование с использованием недорогого интерполирующего фильтра 1:16 является простой задачей. Для завершения аналогового процесса преобразования данные (теперь выбран-

ные с 64-кратной частотой) подаются на полноцифровой Σ-Δ-модулятор и однобитовый ЦАП. Эта структура изображена на рис. 13.29.

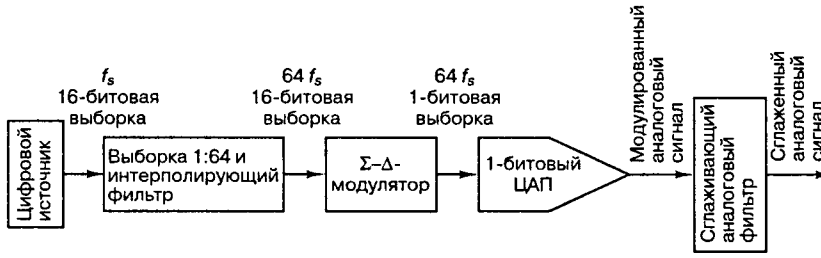


Рис. 13.29. Схема прохождения сигнала в Σ-Δ-цифро-аналоговом преобразователе

Существует много сигналов, которые по отношению к полосе частот сигнала выбираются с очень большой частотой. Эти сигналы могут быть легко преобразованы в аналоговую форму с использованием Σ-Δ-модулятора и 1-битового ЦАП. Примерами являются контрольные сигналы схем АРУ, несущие ГУН и сигналы синхронизации ГУН. Многие системы используют Σ-Δ-модулятор и 1-битовый ЦАП для генерации и формирования аналоговых сигналов управления.

13.4. Адаптивное предсказание

Усиление предсказания, которое получается в классических кодерах с предсказанием, пропорционально отношению дисперсии сигнала к дисперсии ошибки предсказания. Это объясняется тем, что при фиксированном уровне шума квантования требуется меньше бит для описания сигнала с меньшей энергией. Полезность кодера с предсказанием ограничена возможными рассогласованиями между сигналом источника и предсказывающим фильтром. Источники рассогласования связаны с переменным во времени поведением (т.е. нестационарностью) распределения амплитуды и спектральных или корреляционных свойств сигнала. Адаптивные кодеры (медленного действия) включают вспомогательные схемы для оценки параметров, требуемых для получения локальной оптимальной производительности. Эти вспомогательные цепи периодически программируют модификации для предсказания параметров цепи и таким образом избегают рассогласования предсказания. Комитет CCITT (International Telegraph and Telephone Consultative Committee — Международный консультативный комитет по телеграфии и телефонии, МККТТ) в качестве стандарта качественной телефонной связи выбрал адаптивную дифференциальную импульсно-кодовую модуляцию (Adaptive Differential Pulse Code Modulation — ADPCM) со скоростью 32 Кбит/с. Это дает экономию скорости передачи бит 2:1 по сравнению с 64 Кбит/с схемы PCM с логарифмическим сжатием.

13.4.1. Прямая адаптация

В алгоритмах прямой адаптации входные данные, которые должны быть закодированы, буферизуются и обрабатываются с целью получения локальных статистик, таких как первые N выборочных значений автокорреляционной функции. Корреляционное значение $R_x(0)$ с нулевым запаздыванием является кратковременной оценкой

дисперсии. Эта оценка используется для согласования автоматической регулировки усиления с целью получения оптимального согласования масштабированного входного сигнала с динамической областью устройства квантования. Этот процесс обозначается "AQF" от "adaptive quantization forward control" — *контроль прямым адаптивным квантованием.* Остающиеся $N-1$ корреляционных оценок используются для получения новых коэффициентов для фильтра с предсказанием. Этот процесс называется контролем прямым адаптивным предсказанием (adaptive prediction forward — APF). На рис. 13.30 изображена эта форма адаптивного алгоритма. Это расширение структуры, представленной на рис. 13.20. Здесь предсказывающие коэффициенты выводятся из входных данных, теперь называемых *побочной информацией* (side information). Они должны быть переданы вместе с ошибками предсказания с кодера на декодер. Скорость изменения этих адаптивных коэффициентов связана со временем, в течение которого входной сигнал может считаться локально стационарным. Например, речь, вызываемая механическим смещением речевых артикуляторов (язык, губы, зубы и т.д.), не может изменять характеристики быстрее, чем 10 или 20 раз за секунду. Это дает интервал обновления от 50 до 100 мс. Использование арифметически простых, но субоптимальных алгоритмов оценивания для вычисления локальных параметров фильтра делает необходимым более высокую скорость изменения. Для вычисления параметров 10–12-отводного фильтра принят интервал изменения 20 мс. На 10-отводных фильтрах можно получить усиление предсказания от 10 до 16 дБ, если используется адаптация с прямой связью и кодеры с предсказанием [13].

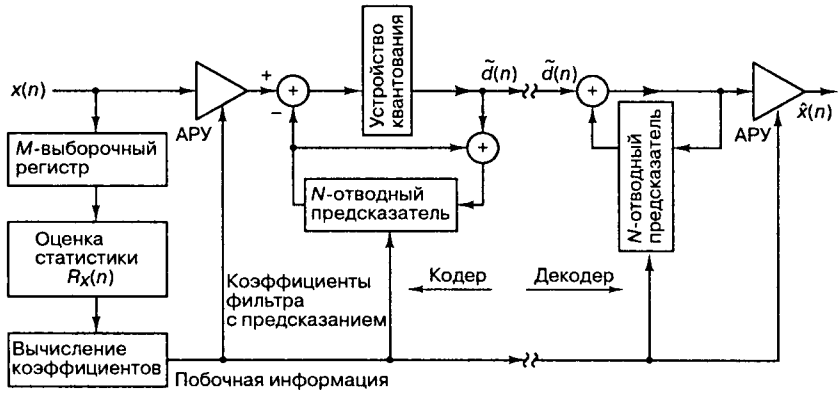


Рис. 13.30. Прямое адаптивное предсказание и кодирование квантования

13.4.2. Синтетическое/аналитическое кодирование

Изучаемые до сих пор схемы кодирования можно назвать *кодерами формы сигналов.* Они создают аппроксимации входных сигналов, минимизирующие некоторую меру расстояния между сигналом и аппроксимацией. Эти технологии являются общими и могут применяться к любому источнику сигнала. С другой стороны, синтетические/аналитические кодеры являются сильно сигнально-зависимыми. В частности, они созданы в основном для речевых сигналов. Эти кодеры играют на том, что слуховой механизм реагирует на амплитудное содержание кратковременного спектра сигнала, но при этом почти нечувствителен к его фазовой структуре.

Таким образом, этот класс кодеров формирует восстановленный сигнал, аппроксимирующий амплитуду и изменяющуюся во времени характеристику последовательности кратковременного спектра сигнала, но не делает попыток сохранить его относительную фазу.

Спектральные характеристики речи кажутся стационарными в течение порядка 20–50 мс. Существует множество технологий, которые анализируют спектральные характеристики голоса каждые 20 мс и используют результаты этого анализа для синтеза сигнала, дающего тот же кратковременный спектр мощности. Некоторые методы применяют модель механизма генерации речи, для которого параметры модели должны быть оценены с частотой обновления. Этот тип кодера наилучшим образом представлен в своих различных формах как линейный кодер с предсказанием (linear predictive coder — LPC). Разновидности кодеров LPC оперируют сигналом с помощью комбинаций спектральных модификаций и временных делений, которые, используя побочную информацию, сокращают количество временных выборок, требуемых для правильного воссоздания исходного спектра. Общим для всех синтетических/аналитических кодеров, используемых для речевых сигналов, является отсутствие необходимости в том, чтобы голосовой сигнал "выглядел" как оригинальный; достаточно, чтобы он "звучал" подобно ему.

13.4.2.1. Линейное кодирование с предсказанием

Адаптивные предсказатели, описанные в разделе 13.3.2, были созданы для предсказания или создания хороших оценок входного сигнала. В адаптивной форме предсказываемые коэффициенты вычисляются как побочная информация на основе периодического изучения входных данных. Затем разность между входом и предсказанием передается получателю для разрешения ошибки предсказания. *Линейные кодеры с предсказанием* (linear predictive coder — LPC) являются естественным расширением N-отводных кодеров с предсказанием. Если коэффициенты фильтра периодически вычисляются с помощью оптимального алгоритма, предсказание является настолько хорошим, что (в основном) информации об ошибке предсказания, которую нужно передавать приемнику, не существует. Вместо того чтобы передавать эти ошибки предсказания, система LPC передает коэффициенты фильтра и озвученное/неозвученное руководство к действию для фильтра. Таким образом, единственными данными, посланными в LPC, является высококачественная побочная информация классического адаптивного алгоритма. Модель LPC для синтеза голоса изображена на рис. 13.31. Кодеры LPC представляют собой ядро из смешанных кодеров, которое включает в себя кодер и управляющий генератор в контуре анализа через синтез, предназначенном для минимизации разности между входным и синтезированным сигналами. В сотовых телефонах для получения качественной связи со скоростью передачи данных ниже 9,6 Кбит/с используются кодеры PRE (Regular-Pulse Excited — активация регулярными импульсами) и CELP (Codebook-Excited Linear Predictive — линейное предсказание, активируемое кодовой книгой). В системе GSM (Global Systems for Mobile — глобальная система мобильной связи) используется сжатие RPE, тогда как для мобильных телефонных систем, созданных согласно стандарту IS-95 относительно множественного доступа с кодовым разделением каналов (code division multiple access — CDMA), применяется вариант CELP. Дополнительный материал по CELP представлен в разделе 13.8.1.3.

Эта модель, использующая 12-отводный синтезатор речи, нашла применение в детских говорящих играх. Дальнейшее рассмотрение методов LPC, используемых для речи, приводится в разделе 13.8.1.

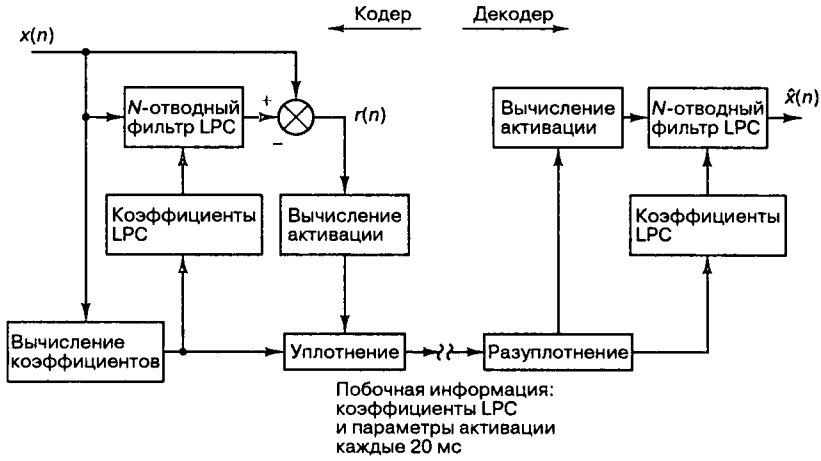


Рис. 13.31. Блочная диаграмма: моделирование речи с помощью линейного кодера с предсказанием

13.5. Блочное кодирование

Изучаемые до сих пор устройства квантования были *скалярными* по своей природе, поскольку они образовывали *единственную выходную выборку*, основанную на настоящей входной выборке и (возможно) N предшествующих выходных выборках. С другой стороны, блочные кодеры образуют *вектор выходных выборок*, основанный на настоящей и N предшествующих входных выборках. *Эффективность кодирования* (coding gain) сигнала представляет собой отношение входного SNR кодера к выходному. Если дисперсии шума на входе и выходе равны, эта эффективность просто представляет собой отношение входной дисперсии сигнала к выходной. Из данного отношения следует, что каждый бит разности между числом входных бит на выборку и средним числом выходных бит на выборку равносилен изменению эффективности на 6 дБ. Блочные кодеры могут давать впечатляющую эффективность кодирования. В среднем они могут представлять последовательности, квантованные по 8 бит, всего с 1 или 2 бит на выборку [8]. Технология блочного кодирования меняется, но общим является отображение входной последовательности в альтернативную систему координат. Это может быть отображение в подпространство большего пространства, так что отображение может быть необратимым [8]. В качестве альтернативы может быть использована информационно-зависимая схема редактирования для идентификации подпространства отображения, из которого получены квантованные данные. Технологии блочного кодирования часто классифицируются по своим схемам отображения, которые включают, например, векторные устройства квантования, кодеры различных ортогональных преобразований, кодеры с разделением по каналам, такие как кодер с многополосным кодированием. Блочные кодеры далее описываются через свои алгоритмические структуры, такие как кодовая книга, дерево, решетка, дискретное преобразование Фурье, дискретное косинус-преобразование, дискретное преобразование Уолша-

Адамара (Walsh-Hadamard), дискретное преобразование Карунена-Лоэва (Karhunen-Loeve) и кодеры с блоком квадратурных зеркальных фильтров. Итак, изучим некоторые схемы блочного кодирования.

13.5.1. Векторное квантование

Векторные устройства квантования представляют собой обобщение общепринятых скалярных устройств квантования. При скалярном квантовании для представления входной выборки скалярное значение выбирается из конечного множества возможных значений. Значение выбирается близким (в некотором смысле) к выборке, которую оно представляет. Мерой точности являются различные взвешенные среднеквадратические меры, которые поддерживают интуитивную концепцию расстояния в терминах обычной векторной длины. Обобщая, имеем, что в векторном квантовании вектор выбирается из конечного перечня возможных векторов, представляющих входной вектор выборки. Вектор выборки является близким (в некотором смысле) к вектору, который он представляет.

Каждый входной вектор может быть представлен точкой в N-мерном пространстве. Устройство квантования определяется с помощью деления этого пространства на множество неперекрывающихся объемов [14]. Эти объемы называются интервалами, полигонами и политопами, соответственно, для одно-, двух- и N-мерных векторных пространств. Задача векторного квантующего устройства состоит в определении объема, в котором расположен входной вектор. Выходом оптимального квантующего устройства является вектор, определяющий центр тяжести этого объема. Как и в одномерном квантующем устройстве, среднеквадратическая ошибка зависит от расположения границы деления и многомерной функции плотности вероятности входного вектора.

Описание векторного устройства квантования может рассматриваться как две точные задачи. Первая — это задача создания кода. Она связана с созданием многомерного объема квантования (или деления) и выбором допустимых выходных последовательностей. Вторая задача состоит в использовании кода и связана с поиском определенного объема при данном делении, который соответствует (согласно некоторому критерию точности) наилучшему описанию источника. Форма алгоритма, выбранного для контроля сложности кодирования и декодирования, может объединять две задачи — деление и поиск. Стандартными методами векторного кодирования являются алгоритмы кодовых книг, древовидные и решетчатые алгоритмы кодирования [15, 16].

13.5.1.1. Кодовые книги, древовидные и решетчатые кодеры

Кодеры, использующие кодовые книги, — это, по сути, алгоритмы поиска в таблице. Перечень возможных шаблонов (кодовых слов) внесен в память кодовой книги. Каждый шаблон снабжен адресом или точечным индексом. Программа кодирования ищет среди шаблонов тот, что расположен ближе всего к входному шаблону, и передает получателю адрес, сообщающий, где этот шаблон может быть найден в его кодовой книге. Древовидные и решетчатые кодеры являются последовательными. Таким образом, допустимые кодовые слова кода не могут выбираться независимо, они должны иметь структуру, которой можно управлять с помощью узловых точек. Это подобно структуре последовательных алгоритмов обнаружения-коррекции ошибок, которые обходят граф при образовании ветвящейся весовой аппроксимации входной последовательности (см. раздел 6.5.1). Древовидный граф подвержен экспоненциальному рос-

ту памяти при увеличении размерности или глубины. Решетчатый граф снижает проблему размерности, поскольку позволяет одновременно отслеживать выбранные траектории и связанные с ними траекторно-весовые метрики, называемые *интенсивностью* (см. раздел 6.3.3).

13.5.1.2. Совокупность кода

Кодовые векторы, внесенные в кодовую книгу, дерево или решетку, являются подобными или типичными векторами. Первый этап создания кода, в котором определяются вероятные кодовые векторы, называется *заселением* кода. Классические методы определения совокупности кодов есть *детерминированными*, *стохастическими* и *итеративными*. Детерминированная совокупность является перечнем предопределенных возможных выходов, основанных на простом субоптимальном или принятом пользователем критерии точности или на простом алгоритме декодирования. Примером детерминированного метода может служить кодирование выборок в трехмерном пространстве красного, зеленого и синего (RGB) компонентов цветного телевизионного сигнала. Для глаза не характерна одинаковая разрешающая способность для каждого цвета, так что кодирование может быть применено независимо к каждому цвету, чтобы отразить эту особенность восприимчивости. Результирующими объемами квантования могут быть прямоугольные параллелепипеды. Проблемой при независимом квантовании является то, что образы видны не в этой системе координат, а в координатах яркости, оттенка и насыщенности. Например, черно-белая фотография использует только координату яркости. Таким образом, независимо квантованные координаты RGB не приводят к уменьшению объема воспринимаемого пользователем искажения данного числа бит. Чтобы получить уменьшенное искажение, квантующие устройства должны разделить свое пространство на области, которые отражают деление в альтернативном пространстве. В качестве альтернативы, квантование может производиться независимо в альтернативном пространстве с использованием преобразующего кодирования, изучаемого в разделе 13.6. Детерминированное кодирование является наиболее простым для реализации, но дает наименьшую эффективность кодирования (наименьшее сокращение в скорости передачи бит при данном SNR).

Стохастическая совокупность должна выбираться на основании предполагаемой функции плотности вероятности входных выборок. Итеративные решения для оптимальных делений существуют и могут быть определены для любых предполагаемых функций плотности вероятности. Общие выборки моделируются с помощью предполагаемых функций плотности вероятности. При отсутствии таких функций могут использоваться итеративные методы, основанные на большой совокупности последовательностей испытаний, для получения разбиения и выходной совокупности. Последовательности испытаний могут включать в себя десятки тысяч входных выборок.

13.5.1.3. Поиск

При данном входном векторе и заселенной кодовой книге, дереве или решетке, алгоритм кодера должен производить поиск для определения наиболее адекватного векторного представителя. Исчерпывающий поиск среди всех возможных представителей будет гарантировать наилучшее отображение. Работа кодера улучшается для пространств большей размерности, но это приводит к росту сложности. Исчерпывающий поиск в пространстве большей размерности может быть весьма трудоемким. Альтернатива — следовать неисчерпывающей, субоптимальной схеме поиска с приемлемо малыми ухудшениями формы оптимальной траектории. Вообще, при выборе алгорит-

мов поиска основными аргументами часто являются требования памяти и вычислительной сложности. Примеры алгоритмов поиска включают в себя алгоритмы единичной траектории (ветвь наилучшего выживания), алгоритмы множественной траектории и двоичные (метод последовательной аппроксимации) алгоритмы кодовой книги. Большинство алгоритмов поиска делают попытку определить и отбросить нежелательные модели без проверки всей модели.

13.6. Преобразующее кодирование

В разделе 13.5.1 изучались векторные устройства квантования в терминах множества вероятных моделей и технологий для определения одной модели во множестве, наиболее близком к входной модели. Одной из мер качества аппроксимации является взвешенная среднеквадратическая ошибка.

$$d(\mathbf{X},\hat{\mathbf{X}}) = (\mathbf{X}-\hat{\mathbf{X}})\mathbf{B}(\mathbf{X})(\mathbf{X}-\hat{\mathbf{X}})^T, \tag{13.76}$$

где $\mathbf{B}(\mathbf{X})$ — это весовая матрица, а $\mathbf{X}^T$— транспонированный вектор $\mathbf{X}$. Минимизация может быть вычислительно проще, если весовая матрица является диагональной. Диагональная весовая матрица дает координатное множество с нарушенной связью (некоррелированное), так что ошибка минимизации вследствие квантования может находиться независимо по каждой координате.

Таким образом, преобразующее кодирование включает следующую последовательность операций, которые изображены на рис. 13.32.

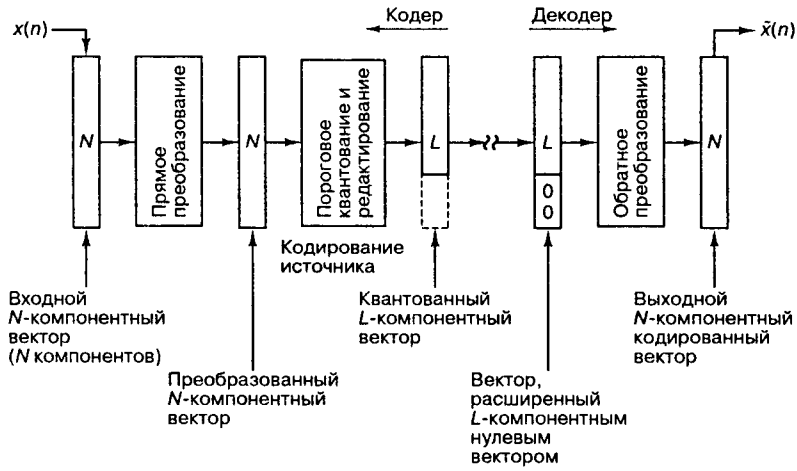


Рис. 13.32. Блочная диаграмма: преобразующее кодирование

1. К входному вектору применяется обратимое преобразование.
2. Коэффициенты преобразования квантуются.
3. Квантованные коэффициенты передаются и получаются.
4. Преобразование обращается с использованием квантованных коэффициентов.

Отметим, что при преобразовании не выполняется никакого кодирования источника; просто допускается более удобное описание вектора сигнала, которое позволяет легче

использовать кодирование источника. Задача преобразования состоит в отображении коррелированной входной последовательности в другую систему координат, в которой координаты имеют меньшую корреляцию. Напомним, что это в точности представляет собой задачу, выполняемую кодером с предсказанием. Кодирование источника происходит посредством присвоения битового значения различным коэффициентам преобразования. Как часть этого присвоения, коэффициенты могут быть разделены на подмножества, которые квантуются с помощью различного числа бит, но *не* с помощью различных размеров шага квантования. Это присвоение отражает динамическую область (дисперсию) каждого коэффициента и может быть взвешено мерой, отражающей важность (относительно человеческого восприятия) элемента, переносимого каждым коэффициентом [17]. Например, подмножество коэффициентов может быть сведено к нулевой амплитуде или может быть квантовано с помощью 1 или 2 бит.

Преобразование может быть независимым от вектора данных. Примерами таких преобразований являются дискретное преобразование Фурье (discrete Fourier transform — DFT, ДПФ), дискретное преобразование Уолша-Адамара (discrete Walsh-Hadamar transform — DWHT), дискретное косинус-преобразование (discrete cosine transform — DCT, ДКП) и дискретное наклонное преобразование (discrete slant transform — DST). Преобразование может быть также получено из вектора данных, как это делается в дискретном преобразовании Карунена-Лоэва (discrete Karhunen-Loeve transform — DKLT), иногда называемом *преобразованием основного компонента* (principal component transform — PCT) [18]. Независимые от данных преобразования являются самыми простыми в реализации, но они не так хороши, как информационно-зависимые. Зачастую вычислительная простота является достаточным оправданием для использования независящих от данных преобразований. При хорошем субоптимальном преобразовании потери эффективности кодирования незначительны (как правило, меньше 2 дБ), и обычно при демонстрации рабочих характеристик упоминается ухудшение качества.

13.6.1. Квантование для преобразующего кодирования

Преобразующие кодеры обычно называются спектральными, поскольку сигнал описывается через свое спектральное разложение (в выбранном базисном множестве). Спектральные члены вычисляются для неперекрывающихся последовательных блоков входных данных. Таким образом, выход преобразующего кодера может рассматриваться как множество временных рядов, один ряд для каждого спектрального члена. Дисперсия каждого ряда может быть определена, и каждый ряд может быть квантован с использованием разного числа бит. Допуская независимое квантование каждого коэффициента преобразования, имеем возможность распределения фиксированного числа бит среди коэффициентов преобразования для получения минимальной ошибки квантования.

13.6.2. Многополосное кодирование

Преобразующие кодеры в разделе 13.6 были описаны как выполняющие деление входного сигнала на множество медленно изменяющихся временных рядов, каждый из которых связан с определенным базисным вектором преобразования. Спектральные члены (скалярные произведения данных с базисными векторами) вычисляются с помощью множества скалярных произведений. Множество скалярных произведений может быть вычислено с помощью множества фильтров с *конечной импульсной характеристикой* [19]. С этой целью преобразующий кодер может рассматриваться как выполняющий разделение полосы частот входных данных на отдельные каналы. Обоб-

щая, получим, что *многополосный кодер*, который выполняет спектральное разделение полосы частот на отдельные каналы с помощью набора непрерывных узкополосных фильтров, может рассматриваться в качестве частного случая преобразующего кодера. (Типичный многополосный кодер изображен на рис. 13.33.)

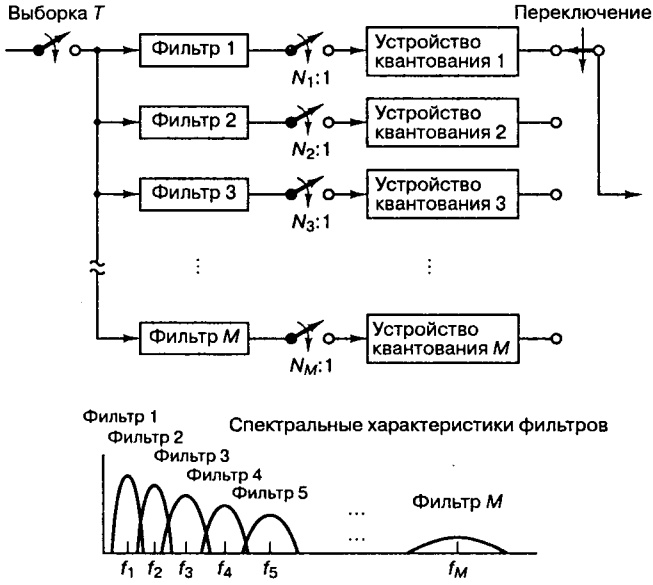


Рис. 13.33. Многополосное кодирование

Спектральное разложение данных (как и фильтрование) допускает различное формирование класса специальных базисных множеств (т.е. спектральных фильтров), в частности базисных множеств, которые отражают приемлемые предпочтения пользователя и модели источника. Например, шум квантования, сгенерированный в полосе частот с большой дисперсией, будет ограничен этой полосой частот; он не будет проникать в соседнюю полосу частот, имеющую низкую дисперсию и, следовательно, уязвимую для низкоуровневых сигналов, которые замаскированы шумом. Имеем также выбор формирующих фильтров с равными или неравными полосами частот (рис. 13.33). Таким образом, можно независимо каждой подполосе приписать выборочную частоту, соответствующую ее ширине полосы частот, и число бит квантования, соответствующее ее дисперсии. Для сравнения, в общепринятом преобразующем кодировании амплитуда каждого базисного вектора выбирается с одинаковой частотой.

Многополосный кодер может быть создан как трансмультиплексор (преобразователь вида уплотнения). Здесь входной сигнал рассматривается в виде составленного из множества базисных функций, моделированных как независимые подканалы узкой полосы частот. Кодер разделяет входной сигнал на множество каналов с низкой скоростью передачи данных, уплотненных с временным разделением (time-division multiplexing — TDM). После квантования и передачи декодер обращает процесс фильтрации и повторной выборки, преобразуя каналы TDM обратно в исходный сигнал. При классическом подходе к этому процессу можно использовать множество узкополосных фильтров с этапами смешивания, фильтрации нижних частот и дискретизации на пониженной частоте (часто называемой *децимацией*, или *прореживанием*). Эта операция фильтрации со-

кращает входную полосу частот до выбранной полосы частот канала и повторно выбирает сигнал до самой низкой частоты, что позволяет избежать наложения разделенных полос частот данных. В приемнике производится обратный процесс. Разделенные на полосы данные для увеличения их частоты до желаемой частоты дискретизации проходят через интерполирующие фильтры и смешиваются обратно до их соответствующего спектрального положения. Чтобы создать исходный смешанный сигнал, они объединяются. Для кодирования речи или, в более общем смысле, для сигналов, которые связаны с механическим резонансом, желательны группы фильтров с неравными центральными частотами и неравными полосами частот. Такие фильтры называются пропорциональными наборами фильтров. Эти фильтры имеют логарифмически расположенные центральные частоты и полосы частот, пропорциональные центральным частотам. При рассмотрении на логарифмической шкале такое пропорциональное размещение выглядит как равномерное расположение полос частот и отражает спектральные свойства многих физических акустических источников.

13.7. Кодирование источника для цифровых данных

Кодирование с целью сокращения избыточности источника данных обычно влечет за собой выбор эффективного двоичного представления этого источника. Часто это требует замены двоичного представления символов источника альтернативным представлением. Замена обычно является временной и производится, для того чтобы достичь экономии при запоминании или передаче символов дискретного источника. Двоичный код, присвоенный каждому символу источника, должен удовлетворять определенным ограничениям, чтобы позволить обращение замены. К тому же код может быть далее ограничен спецификацией системы, например ограничениями памяти и простотой реализации.

Мы настолько привыкли к использованию двоичных кодов для представления символов источника, что можем забыть о том, что это всего лишь один из вариантов присвоения. Наиболее общим примером этой процедуры является двоичное присвоение количественным числительным (даже не будем рассматривать отрицательные числа). Можно прямо переводить в двоичную систему счисления, двоичные коды восьмеричных чисел, двоичные коды десятичных чисел, двоичные коды шестнадцатеричных чисел, десятичные коды "два из пяти", десятичные коды с избытком три и т.д. В этом примере при выборе соответствия учитывается простота вычисления, определения ошибки, простота представления или удобство кодирования. Для определенной задачи сжатия данных основной целью является *сокращение количества бит.*

Конечные дискретные источники характеризуются множеством различных символов, $X(n)$, где $n = 1, 2, \ldots, N$ — *алфавит источника*, а n — индекс данных. Полное описание требует вероятности каждого символа и совместных вероятностей символов, выбранных по два, три и т.д. Символы могут представлять двухуровневый (двоичный) источник, такой как черно-белые уровни факсимильного изображения, или многосимвольный источник, такой как 40 общих знаков санскрита. Еще одним общим многосимвольным алфавитом является клавиатура компьютерного терминала. Эти недвоичные символы отображаются посредством словаря, называемого знаковым кодом, в описание с помощью двоичного алфавита. (На рис. 2.2 представлен код ASCII, а на рис. 2.3 — код EBCDIC.) Стандартные знаковые коды имеют фиксированную длину, такую как 5–7 бит. Длина обычно выбирается так, чтобы существовало достаточно двоичных знаков для того, чтобы присвоить единственную двоичную последовательность каждому вход-

ному знаку алфавита. Это присвоение может включать большие и маленькие буквы алфавита, цифры, знаки пунктуации, специальные знаки и знаки управления, такие как знак забоя, возврата и т.д. Коды фиксированной длины обладают следующим свойством: знаковые границы отделены фиксированным числом бит. Это допускает превращение последовательного потока данных в параллельный простым счетом бит.

Двухкодовые стандарты могут определять один и тот же символ разными способами. Например, (7-битовый) код ASCII имеет достаточно бит, чтобы присвоить различные двоичные последовательности большой и маленькой версиям каждой буквы. С другой стороны, (5-битовый) код Бодо, который обладает только 32 двоичными последовательностями, не может сделать то же самое. Для подсчета полного множества знаков код Бодо определяет два контрольных знака, называемых *переключением на печатание букв* (letter shift — LS) и *переключением на печатание цифр* (figure shift — FS), которые должны использоваться как префиксы. При использовании эти контрольные знаки переопределяют отображение символа в двоичную форму. Это напоминает *клавишу переключения регистра* (shift key) на печатающем устройстве; эта клавиша полностью переопределяет новое множество символов на клавиатуре. Клавиатуры некоторых калькуляторов также имеют две клавиши переключения регистров, так что каждое нажатие клавиши имеет три возможных значения. Кроме того, некоторые команды текстового процессора используют двойные и тройные командные функции. В некотором смысле эти двух- и трехсловные команды представляют собой кодовое присвоение переменной длины. Эти более длинные кодовые слова присваиваются знакам (или командам), которые не встречаются так часто, как присвоенные отдельным кодовым словам. В обмен на использование соответствующих случаю более длинных слов получаем более эффективное запоминание (меньшая клавиатура) или более эффективную передачу источника.

Коды сжатия данных часто имеют переменную длину. Интуитивно ясно, что длина двоичной последовательности, присвоенной каждому символу алфавита, должна обратно зависеть от вероятности этого символа. Из всего сказанного очевидно, что если символ появляется с высокой вероятностью, он содержит мало информации и ему не должен выделяться значительный ресурс системы. Аналогично не будет казаться неразумным, что когда все символы одинаково вероятны, код должен иметь фиксированную длину. Возможно, наиболее известным кодом переменной длины является код (или азбука) Морзе (Morse code). Самуэль Морзе, чтобы определить относительную частоту букв в нормальном тексте, вычислил количество букв в шрифтовой секции печатающего устройства. Кодовое присвоение переменной длины отражает эту относительную частоту.

Если имеется существенное различие в вероятностях символов, может быть получено значительное *сжатие данных*. Чтобы достичь этого сжатия, необходимо достаточно большое число символов. Иногда, чтобы иметь достаточно большое множество символов, образуется новое множество символов, определенное из исходного множества и называемое *кодом расширения*. Эта процедура уже рассматривалась в примере 13.3, а общая технология будет изучена в следующем разделе.

13.7.1. Свойства кодов

Ранее обращалось внимание на свойства, которым должен удовлетворять полезный код. Некоторые из этих свойств являются очевидными, а некоторые — нет. *Желаемые свойства* стоят того, чтобы их перечислить и продемонстрировать. Рассмотрим следующий трехсимвольный алфавит со следующими вероятностными соответствиями.

X_i	$P(X_i)$
a	0,73
b	0,25
c	0,02

Входному алфавиту сопутствуют следующие шесть двоичных кодовых соответствий, где крайний правый бит является наиболее ранним.

Символ	Код 1	Код 2	Код 3	Код 4	Код 5	Код 6
a	00	00	0	1	1	1
b	00	01	1	10	00	01
c	11	10	11	100	01	11

Изучите предлагаемые соответствия и попытайтесь определить, какие коды являются практичными.

Свойство единственности декодирования. Единственным образом декодируемые коды позволяют обратить отображение в исходный символьный алфавит. Очевидно, код 1 в предыдущем примере не является единственным образом декодируемым, так как символам a и b соответствует одна и та же двоичная последовательность. Таким образом, первым требованием полезности кода является то, чтобы каждому символу соответствовала уникальная двоичная последовательность. При этих условиях все другие коды оказываются удовлетворительными до тех пор, пока мы внимательно не изучим коды 3 и 6. Эти коды действительно имеют уникальную двоичную последовательность, соответствующую каждому символу. Проблема возникает при попытке закодировать последовательность символов. Например, попытайтесь декодировать двоичное множество 10111 при коде 3. Это b, a, b, b, b; b, a, b, c или b, a, c, b? Попытка декодировать ту же последовательность в коде 6 вызывает аналогичные сложности. Эти коды не являются единственным образом декодируемыми, даже если отдельные знаки имеют единственное кодовое соответствие.

Отсутствие префикса. Достаточным (но не необходимым) условием того, что код единственным образом декодируем, является то, что никакое кодовое слово не является префиксом любого другого кодового слова. Коды, которые удовлетворяют этому условию, называются кодами, свободными от префикса. Отметим, что код 4 не является свободным от префикса, но он единственным образом декодируем. Свободные от префикса коды также обладают таким свойством — они мгновенно декодируемы. Код 4 имеет свойство, которое может быть нежелательным. Он не является мгновенно декодируемым. Мгновенно декодируемый код — это такой код, для которого граница настоящего кодового слова может быть определена концом настоящего кодового слова, а не началом следующего кодового слова. Например, при передаче символа b с помощью двоичной последовательности 10 в коде 4, получатель не может определить, является ли это целым кодовым словом для символа b или частью кодового слова для символа c. В противоположность этому, коды 2 и 5 являются свободными от префикса.

13.7.1.1. Длина кода и энтропия источника

В начале главы были описаны формальные концепции информационного содержания и энтропии источника. Самоинформация символа X_n в битах была определена следующим образом: $I(X_n) = \log_2[1/P(X_n)]$. С точки зрения того, что информация разрешает неопределенность, было осознано, что информационное содержание символа стремится к нулю, когда вероятность этого символа стремится к единице. Кроме того,

была определена *энтропия* конечного дискретного источника как средняя информация этого источника. Поскольку информация разрешает неопределенность, энтропия является средним количеством неопределенности, разрешенной с использованием алфавита. Она также представляет собой среднее число бит на символ, которое требуется для описания источника. В этом смысле это также нижняя граница, которая может быть достигнута с помощью некоторых кодов сжатия данных, имеющих переменную длину. Действительный код может не достигать граничной энтропии входного алфавита, что объясняется множеством причин. Это включает неопределенность в вероятностном соответствии и ограничения буферизации. Средняя длина в битах, достигнутая данным кодом, обозначается как $\bar{n}$. Эта средняя длина вычисляется как сумма длин двоичных кодов, взвешенных вероятностью этих кодовых символов $P(X_i)$.

$$\bar{n} = \sum_i n_i P(X_i)$$

Когда говорится о поведении кода переменной длины, массу информации можно получить из знания *среднего числа бит*. В кодовом присвоении переменной длины некоторые символы будут иметь длины кодов, превосходящие среднюю длину, в то время как некоторые будут иметь длину кода, меньшую средней. Может случиться, что на кодер доставлена длинная последовательность символов с длинными кодовыми словами. Кратковременная скорость передачи битов, требуемая для передачи этих символов, будет превышать среднюю скорость передачи битов кода. Если канал ожидает данные со средней скоростью передачи, локальный избыток информации должен заноситься в буфер памяти. К тому же на кодер могут быть доставлены длинные модели символов с короткими кодовыми словами. Кратковременная скорость передачи битов, требуемая для передачи этих символов, станет меньше средней скорости кода. В этом случае канал будет ожидать биты, которых не должно быть. По этой причине для сглаживания локальных статистических вариаций, связанных с входным алфавитом, требуется *буферизация* данных.

Последнее предостережение состоит в том, что коды переменной длины создаются для работы со специальным множеством символов и вероятностей. Если данные, поступившие на кодер, имеют существенно отличающийся перечень вероятностей, буферы кодера могут быть не в состоянии поддержать несоответствие и будет происходить недогрузка или перегрузка буфера.

13.7.2. Код Хаффмана

Код Хаффмана (Huffman code) [20] — это свободный от префикса код, который может давать самую короткую среднюю длину кода $\bar{n}$ для данного входного алфавита. Самая короткая средняя длина кода для конкретного алфавита может быть значительно больше энтропии алфавита источника, и тогда эта невозможность выполнения обещанного сжатия данных будет связана с алфавитом, а не с методом кодирования. Часть алфавита может быть модифицирована для получения кода расширения, и тот же метод повторно применяется для достижения лучшего сжатия. Эффективность сжатия определяется *коэффициентом сжатия*. Эта мера равна отношению среднего числа бит на выборку до сжатия к среднему числу бит на выборку после сжатия.

Процедура кодирования Хаффмана может применяться для преобразования между любыми двумя алфавитами. Ниже будет продемонстрировано применение процедуры при произвольном входном алфавите и двоичном выходном алфавите.

Код Хаффмана генерируется как часть процесса образования дерева. Процесс начинается с перечисления входных символов алфавита наряду с их вероятностями (или относительными частотами) в порядке убывания частоты появления. Эти позиции таблицы соответствуют концам ветвей дерева, как изображено на рис. 13.34. Каждой ветви присваивается ее весовой коэффициент, равный вероятности этой ветви. Теперь процесс образует дерево, поддерживающее эти ветви. Два входа с самой низкой относительной частотой объединяются (на вершине ветви), чтобы образовать новую ветвь с их смешанной вероятностью. После каждого объединения новая ветвь и оставшиеся ветви переупорядочиваются (если необходимо), чтобы убедиться, что сокращенная таблица сохраняет убывающую вероятность появления. Это переупорядочение называется *методом пузырька* [21]. Во время переупорядочения после каждого объединения поднимается ("всплывает") новая ветвь в таблице до тех пор, пока она не сможет больше увеличиваться. Таким образом, если образуется ветвь с весовым коэффициентом 0,2 и во время процесса находятся две другие ветви уже с весовым коэффициентом 0,2, новая ветвь поднимается до вершины группы с весовым коэффициентом 0,2, а не просто присоединяется к ней. Процесс "всплытия" пузырьков к вершине группы дает код с уменьшенной дисперсией длины кода, в противном случае — код с такой же средней длиной, как та, которая получена посредством простого присоединения к группе. Эта сниженная дисперсия длины кода уменьшает шанс переполнения буфера.

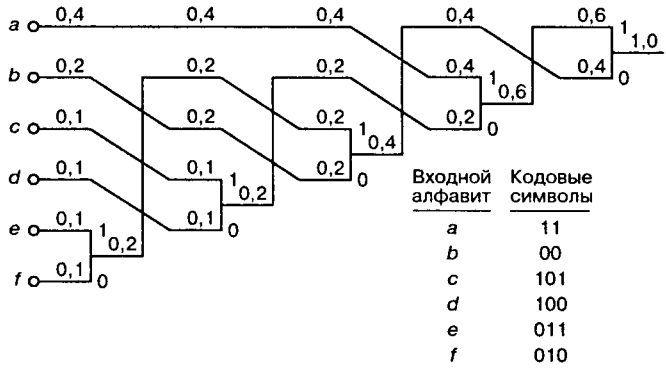


Рис. 13.34. Дерево кодирования Хаффмана для шестизначного множества

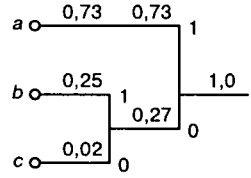
В качестве примера этой части процесса кодирования применим процедуру Хаффмана к входному алфавиту, изображенному на рис. 13.34. Протабулированный алфавит и связанные с ним вероятности изображены на рисунке. После формирования дерева, чтобы различать две ветви, каждая вершина ветви снабжается двоичным решением "1/0". Присвоение является произвольным, но для определенности на каждой вершине будем обозначать ветвь, идущую вверх как "1", и ветвь, идущую вниз как "0". После обозначения вершин ветвей проследим траектории дерева от основания (крайнее правое положение) до каждой выходной ветви (крайнее левое положение). Траектория — это двоичная последовательность для достижения этой ветви. В следующей таблице для каждого конца ветви указана последовательность, соответствующая каждой траектории, где $i = 1, \ldots, 6$.

X_i	$P(X_i)$	Код	n_i	$n_iP(X_i)$
a	0,4	11	2	0,8
b	0,2	00	2	0,4
c	0,1	101	3	0,3
d	0,1	100	3	0,3
e	0,1	011	3	0,3
f	0,1	010	3	0,3
				$\overline{n} = 2,4$

Находим, что средняя длина кода $\overline{n}$ для этого алфавита равна 2,4 бит на знак. Это не означает, что необходимо найти способ передачи нецелого числа бит. Это означает, что для передачи 100 входных символов через канал связи в среднем должно пройти 240 бит. Для сравнения, код фиксированной длины, требуемый для охвата шестизначного входного алфавита, должен иметь длину 3 бит и энтропию входного алфавита (используем формулу (13.32)), равную 2,32 бит. Таким образом, этот код дает коэффициент сжатия 1,25 (3,0/2,4) и достигает 96,7% (2,32/2,40) возможного коэффициента сжатия. В качестве еще одного примера рассмотрим случай, для которого можно продемонстрировать использование кода расширения. Изучим трехзначный алфавит, представленный в разделе 13.6.1.

X_i	$P(X_i)$
a	0,73
b	0,25
c	0,02

Дерево кода Хаффмана для этого алфавита изображено на рис. 13.35, а его элементы протабулированы ниже.



Входной алфавит	Кодовые символы
a	1
b	01
c	00

Рис. 13.35. Дерево кодирования Хаффмана для трехзначного множества

X_i	$P(X_i)$	Код	n_i	$n_iP(X_i)$
a	0,73	1	1	0,73
b	0,27	01	2	0,54
c	0,02	00	2	0,04
				$\overline{n} = 1,31$

Здесь $i = 1, 2, 3$. Средняя длина кода в приведенном примере равна 1,31 бит; она будет равна 2 бит для кода Хаффмана фиксированной длины. Коэффициент сжатия для этого кода равен 1,53. И снова, используя равенство (13.32), получим, что энтропия для алфавита равна 0,9443 бит, так что эффективность кода (0,944/1,31 = 72%) значительно меньше, чем для предыдущего примера.

Чтобы улучшить эффективность кодирования или достичь большего сжатия, следует переопределить алфавит источника. Больший алфавит источника увеличивает разнообразие, что является одним из требований при сокращении средней длины кода и увеличении числа ветвей дерева для присвоения кода переменной длины. Это делается посредством выбора знаков (по два) из алфавита источника, которые становятся новыми знаками в расширенном алфавите. Если предположить, что символы независимы, вероятность каждого нового элемента является произведением отдельных вероятностей. Алфавит расширения имеет следующий вид.

X_i	$P(X_i)$	Код	n_i	$n_iP(X_i)$
aa	0,5329	1	1	0,5329
ab	0,1825	00	2	0,3650
ba	0,1825	011	3	0,5475
bb	0,0625	0101	4	0,2500
ac	0,0146	01000	5	0,0730
ca	0,0146	010011	6	0,0876
bc	0,0050	0100100	7	0,0350
cb	0,0050	01001011	8	0,0400
cc	0,0002	01001010	8	0,0016
				$\bar{n}$ = 1,9326 бит/два символа
				= 0,9663 бит/символ

Здесь $i = 1, ..., 9$, а кодовая последовательность для каждого X_i была найдена с использованием выше приведенной процедуры Хаффмана. Коэффициент сжатия для этого кода расширения равен 2,076, а эффективность кодирования — 97,7%. Коды расширения предлагают очень мощную технологию включения эффектов множеств символов, которые не являются независимыми. Например, в английском алфавите соседние буквы являются высоко коррелированными. Очень часто встречаются следующие пары букв.

th	re	in
sh	he	e_
de	ed	s_
ng	at	r_
te	es	d_

Здесь подчеркивание представляет пробел. Наиболее общими английскими наборами трех букв являются следующие.

the	and	for
ing	ion	ess

Таким образом, вместо того чтобы производить кодирование Хаффмана отдельных букв, более эффективно расширить алфавит, включив все 1-кортежи плюс распространенные 2- и 3-кортежи, а затем произвести кодирование с помощью кода расширения.

13.7.3. Групповые коды

Во многих приложениях последовательность символов, которую необходимо передать или запомнить, характеризует последовательное кодирование определенных символов. Иногда, вместо того чтобы кодировать каждый символ последовательности, есть смысл описать группу с помощью подстановочного кода. В качестве примера рассмотрим случай, когда последовательности пробелов (наиболее употребимый символ в

тексте) кодируются во многих протоколах связи с помощью символа управления, за которым следует счетчик символов. Протокол IBM 3780 BISYNC имеет опцию замены последовательности пробелов с помощью знака "IGS" (если имеем дело с EBCDIC) или "GS" (если имеем дело с ASCII), за которым следует счетчик от 2 до 63. Более длинные последовательности делятся на серии по 63 знака.

Групповое подстановочное кодирование может быть применено к исходному алфавиту символов или двоичному представлению этого алфавита. В частности, групповое кодирование является удачным для двоичных алфавитов, полученных от специфических источников. Наиболее важным коммерческим примером является факсимильное кодирование, используемое для передачи документов по мгновенной электронной почте [22].

13.7.3.1. Кодирование Хаффмана для факсимильной передачи

Факсимильная передача — это процесс передачи двухмерного образа как последовательности последовательных строчных разверток. В действительности наиболее распространенными образами являются документы, содержащие текст и цифры. Положение строчной развертки и положение вдоль развертки квантуются в пространственные расположения, которые определяют двухмерную координатную сетку элементов картинки, называемых пикселями. Ширина стандартного документа МККТТ определяется равной 8,27 дюймов (20,7 см), а длина — 11,7 дюймов (29,2 см), почти 8,5 дюймов на 11,0 дюймов. Пространственное квантование для нормального разрешения составляет 1728 пикселей/строку и 1188 строк/документ. Стандарт также определяет квантование с высоким разрешением с теми же 1728 пикселями/строку, но с 2376 строками/документ. Общее число отдельных пикселей для факсимильной передачи с нормальным разрешением составляет 2 052 864, и оно удваивается для высокого разрешения. Для сравнения, число пикселей в стандарте NTSC (National Television Standards Committee — Национальный комитет по телевизионным стандартам) коммерческого телевидения составляет 480 × 460, или 307 200. Таким образом, факсимильное изображение имеет разрешение в 6,7 или 13,4 раза больше разрешения стандартного телевизионного образа.

Относительная яркость или затемненность развернутого образа в каждом положении на строке развертки квантуется в два уровня: Ч (черный) и Б (белый). Таким образом, сигнал, наблюдаемый на протяжении линии развертки, — это двухуровневая модель, представляющая элементы Ч и Б. Легко видеть, что горизонтальная развертка данной страницы будет представлять последовательность, состоящую из длинных групп уровней Ч и Б. Стандарт МККТТ схемы группового кодирования для сжатия отрезков Ч и Б уровней базируется на модифицированном коде Хаффмана переменной длины, который приведен в табл. 13.1. Определяются два типа шаблонов, группы Б и Ч. Каждый отрезок описывается *кодовыми словами деления*. Первое деление, названное *созданное кодовое слово*, определяет группы с длинами, кратными 64. Второе деление, именуемое *оконечное кодовое слово*, определяет длину оставшейся группы. Каждая серия из знаков Ч (или Б), длиной от 0 до 63, обозначает единственное кодовое слово Хаффмана, как и каждая группа длины $64 \times K$, где $K = 1, 2, \ldots, 27$. Также кодом определен уникальный символ конца строки (end of line — EOL), который указывает, что дальше не следуют черные пиксели. Следовательно, должна начаться следующая строка, что подобно возврату каретки пишущей машинки [23].

Таблица 13.1. Модифицированный код Хаффмана для стандарта факсимильной связи МККТТ

Длина группы	Белые	Черные	Длина группы	Белые	Черные
		Созданные кодовые слова			
64	11011	0000001111	960	011010100	0000001110011
128	10010	000011001000	1024	011010101	0000001110100
192	010111	000011001001	1088	011010110	0000001110101
256	0110111	000001011011	1152	011010111	0000001110110
320	00110110	000000110011	1216	011011000	0000001110111
384	00110111	000000110100	1280	011011001	0000001010010
448	01100100	000000110101	1344	011011010	0000001010011
512	01100101	0000001101100	1408	011011011	0000001010100
576	01101000	0000001101101	1472	010011000	0000001010101
640	01100111	0000001001010	1536	010011001	0000001011010
704	011001100	0000001001011	1600	010011010	0000001011011
768	011001101	0000001001100	1664	011000	0000001100100
832	011010010	0000001001101	1728	010011011	0000001100101
896	011010011	0000001110010	EOL	000000000001	000000000001
Длина группы	Белые	Черные	Длина группы	Белые	Черные
		Оконечные кодовые слова			
0	00110101	000110111	32	00011011	000001101010
1	000111	010	33	00010010	000001101011
2	0111	11	34	00010011	000011010010
3	1000	10	35	00010100	000011010011
4	1011	011	36	00010101	000011010100
5	1100	0011	37	00010110	000011010101
6	1110	0010	38	00010111	000011010110
7	1111	00011	39	00101000	000011010111
8	10011	000101	40	00101001	000001101100
9	10100	000100	41	00101010	000001101101
10	00111	0000100	42	00101011	000011011010
11	01000	0000101	43	00101100	000011011011
12	001000	0000111	44	00101101	000001010100
13	000011	00000100	45	00000100	000001010101
14	110100	00000111	46	00000101	000001010110
15	110101	000011000	47	00001010	000001010111
16	101010	0000010111	48	00001011	000001100100
17	101011	0000011000	49	01010010	000001100101
18	0100111	0000001000	50	01010011	000001010010

Окончание табл. 13.1

Длина группы	Белые	Черные	Длина группы	Белые	Черные
		Оконечные кодовые слова			
19	0001100	00001100111	51	01010100	000001010011
20	0001000	00001101000	52	01010101	000001000100
21	0010111	00001101100	53	00100100	000000110111
22	0000011	00000110111	54	00100101	000000111000
23	0000100	00000101000	55	01011000	000000100111
24	0101000	00000010111	56	01011001	000000101000
25	0101011	00000011000	57	01011010	000001011000
26	0010011	000011001010	58	01011011	000001011001
27	0100100	000011001011	59	01001010	000000101011
28	0011000	000011001100	60	01001011	000000101100
29	00000010	000011001101	61	00110010	000001011010
30	00000011	000001101000	62	00110011	000001100110
31	00011010	000001101001	63	00110100	000001100111

Пример 13.8. Код группового кодирования

Используйте модифицированный код Хаффмана для сжатия строки

200 Б, 10 Ч, 10 Б, 84 Ч, 1424 Б,

состоящей из 1728 пиксельных элементов.

Решение

Используя табл. 13.1, определим, каким должно быть кодирование для этой модели (для удобства чтения введены пробелы).

010111	10011	0000100	00111	0000001111	00001101000	000000000001
192Б	8Б	10Ч	10Б	64Ч	20Ч	EOL

Итак, требуется всего 56 бит, чтобы послать эту строку, содержащую последовательность 1188 бит.

13.7.3.2. Коды Лемпеля-Зива (ZIP)

Основной сложностью при использовании кода Хаффмана является то, что вероятности символов должны быть известны или оценены и как кодер, так и декодер должны знать дерево кодирования. Если дерево строится из необычного для кодера алфавита, канал, связывающий кодер и декодер, должен также отправлять кодирующее дерево как заголовок сжатого файла. Эти служебные издержки уменьшат эффективность сжатия, реализованную с помощью построения и применения дерева к алфавиту источника. Алгоритм Лемпеля-Зива (Lempel-Ziv) и его многочисленные разновидности используют текст сам по себе для итеративного построения синтаксически выделенной последовательности кодовых слов переменной длины, которые образуют кодовый словарь.

Код предполагает, что существующий словарь содержит уже закодированные сегменты последовательности символов алфавита. Данные кодируются с помощью просмотра существующего словаря для согласования со следующим коротким сегментом кодируемой последовательности. Если согласование найдено, код следует такой фило-

софии: поскольку получатель уже имеет этот сегмент кода в своей памяти, нет необходимости пересылать его, требуется только определить адрес, чтобы найти сегмент. Код ссылается на расположение последовательности сегмента и затем дополняет следующий символ в последовательности, чтобы образовать новую позицию в словаре кода. Код начинается с пустого словаря, так что первые элементы являются позициями, которые не ссылаются на более ранние. В одной форме словаря рекуррентно формируется выполняемая последовательность адресов и сегмент символов алфавита, содержащийся в ней. Закодированные данные состоят из пакета <адрес словаря, следующий знак данных>, а каждый новый входной элемент словаря образован как пакет, содержащий адрес того словаря, за которым следует следующий символ. Рассмотрим пример такой технологии кодирования.

Закодируйте последовательность символов [a b a a b a b b b b b b a b b b b a]

Закодированные пакеты:	<0,a>,	<0,b>,	<1,a>,	<2,a>,	<2,b>,	<5,b>,	<5,a>,	<6,b>, <4,–>
Адрес:	1	2	3	4	5	6	7	8
Содержимое:	a	b	aa	ba	bb	bbb	bba	bbbb

Начальный пакет <0,a> показывает нулевой адрес, потому что в словаре еще нет ни одной позиции. В этом пакете знак “a” является первым в последовательности данных, и он приписан к адресу 1. Следующий пакет <0,b> содержит второй знак данных b, который еще не был в словаре (следовательно, адресное значение есть 0); b приписывается адресу 2. Пакет <1,a> представляет кодирование следующих двух знаков “aa” с помощью вызова адреса 1 для первого и присоединения к этому адресу следующего знака “a”. Пара знаков “aa” приписывается адресу 3. Пакет <2,a> представляет кодирование следующих двух знаков данных “ba” с помощью вызова адреса 2 для знака “b” и присоединения к этому адресу следующего знака “a”. Пара знаков данных “ba” приписывается адресу 4 и т.д. Отметим, как завершается групповое кодирование. Восьмой пакет составлен из адреса 6, содержащего три знака “b”, за которыми следует другой знак “b”. В этом примере закодированные данные могут быть описаны с помощью трехбитового адреса с последующим битом 0 или 1 для определения присоединенного знака. В закодированной последовательности существует последовательность из 9 символов для общего содержимого в 36 бит для кодирования данных, содержащих 20 знаков. Как во многих схемах сжатия, эффективность кодирования не достигается для коротких последовательностей, как в этом примере, и имеется только для длинных последовательностей.

В другой форме алгоритма Лемпеля-Зива закодированные данные представлены как три словесных пакета вида <число знаков сзади, длина, следующий знак>. Здесь концепция адреса не используется. Наоборот, имеются ссылки на предшествующие последовательности данных, а также допускаются рекуррентные ссылки на параметр длины. Это показано в следующем примере, представленном как позиция <1,7,a>.

Закодируйте последовательность символов [a b a a b a b b b b b b a b b b b a]

Закодированные пакеты:	<0,0,a>,	<0,0,b>,	<2,1,a>,	<3,2,b>,	<1,7,a>,	<6,5,a>
Содержимое:	a	b	aa	bab	bbbbbbba	bbbbba
Текущий текст:	a	ab	abaa	abaabab	abaabbbbbbbbba	вся последовательность

Здесь также не видно эффективности кодирования для короткой серии данных. Разновидности кода ограничивают размер обратной ссылки, например 12-битовая для максимума в 4 096 пунктов обратной ссылки. Это ограничение уменьшает размер памяти, требуемой для словаря, и сокращает вероятность перегрузки памяти. Возможны также модификации кода, ограничивающие длину префикса или фразы, определенной первыми двумя аргументами <назад $n1$, вперед $n2$, xxx>, которые должны быть меньше некоторого значения (например, 16) с целью ограничения сложности обратного поиска во время кодирования. Алгоритм Лемпеля-Зива присутствует во многих коммерческих и пробных программах, которые включают *сжатие* LZ77, Gzip, LZ78, LZW и UNIX.

13.8. Примеры кодирования источника

Кодирование источника стало основной подсистемой в современных системах связи. Высокие требования к полосе частот и возможность запоминания явились мотивом его развития, в то время как интегрированные схемы и методы обработки сигналов предоставили такую возможность. Вторичной причиной широкого внедрения процесса в систему связи является определение общеиндустриальных *стандартов*, которые позволяют множественным поставщикам проводить рентабельную и конкурентоспособную реализацию процесса кодирования. Существуют стандарты МККТТ для кодирования источника или алгоритмов сжатия речи, аудио, неподвижных образов и движущихся изображений. В этом разделе будет изучено множество алгоритмов кодирования источника, основанных на стандартах, что должно продемонстрировать широкую применимость кодирования источника в системах связи и проиллюстрировать типичные уровни производительности.

13.8.1. Аудиосжатие

Аудиосжатие широко применяется в потребительских и профессиональных цифровых аудиопродуктах, таких как компакт-диски (compact disc — CD), цифровая аудиолента (digital audio type — DAT), мини-диск (mini-disk — MD), цифровая компакт-кассета (digital compact cassette — DCC), универсальный цифровой диск (digital versatile disc — DVD), цифровое аудиовещание (digital audio broadcasting — DAB) и аудиопродукция в формате MP3 от экспертной группы по вопросам движущегося изображения (Motion Picture Experts Group — (MPEG). К тому же сжатие речи в телефонии, в частности сотовой телефонии, требуемое для экономии полосы частот и сбережения времени жизни батареи, дало начало процессу разработки множества стандартов сжатия речи. Различные алгоритмы применимы к речевым и потребительским сигналам более широкой полосы частот. Аудио- и речевые схемы сжатия можно для удобства разделить согласно приложениям, что отражает некоторую меру приемлемого качества. Рассмотрим параметры, описывающие это деление [24, 25].

Типичные значения параметров для трех классов аудиосигналов

	Диапазон частот	Частота дискретизации	Бит PCM/выборку	Скорость передачи битов PCM
Телефонная речь	300–3 400 Гц	8 кГц	8	64 Кбит/с
Широкополосная речь	60–7 000 Гц	16 кГц	14	224 Кбит/с
Широкополосное аудио	10–20 000 Гц	48 кГц	16	768 Кбит/с

13.8.1.1. Адаптивная дифференциальная импульсно-кодовая модуляция

Начнем наше обсуждение с обработки телефонной речи. Один из стандартов этой области — адаптивная дифференциальная импульсно-кодовая модуляция (adaptive differential pulse-code modulation — ADPCM) G.726 от МККТТ. Этот стандарт кодирует выборку за выборкой, предсказывая значение каждой выборки из восстановленной речи предшествующих выборок, с использованием адаптивного предсказателя с обратной связью. Он принимает качественную речь, преобразованную посредством 8-битового линейного преобразования с использованием *A*- или μ-закона со скоростью 64 Кбит/с, и выдает сжатую речь со скоростью 16, 24, 32 и 40 Кбит/с. Кодер применяет декодер в контуре обратной связи для анализа и модификации параметров алгоритма с целью минимизации ошибки восстановления. Предсказатель использует фильтр шестого порядка для моделирования нулей и фильтр второго порядка — для моделирования полюсов источника входного сигнала. Блочная диаграмма кодера изображена на рис. 13.36.

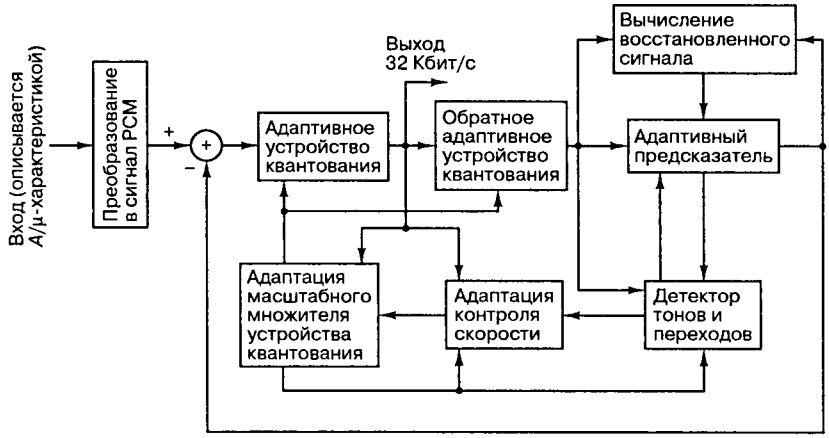


Рис. 13.36. Речевой кодек ADPCM (G.726)

13.8.1.2. Адаптивная дифференциальная импульсно-кодовая модуляция с разделением на подполосы

Стандарт МККТТ G.722 является стандартом кодирования широкополосной речи. Широкополосное сжатие приводит к значительному улучшению качества телефонной речи, которое приближается к качеству речи при радиовещании и в музыкальных сигналах. Данный кодер использует дополнительные фильтры нижних и верхних частот для отделения входной полосы частот в 7 кГц, после чего речь дискретизуется с частотой 16 кГц в более высокую и более низкую подполосы, каждая из которых выбирается с частотой 8 кГц. Функции обоих фильтров и операция повторной дискретизации реализованы в цифровом фильтре, известном как *квадратурный зеркальный фильтр* (quadrature mirror filter). Независимые кодеры ADPCM обрабатывают временные ряды сокращенных полос частот от двух фильтров и выдают скорости в 48 Кбит/с и 16 Кбит/с, соответственно, на выходе низкой и высокой полос. Эти кодеры представляют собой модифицированную версию речевых кодеров ADPCM МККТТ В.721, которые используют фильтры с обратным предсказанием, основанные на закодированном разностном сигнале. Отбрасывание младшего бита коэффициентов предсказывающего фильтра позволяет этому кодеру работать со скоростью 56 и 48 Кбит/с,

как и с номинальной скоростью 64 Кбит/с. При сниженной скорости передачи битов система связи может присваивать неиспользованные биты вспомогательному потоку данных, который передается со скоростью 8 и 16 Кбит/с, если канал поддерживает фиксированную выходную скорость в 64 Кбит/с. Предсказатель использует структуру с 6 нулями и 2 полюсами. Блочная диаграмма широкополосного аудиокодера, работающего со скоростью 64 Кбит/с, изображена на рис. 13.37.

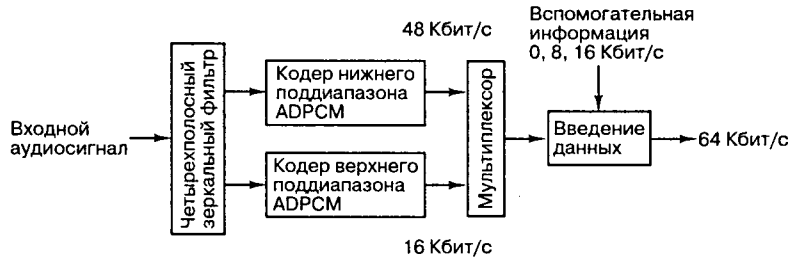


Рис. 13.37. Широкополосный кодек QMF-ADPCM (64 Кбит/с) (G.722)

13.8.1.3. Схема CELP

Речевые кодеры, использующие линейные фильтры с предсказанием (linear predictive filter — LPF), могут давать высокое качество речи, закодированной со скоростью выше 16 Кбит/с, однако при снижении скорости качество быстро падает. Кодеры LPC могут быть модифицированы с целью получения высококачественного сжатия речи со скоростями порядка от 4,8 до 9,6 Кбит/с посредством приведения задачи синтеза к двухэтапной процедуре, названной *синтез через анализ* (synthesis by analysis). На первом этапе образуется модель LPC 10-го порядка для сигнала, действительного на протяжении короткого интервала, скажем каждые 20 мс. На втором этапе находится волновой сигнал, который, будучи примененным к модели LPC, образует выходной сигнал, по возможности близкий к исходному синтезируемому сигналу. Завершается эта задача с помощью последовательного применения подходящего сигнала активизации к модели и сравнения каждой синтезированной формы сигнала с исходным сигналом с последующим выбором того, который минимизирует ошибку между исходным сигналом и выходом управляемой модели.

Из теории процесса формирования речи известно, что активизация речи часто состоит из периодических импульсов (образованных посредством вибрации речевых связок). Период периодических импульсов P связан с голосом говорящего. Одноотводный рекурсивный фильтр определяется двумя параметрами: P — число интервалов запаздывания в контуре обратной связи и g — коэффициент обратной связи. Импульсная характеристика этого фильтра представляет собой затухающую последовательность с P равными нулю выходными выборками между последовательными ненулевыми выходными выборками. Выход этого фильтра генерирует периодический сигнал активизации, подаваемый на вход модели LPC (см. раздел 13.3.2). Алгоритм синтеза должен проверять возможные значения P из перечня подходящих. Два параметра голоса оцениваются каждые 5 мс. Вход в речевой фильтр извлекается из таблицы подходящих последовательностей активизации. Выход фильтра, в свою очередь, управляет моделью LPC. Таблица, содержащая, как правило, 1 024 позиции, называется кодовой книгой. Кодовая книга посещается каждые 2,5 мс. Когда наилучшая комбинация позиций кодовой книги и период голоса определены с помощью полного

поиска, формируется группа, содержащая последовательность параметров голоса, последовательность адресов кодовой книги и информацию о коэффициентах LPC.

Кодер должен доставить параметры, описывающие модель LPC, на декодер. Спектральная характеристика фильтра LPC очень чувствительна к квантованию коэффициентов и как таковая должна бы представляться с помощью неприемлемо большого числа бит. Поэтому коэффициенты LPC преобразуются в иное множество параметров, названных *линейными спектральными парами* [10], которые являются нечувствительными к квантованию.

Системы, созданные согласно стандарту IS-95, используют следующий формат кадра LPC. Кадр, требуемый для описания 2 мс данных, содержит 192 бит, присвоенных представителю закодированных параметров.

10 коэффициентов LPC	40 бит
4 параметра запаздывания и опережения	40 бит
8 адресов кодовой книги	80 бит
Биты четности, проверочные биты и прочая служебная информация	32 бит

Общая скорость передачи битов для этой системы составляет 192 бит за 20 мс, или 9600 бит/с. Скорость передачи может быть снижена, если кодер обнаруживает речевые паузы.

13.8.1.4. Уровни I, II и III стандарта MPEG

Международная организация по стандартизации (International Organization for Standardization —ISO) и экспертная группа по вопросам движущегося изображения (Motion Picture Experts Group — MPEG) разработали стандарт аудиосжатия для сигнала, синхронизированного с сжатым видеосигналом, известный как MPEG. В этой схеме объединены свойства MUSICAM (Masking pattern adaptive Universal Subband Integrated Coding And Multiplexing — универсальные интегральные средства кодирования и уплотнения по поддиапазонам с маскировкой и адаптацией к кодограмме) и ASPEC (Adaptive Spectral Perceptual Entropy Coding — адаптивное спектрально-восприимчивое кодирование энтропии). В схеме использованы три уровня (коды) увеличивающейся сложности и улучшающейся субъективной производительности, входные частоты дискретизации равны 32, 44,1 и 48 кГц, а биты на выход подаются со скоростью от 32 до 192 Кбит/с (монофонический канал) или со скоростью от 64 до 384 Кбит/с (стереофонический канал). Стандарт поддерживает режим работы единственного канала, стереорежим, двойственный режим работы канала (для двуязычных аудиопрограмм) и дополнительный совместный стереорежим. В последнем режиме два кодера для левого и правого каналов могут поддерживать друг друга, используя общие статистики с целью снижения скорости передачи бит аудиосигнала, даже большего, чем это возможно при монофонической передаче [26].

Кодер действует в соответствии с моделью реального времени порога *спектральной восприимчивости человека*. Этот порог представляет собой зависящую от частоты границу или порог, который отмечает уровни звукового давления, ниже которых человеческое ухо не может воспринимать сигналы. Эта кривая, названная *порогом остроты слуха*, генерируется во время слухового теста. Порог остроты обычно присутствует на уровнях амплитуды как функция спектрального положения и во многом подобен кривой спектра мощности. Этот порог представляет собой изменяющуюся во времени функцию кратковременной спектральной плотности мощности и имеет локальные максимумы в соответствии с тонами высокого уровня и тонообразными сигналами (называемыми *тонала-*

ми). Повышение порога вследствие наличия сильных тоналов, приводит к локальной маскировке спектральных компонентов ниже нового порогового уровня. Спектральные компоненты сигнала, лежащие ниже порога слышимости, объявляются несущественными и не кодируются в процессе сжатия. Сигналы, превышающие зависящий от частоты порог, кодируются с достаточной точностью, позволяющей удерживать ошибку аппроксимации ниже уровня остроты. Этот процесс завершается делением спектра множеством узкополосных фильтров и присвоением достаточного числа бит для описания каждого выхода фильтра относительно его амплитуды, которая расположена выше порога. Таким образом, сигналу, в определенной полосе составляющему 30 дБ выше порога, будет при квантовании выделено 5 бит. В этом случае шум квантования падает ниже порога, так как отношение шум/сигнал квантования сократилось на 6 дБ на бит. Типичный график порога остроты представлен на рис. 13.38.

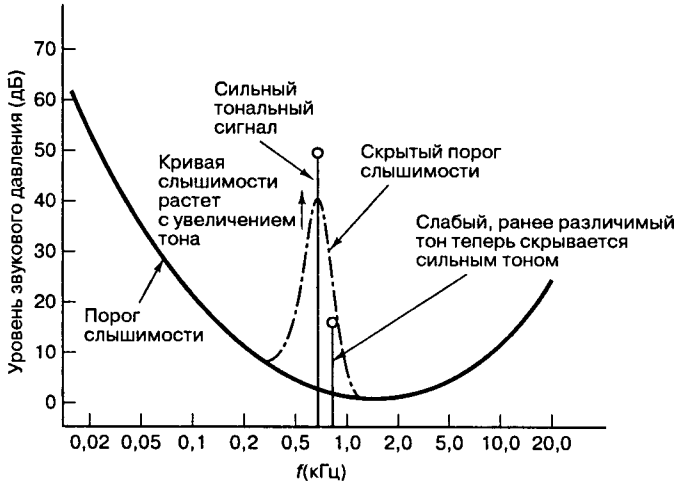


Рис. 13.38. Порог остроты и маскировка

Кодер работает следующим образом. Стандартный 16-битовый аудиосигнал PCM усекается и преобразуется в компоненты спектральной подполосы с помощью группы многофазных фильтров, состоящей из 32 равномерно расположенных полосовых фильтров. Блок фильтра создается с помехами соседнего канала, превосходящими 96 дБ, — уровень, требуемый для подавления искажения восприимчивости, вызванного шумом квантования. Фильтрованные выходные сигналы выбираются с частотой Найквиста для каждой полосы пропускания диапазона частот. В декодере этот процесс обращается. Частота дискретизации каждого многополосного фильтра увеличивается до частоты исходного сигнала источника с помощью интерполирования сигналов подполосы, образованных на выходах полосы пропускания блока синтетических фильтров. На рис. 13.39 представлена блочная диаграмма аудиокодера и декодера уровней I и II стандарта MPEG.

На уровне III стандарта MPEG/ISO (MP3) достигается разрешение более высокой частоты, которое весьма точно соответствует критической разрешающей способности человека. Это усовершенствованное деление достигается посредством дальнейшей обработки 32 подполосных сигналов с помощью перекрывающегося или усеченного 6-точечного или 18-точечного модифицированного дискретного косинус-преобразования (modified discrete cosine transform — MDCT). (Короткое описание ДКП представлено в следующем разделе, посвященном сжатию изображений.) Результирующее число полос частот, которое может

быть разрешено на уровне III, равно 32 × 18, или 576, где каждый фильтр представляет полосу частот в 24 000/576 или 41,67 Гц. Уровень III отличается от уровней I и II дополнительным введением модифицированного ДКП в блок анализа, кодера Хаффмана на выход квантующего устройства и канала побочной информации.

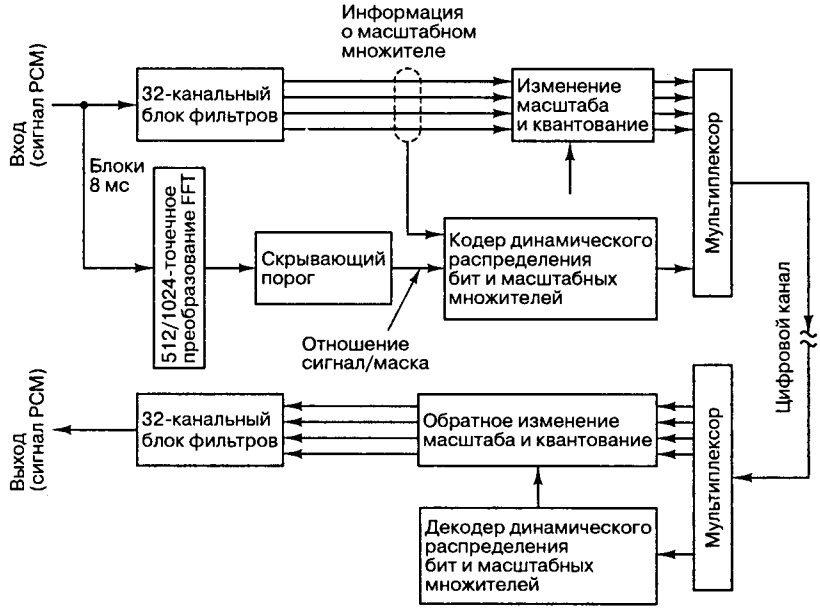


Рис. 13.39. Блочная диаграмма аудиокодера и декодера, уровни I и II

13.8.2. Сжатие изображения

Мы часто слышали старое высказывание: *Картина стоит тысячи слов*[1]. Верно ли оно? 1 000 слов содержит 6 000 знаков, которые, будучи закодированы как 7-битовые символы ASCII, требуют в общей сложности 42 000 бит. Какого размера образ (или картина) может быть описан с помощью 42 000 бит? Если используется монохромный (т.е. черный и белый) образ со стандартной 8-битовой шкалой оттенков серого, образ будет ограничен 5 250 пикселями (или элементами изображения). Этот образ может иметь размерность 70 × 75 пикселей, и если предположить, что образ среднего качества (разрешение 300 пикселей на дюйм), в результате получаем, что наш образ составляет примерно $\frac{1}{4}$ дюйма на $\frac{1}{4}$ дюйма. Определенно, требуется какое-то кодирование изображения.

Подойдем к проблеме с другой стороны. Насколько большим является изображение? Выбирая лист бумаги размером 8,5 × 11,0 дюймов, содержащий изображение с разрешением 300 пикселей на дюйм, получаем образ, содержащий 8,5 × 300 × 11,0 × 300 или 8,4 × 10^6 элементов изображения. Если это полноцветная картина с тремя цветами на элемент, каждый из которых описывается с помощью 8-битовых слов, находим, что образ содержит 2 × 10^8 бит, что эквивалентно 4,8 × 10^6 6-знаковых слов ASCII. Возможно, старое высказывание стоит обновить в соответствии с совре-

[1] Более привычным является все же выражение "лучше один раз увидеть, чем сто раз услышать", но в целях дальнейшего обсуждения приведен дословный перевод. — *Примеч. пер.*

менным положением дел, сказав, что: *Картина стоит порядка пяти миллионов слов.* Для сравнения с другими форматами изображения отметим, что отдельный кадр телевизионного изображения высокой четкости содержит примерно $1{,}8 \times 10^6$ пикселей, стандартное телевизионное изображение — это примерно $0{,}33 \times 10^6$ пикселей, а мониторы компьютера высшего класса содержат от 1,2 до $3{,}1 \times 10^6$ элементов изображения.

Технология дала нам принтеры низкой стоимости с высокой разрешающей способностью, сканеры, камеры и мониторы, позволяющие схватывать и представлять изображения с коммерческой и развлекательной целью. Хранение и передача этих образов существенно зависит от кодирования источника, призванного снизить требования к полосе частот и памяти. Существует множество стандартов, которые были разработаны для сжатия изображений. В следующем разделе будут изучены элементы двух основных схем сжатия [26, 27].

13.8.2.1. JPEG

JPEG (Joint Photography Experts Group — объединенная группа экспертов в области фотографии) — это общее название, которое дано стандарту ISO/JPEG 10918-1 и стандарту ITU-T Recommendation T.81 "Цифровое сжатие постоянных изображений непрерывного тона". JPEG, в основном, известен как основанная на преобразовании схема сжатия с потерями. Сжатие с потерями допускает ошибки в построении сигнала. Уровни ошибок должны быть ниже порога восприимчивости человеческого глаза. JPEG поддерживает три режима работы, связанных с дискретным косинус-преобразованием (discrete cosine transform — DCT, ДКП): последовательное ДКП, прогрессивное ДКП и иерархическое, а также режим без потерь с использованием дифференциального предсказания и энтропии кодирования ошибки предсказания. ДКП — это численное преобразование, связанное с дискретным преобразованием Фурье (discrete Fourier transform — DFT, ДПФ) и предназначенное для получения спектрального разложения четносимметричных последовательностей. Если входная последовательность является четносимметричной, нет необходимости в синусоидальных компонентах преобразования. Следовательно, ДКП может заменить ДПФ.

Начнем с введения двухмерного преобразования ДКП 8×8. Сначала прокомментируем использование ДКП для образования спектрального описания блока 8×8 пикселей. Двухмерное ДКП — это сепарабельное преобразование, которое может быть записано в виде двойной суммы по двум размерностям. Сепарабельное ДКП производит восемь 8-точечных ДКП в каждом направлении. Следовательно, основной компоновочный блок представляет собой единичное 8-точечное ДКП. Возникает вопрос, почему используется ДКП, а не какое-либо другое преобразование, например ДПФ. Ответ связан с теоремой о дискретном представлении и преобразованием Фурье. Преобразование в одной области приводит к периодичности в другой. Если преобразуется временной ряд, его спектр становится периодичным. С другой стороны, если преобразуется спектр временного ряда, временной ряд периодически продолжается. Этот процесс известен как *периодическое расширение* и обозначается результирующей *периодограммой.* Периодическое расширение исходных данных (рис. 13.40) демонстрирует разрыв на границах, который ограничивает степень спектрального затухания в спектре величиной $1/f$. Можно образовать четное расширение данных, отображая данные относительно одной из границ. Если данные являются периодически расширенными, как показано на рис. 13.40, разрывность уже свойственна не амплитуде данных, а ее первой производной, так что степень спектрального затухания увеличивается до $1/f^2$. Более быстрая скорость спектрального затухания приводит к меньшему

числу значимых спектральных членов. Еще одним преимуществом ДКП есть то, что поскольку данные четно-симметричные, их преобразование также является действительным и симметричным; следовательно, отсутствует необходимость в нечетно-симметричных базисных членах — функциях синуса.

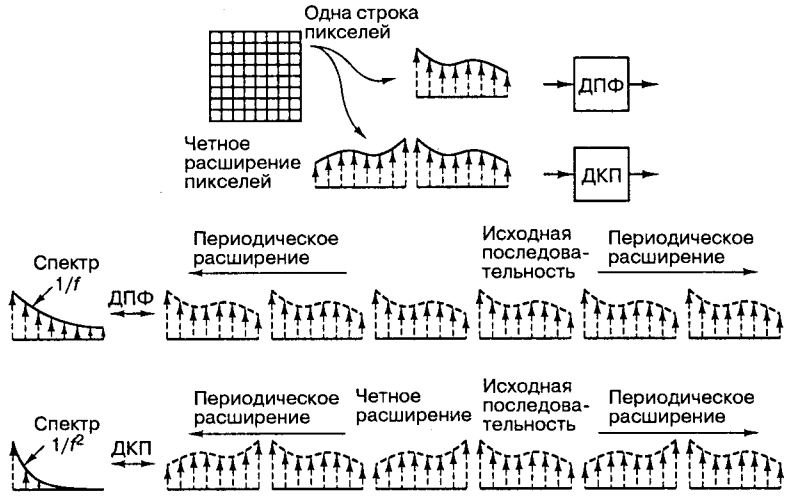


Рис. 13.40. Спектральное затухание и периодическое расширение временного ряда с помощью ДПФ и ДКП

Поскольку амплитуда образа имеет сильную корреляцию на небольших пространственных интервалах, значение ДКП блока 8×8 пикселей определяется, в основном, окрестностью постоянной составляющей и относительно небольшим числом иных значимых членов. Типичное множество амплитуд и их преобразование ДКП представлено на рис. 13.41. Отметим, что спектральные члены убывают, по крайней мере, как $1/f^2$ и большинство членов высокой частоты, в основном, нулевые. Спектр посылается на устройство квантования, которое использует стандартные таблицы квантования для присвоения бит спектральным членам согласно их относительным амплитудам и их психовизуальному значению. Для компонентов яркости и цветности используются различные таблицы квантования.

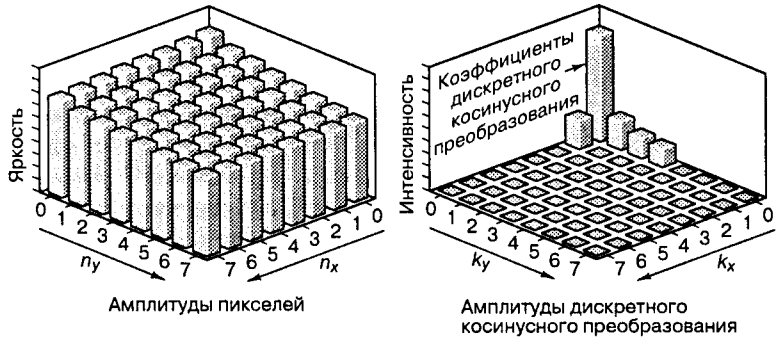


Рис. 13.41. Пиксели и амплитуды ДКП, описывающие один и тот же блок 8×8 пикселей

Чтобы использовать преимущество большого числа нулевых позиций в квантованном ДКП, спектральные адреса ДКП сканируются зигзагообразным образом, как изображено на рис. 13.42. Зигзагообразная модель обеспечивает длинную последовательность нулей. Это улучшает эффективность кодирования группового кода Хаффмана, описывающего спектральные выборки. На рис. 13.43 представлена блочная диаграмма кодера JPEG. Сигнал, доставленный на кодер, обычным образом представлен в виде растровой развертки с дискретными основными аддитивными цветами: красным, зеленым и синим (RGB). Цветная плоскость преобразуется в сигнал яркости (Y) и цветности $0{,}564 \times (B - Y)$ (обозначено как C_B) и $0{,}713 \times (R - Y)$ (обозначено как C_R), используя преобразование цветового контраста, разработанное для цветного ТВ. Это отображение описывается следующим образом.

$$\begin{bmatrix} Y \\ C_B \\ C_R \end{bmatrix} = \begin{bmatrix} 0{,}299 & 0{,}587 & 0{,}114 \\ -0{,}169 & -0{,}331 & 0{,}500 \\ 0{,}500 & -0{,}419 & -0{,}081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Здесь компонент Y образован для отражения чувствительности человеческого глаза к основным цветам.

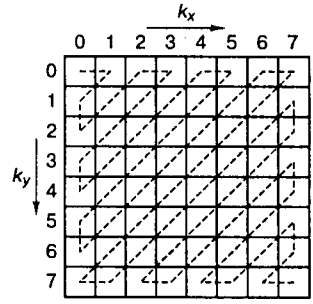


Рис. 13.42. Зигзагообразное сканирование спектральных составляющих ДКП

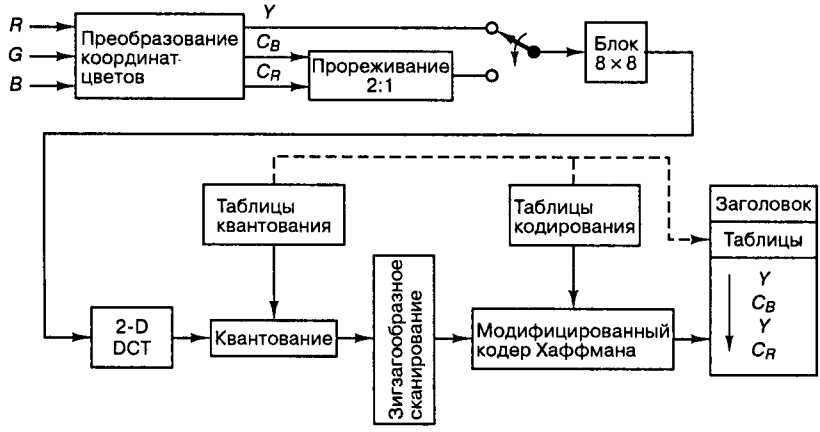


Рис. 13.43. Блочная диаграмма кодера JPEG

Глаз человека имеет разную чувствительность к цветным компонентам и компонентам яркости (черное и белое). Эта разница в способности к разрешению является следствием распределения рецепторов цвета (палочек) и рецепторов яркости (колбочек) на сетчатке. Человеческий глаз может различать 1-дюймовые чередующиеся черные и белые полоски со 180 футов (1/40 градуса). Для сравнения, 1-дюймовые сине-красные или сине-зеленые цветные полоски невозможно различить с расстояний, больших 40 футов (1/8 градуса). Следовательно, трехцветные образы требуют примерно на 1/25 (1/5 в каждом направлении) больше данных, чем нужно для получения черно-белого изображения. В далеком прошлом фотографы знали, что глаз требует очень малого числа цветных деталей. Чтобы придать образу цвет, существовала живая индустрия, в которой от руки раскрашивали черно-белые фотографии и почтовые открытки. Большинство аналоговых и цифровых цветных ТВ используют преимущество этой разницы в остроте восприятия для доставки дополнительных цветных компонентов через значительно сокращенную полосу частот. Стандарт NTSC определяет доставку всех трех цветов через полосу частот в 0,5 МГц, а не 4,2 МГц, действительно требуемую яркостным компонентом. Аналогично JPEG использует преимущество разницы в восприятии и выбирает компоненты цветового контраста с половинной частотой в направлении сканирования (x), но не в направлении поперек линий развертки (y).

Сигналы цветового контраста и сигналы с пониженной частотой дискретизации последовательно представлены как блоки 8×8 в двухмерном ДКП. Выходы ДКП квантуются с помощью соответствующей таблицы и затем зигзагообразно сканируются для передачи на кодер Хаффмана. JPEG использует кодер Хаффмана для кодирования коэффициентов переменной составляющей сигнала, но поскольку компоненты постоянной составляющей имеют высокую корреляцию между соседними блоками, для них используется дифференциальное кодирование. Разумеется, для формирования образа декодер обращает эти операции.

13.8.2.1.1. Варианты декодирования с помощью JPEG

Во время реконструкции образа декодер может работать последовательно, начиная с верхнего левого угла изображения и образуя блоки 8×8 пикселей по мере их поступления. Это последовательный режим JPEG. В прогрессивном режиме кодирования образ сначала объединяется в блоки 8×8, образованные только компонентом постоянной составляющей в каждом блоке. Это очень быстрый процесс, который представляет крупноблочный, но распознаваемый в результате предварительного просмотра образ, — процесс, часто демонстрируемый в Internet при загрузке файлов GIF (Graphic Interchange Format), которые в начале передачи данных доставляют только компоненты постоянной составляющей. Затем изображение обновляется в каждом блоке 8×8, образованном из компонентов постоянной составляющей и первых двух соседних компонентов, представляющих следующее множество данных, доставленных на декодер. И наконец, образ обновляется при полном разрешении посредством полного множества коэффициентов, связанных с каждым блоком 8×8.

При иерархическом кодировании образ кодируется и декодируется как перекрывающиеся кадры. Изображение с низким разрешением, выбранное с пониженной частотой (4:1 в каждом направлении), кодируется с использованием ДКП и квантованного коэффициента, образуя первый кадр. Изображение, полученное с помощью этого кадра, выбирается с более высокой частотой и сравнивается с версией исходного изображения большего разрешения (2:1 в каждом направлении), и разность, представляющая ошибку в формировании образа, снова кодируется как изображение MPEG. Два кадра, образо-

ванные двумя уровнями кодирования, используются для создания составного образа, который увеличивается и сравнивается с исходным образом. Разность между исходным образом и двумя уровнями реконструкций с более низкой разрешающей способностью формируется с наивысшей доступной разрешающей способностью, и снова применяется кодирование JPEG. Этот процесс полезен при доставке образов с последовательно высоким качеством реконструкции, подобно прогрессивному кодированию. Разница заключается в том, что имеется дополнительная разрешающая способность, но она не может быть послана до тех пор, пока не будет востребована. Пример: сканирование пользователем библиотеки изображений и требование окончательного качества после просмотра множества изображений. Еще одним примером может быть доставка одного уровня качества на дисплей персонального компьютера и более высокого уровня на дисплей рабочей станции с высокой разрешающей способностью.

В заключение отметим, что JPEG-2000 — это предложенный стандарт для определения *новой системы кодирования изображения*, предназначенной для Internet-приложений и мобильных приложений. В этой системе предлагается узкая полоса частот, множественная разрешающая способность, устойчивость к ошибкам, защищенность изображения и низкая сложность. Она базируется на алгоритмах волнового сжатия, и по отношению к JPEG в ней предлагается улучшенная эффективность сжатия со многими возможностями разрешения [28].

13.8.2.2. MPEG

MPEG (Motion Picture Experts Group — экспертная группа по вопросам движущегося изображения) представляет собой стандарты, созданные для поддержания *кодирования движущихся изображений и ассоциированного аудио для среды цифрового запоминания со скоростями до 1,5 Мбит/с*. MPEG-1, стандарт ISO 11172, был принят в ноябре 1992 года для разрешения записи полномасштабного видео на CD-плейерах, первоначально созданных для стерео-аудиовоспроизведения. MPEG-2, стандарт ISO 13818 или рекомендация ITU-T H.262, *Универсальное кодирование движущихся изображений и ассоциированного аудио*, принятый в ноябре 1994 года, дает большую гибкость форматов входа/выхода, большую скорость передачи данных и уделяет больше внимания таким системным требованиям, как передача и синхронизация, темам, не рассмотренным в MPEG-1. MPEG-2 поддерживает разновидности цифрового ТВ, охватывающие оцифрованное видео, которое отображает существующий аналоговый формат с определенным качеством посредством DVD (цифровой видеодиск) и HDTV (телевидение высокой четкости) с различными форматами изображения, частоты развертки, скорости сканирования пикселей, опций обратного сканирования и различными опциями выборки на повышенной частоте для компонентов цветового контраста. Ниже описывается основная теория работы простейшей версии MPEG-2.

MPEG-2. MPEG сжимает последовательность движущихся образов, используя преимущество высокой корреляции между последовательными движущимися изображениями. MPEG создает три типа изображений: интра-изображения (*I*-изображения), предсказанные (*P*-изображения) и изображения двунаправленного предсказания (*B*-изображения). В MPEG каждое *M*-е изображение в последовательности может быть полностью сжато с использованием стандартного алгоритма JPEG; это *I*-изображения. Затем процесс сравнивает последовательные *I*-изображения и идентифицирует часть образа, которая была перемещена. Части образа, которые не были перемещены, переносятся в промежуточное изображение с помощью памяти декодера. После этого процесс отбирает подмножество промежуточных изображений, а затем предсказывает

(посредством линейной интерполяции между *I*-изображениями) и корректирует расположение частей образа, которые были перемещены. Эти предсказанные и скорректированные образы являются *P*-изображениями. Между *I*- и *P*-изображениями находятся *B*-изображения, которые включают стационарные части образа, не охваченные движущимися частями. Относительное расположение этих изображений показано на рис. 13.44. Отметим, что *P*- и *B*-изображения допускаются, но не требуются, и их количество является переменным. Последовательность может быть образована без каких бы то ни было *P*- или *B*-изображений, но последовательность, содержащая только *P*- или *B*-изображения, не может существовать.

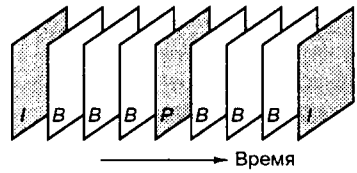


Рис. 13.44. Последовательность изображений при сжатии MPEG

I-изображения сжимаются так, как если бы они были изображениями JPEG. Это сжатие применяется к четырем непрерывным блокам 8×8, называемым макроблоками. Макроблоки могут быть выбраны с пониженной частотой для последовательного сжатия цветных компонентов. Макроблоки и их опции выборки с пониженной частотой изображены на рис. 13.45. Сжатие *I*-кадра производится независимо от ранних или поздних изображений в последовательности кадров. Расстояние в последовательности, рассчитанное между *I*-изображениями, является регулируемым, и оно может быть сделано малым порядка 1 либо настолько большим, насколько позволяет память. Редактирование сечений в последовательности изображений и локальная программная вставка могут производиться только с *I*-изображениями. Поскольку одна вторая секунды — это приемлемая временная точность для производства такого дополнения, расстояние между *I*-изображениями обычно ограничено примерно 15 изображениями для стандарта NTSC (30 изображений в секунду) или 12 изображениями для Британского стандарта PAL (25 изображений в секунду).

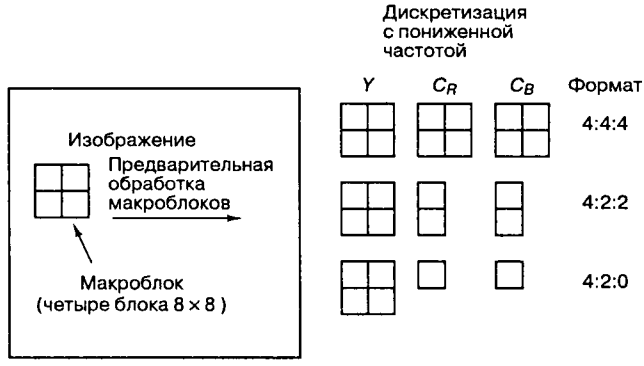


Рис. 13.45. Обработка макроблока для выборки цветности с пониженной частотой

Первым этапом обработки, производимым MPEG, является определение, какой из макроблоков перемещается между *I*-изображениями. Это выполняется путем переноса каждого макроблока из одного *I*-кадра вперед к следующему и вычисления двухмерной взаимной корреляции в окрестности его исходного расположения. Для каждого сдвинутого макроблока определяются векторы движения, которые указывают направление и величину перемещения. Макроблоки, которые не сдвигались, являются стационарными в картинах между *I*-изображениями и могут быть вынесены вперед в промежуточных изображениях.

Следующий этап обработки в MPEG состоит в образовании *P*-кадра между *I*-изображениями. Сначала предположим, что сдвинутые макроблоки перемещались линейно во времени между двумя положениями, определенными на первом этапе обработки. Каждый макроблок помещается на свое предсказанное положение в *P*-кадре. Вычисляется взаимная корреляция в окрестности этого блока для определения истинного расположения макроблока в *P*-кадре. Разность между предсказанным и истинным положениями макроблока является ошибкой предсказания, и эта ошибка сжимается с помощью ДКП и используется для коррекции *P*-кадра. Та же информация передается на декодер, так что он может корректировать свои предсказания. На рис. 13.46 представлен сдвиг макроблока между *I*-изображениями и промежуточное *P*-изображение.

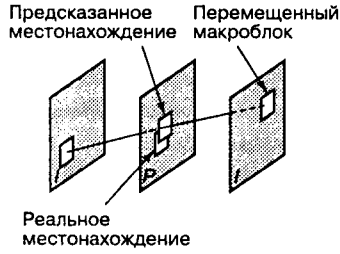


Рис. 13.46. Движение макроблока между I- и P-изображениями

B-изображения расположены между *I*- и *P*-изображениями. В этих изображениях векторы движения передвигают сдвинутые макроблоки линейно во времени к их двунаправленным интерполированным положениям в каждом последовательном *B*-кадре в последовательности. *I*-изображения требуют максимального количества данных для описания их содержания, сжатого с помощью ДКП. *P*-изображения требуют меньше данных. Они служат только для описания пикселей, ошибочно предсказанных на основании движения макроблоков в кадре. Остаток пикселей в кадре выносится вперед в память из предшествующего *I*-кадра. *B*-изображения являются наиболее эффективными изображениями множества. Они должны только *линейно сдвинуть* и *скорректировать* пиксели, охваченные и неохваченные в результате движения макроблоков через кадры.

Реконструкция образов на декодере требует того, чтобы последовательность образов была доставлена в порядке, необходимом для соответствующей обработки. Например, поскольку вычисление *B*-изображений требует информации от *I*- и *P*-изображений или *P*- и *P*-изображений с обеих сторон, *I*- и *P*-изображения должны быть доставлены первыми. Рассмотрим следующий пример требуемого порядка кадров на входе и выходе кодера и декодера.

Порядок изображений на входе кодера

1	2	3	4	5	6	7	8	9	10	11	12	13
I_0	B_1	B_2	P_1	B_3	B_4	P_2	B_5	B_6	I_{n+1}	B_1	B_2	P_1

Порядок закодированных изображений на выходе кодера и входе декодера

1	2	3	4	5	6	7	8	9	10	11	12	13
I_0	P_1	B_1	B_2	P_2	B_3	B_4	I_{n+1}	B_5	B_6	P_1	B_1	B_2

Порядок изображений на выходе декодера

1	2	3	4	5	6	7	8	9	10	11	12	13
I_0	B_1	B_2	P_1	B_3	B_4	P_2	B_5	B_6	I_{n+1}	B_1	B_2	P_1

На рис. 13.47 представлена блочная диаграмма кодера MPEG. Отметим, что его структура представляет собой стандартную модель предсказания-коррекции. Отметим интересное соотношение между воспринимаемой глазом мерой качества изображения и мерой его активности. С одной стороны, когда образ содержит значительное движение, глаз воспринимает образы более низкого качества. С другой стороны, когда образ содержит мало движения, глаз чувствителен к помехам изображения. В кодере отсутствие движения влияет на активность кодирования и приводит к тому, что данные доставляются на выход буфера с более низкой скоростью. Буфер считает это индикатором стационарности образов и контролирует образ, допуская квантование ДКП более высокого качества. Скорость на выходе буфера фиксируется согласно требованиям линий связи. Для отображения средней входной скорости в фиксированную выходную применяется текущий контроль. Текущий контроль регистрирует низкую активность кодера, замечая, что его буфер опустошается быстрее, чем наполняется. Простой индикатор разности между входной и выходной скоростями — это расположение выходного адресного указателя. Если указатель движется по направлению к началу памяти буфера, указателю опустошения памяти, система увеличивает входную скорость, выбирая таблицу квантования, которая дает большее число бит на ДКП. Аналогично, если указатель движется по направлению к концу памяти буфера, указателю переполнения, система увеличивает выходную скорость, выбирая таблицу квантования, которая дает меньшее число бит на ДКП. Этот процесс согласовывает качество изображения с порогом качества, воспринимаемым глазом, сохраняя при этом среднюю выходную скорость канала.

13.9. Резюме

В этой главе представлены некоторые основные моменты кодирования источника. Здесь показано, что кодирование источника может быть применено к цифровым данным и к волновым сигналам. Цифровые данные могут быть точно восстановлены путем сокращенного описания данных источника, если источник демонстрирует корреляцию между элементами алфавита или элементы не являются равновероятными. Вообще говоря, волновые сигналы, представленные в цифровой форме, искажены. Это искажение может быть сделано произвольно малым посредством соответствующего увеличения скорости передачи битов, требуемой для описания источника. Кодирование источника может быть также применено к волновым источникам для получения описания с меньшей скоростью передачи данных, если для источника характерен большой радиус корреляции или возможные амплитуды не являются равновероятными.

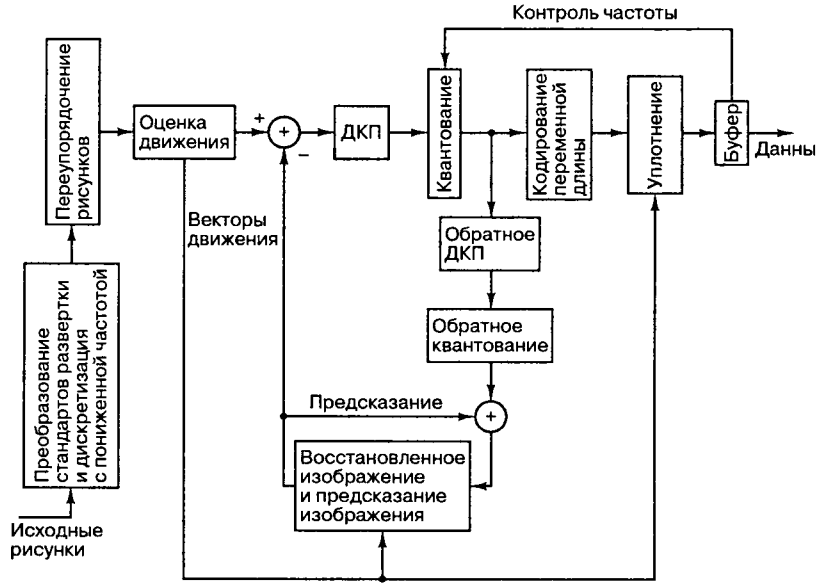


Рис. 13.47. Блочная диаграмма кодера MPEG с встроенным декодером

Преимущество системы кодирования источника состоит в сокращении необходимой полосы частот системы и/или энергии на бит, требуемых для получения описания источника. Это преимущество имеется и при определении компромиссов с еще один ресурсом системы — сложность вычисления и память. За счет этих ресурсов, стоимость которых в последние десятилетия продолжает падать, кодирование источника обещает получить постоянно возрастающую роль в системах связи и запоминания. Заинтересованный читатель может ознакомиться с работами [8, 17, 24–26], в которых кодирование источника рассмотрено весьма подробно.

Литература

1. Papoulis A. *Probability, Random Variables, and Stochastic Processes* McGraw-Hill Book Company, New York, 1965.
2. Harri F. J. *Windows, Harmonic Analysis, and the Discrete Fourier Transform.* Proc. IEEE, vol. 67, January, 1979.
3. Martin G. *Gyroscopes May Cease Spinning.* IEEE Spectrum, vol. 23, n. 2, February, 1986, pp. 48–53.
4. Vanderkooy J. and Lipshitz S. T. *Resolution beyond the Least Significant Bit with Dither.* J. Audio Eng. Soc., n. 3, March, 1984, pp. 106–112.
5. Blesser B. A. *Digitization of Audio: A Comprehensive Examination of Theory, Implementation, and Current Practice.* J. Audio Eng. Soc., vol. 26, n. 10, October, 1978, pp.739–771.
6. Sluyter R. J. *Digitization of Speech.* Phillips Tech. Rev., vol. 41, n. 7-8, 1983-84, pp. 201–221.
7. Bell Telephone Laboratories Staff. *Transmission Systems for Communications.* Western Electric Co. Technical Publications, Winston-Salem, N. C., 1971.
8. Jayant N. S. and Noil P. *Digital Coding of Waveforms/* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1984.
9. Marcel J. D. and Gray A. H. Jr. *Linear Prediction of Speech* Springer-Verlag, New York, 1976.
10. Deller J., Proakis J. and Hansen J. *Discrete-Time Processing of Speech Signals.* Macmillan, New York, 1993.
11. Candy J. and Temes G. *Oversampling Delta-Sigma Data Converters.* IEEE Press, 1991.

12. Dick C. and Harris F. FPGA *Signal Processing Sigma-Delta Modulation* IEEE Signal Proc. Mag., Vol. 17., n. 1, January, 2000, pp. 20–35.
13. Cummisky P., Jayant N. and Flanagan J. *Adaptive Quantization in Differential PCM Coding of Speech/* Bell Syst. Tec J., Vol. 52, 1973, pp. 115–119.
14. Gersho A. *Asymptotically Optimal Block Quantization.* IEEE Trans. Inf. Theory, vol. IT25, n. 4, July, 1979, pp. 373–380.
15. Gersho A. *On the Structure of Vector Quantizers.* IEEE Trans. Inf. Theory, vol. IT28, n. 2, March, 1982, pp. 157–166.
16. Abut H. *Vector Quantization.* IEEE Press, 1990.
17. Jeffress L. *Mascing*; in J. Tobias, ed., Foundations of Modern Auditory Theory. Academic Press, Inc., New York, 1970.
18. Lynch T. J. *Data Compression Techniques and Applications.* Lifetime Learning Publications, New York, 1985.
19. Schafer R. W. and Rabiner L. R. *Design of Digital Filter Banks for Speech Analysis.* Bell Syst. Tech. J., vol. 50, n. 10, December, 1971, pp. 3097–3115.
20. Huffman D. A. *A Method for the Construction of Minimum Redudancy Codes.* Proc. IRE, vol. 40, September, 1952, pp. 1098–1101.
21. Hamming R. W. *Coding and Information Theory.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1980.
22. Hunter R. and Robinson A. *International Digital Facsimile Coding Standard.* Proc. IEEE, Vol. 68, n. 7, July, 1980, pp. 854–867.
23. McConnel K., Bodson D. and Urban S. *FAX: Facsimile Technology and Systems.* Artech House, 1999.
24. Cox R. *Three New Speech Coders From the ITU Cover a Range of Applications.* IEEE Comm. Mag., Vol. 35, n. 9, September, 1997, pp. 40–47.
25. Noll P. *Wideband Speech and Audio Coding.* IEEE Comm. Mag., Vol. 31, n. 11, November, 1993, pp. 34–44.
26. Solari S. *Digital Video and Audio Compression.* McGraw-Hill, New York, 1997.
27. Rzeszewski T. *Digital Video: Concepts and Applications Across Industries.* IEEE Press, 1995.
28. Ebrahimi T., Santa Cruz. D., Christopoulos C., Askelof J., Larsson M. *JPEG 2000 Still Image Coding Versus Other Standards.* SPIE International Symposium, 30 July–4 August 2000, Special Session on JPEG2000, San Diego, CA.

Задачи

13.1. Дискретный источник генерирует три независимых символа A, B и C с вероятностями 0,9, 0,08 и 0,02. Определите энтропию источника.

13.2. Дискретный источник генерирует два независимых символа A и B с следующими условными вероятностями.

$$P(A|A) = 0{,}8 \quad P(B|A) = 0{,}2$$
$$P(A|B) = 0{,}6 \quad P(B|B) = 0{,}4$$

а) Определите вероятности символов A и B.

б) Определите энтропию источника.

в) Определите энтропию источника, если символы независимы и имеют те же вероятности.

13.3. 16-битовый аналого-цифровой преобразователь работает с входным диапазоном в ±5,0 В.

а) Определите размер квантили.

б) Определите среднеквадратическое напряжение шума квантования.

в) Определите среднее SNR (вследствие квантования) для полномасштабного входного синусоидального сигнала.

г) Считайте, что расстояние в 100 миль, пройденное автомобилем, измеряется с той же точностью, что и в 16-битовом преобразователе. Чему равна среднеквадратическая ошибка в футах?

13.4. 10-битовый АЦП работает с входным диапазоном в ±5,0 В.

а) Определите размер единичного шага квантили.

б) Для (полномасштабной) синусоиды в 5,0 В определите выходное отношение сигнала к шуму квантования.

в) Для синусоиды ($\frac{1}{100}$ полного масштаба) в 0,050 В определите выходное отношение сигнала к шуму квантования.

г) Для входного сигнала, имеющего гауссово распределение амплитуд, вероятность насыщения контролируется присоединением входного аттенюатора, так что уровень насыщения соответствует четырем среднеквадратическим отклонениям. Определите выходное отношение сигнала к шуму квантования.

д) Определите вероятность насыщения сигнала, описанного в п. г.

13.5. Определите оптимальную характеристику сжатия для входной функции плотности (аппроксимации непрерывной функции плотности), изображенной на рис. 313.1.

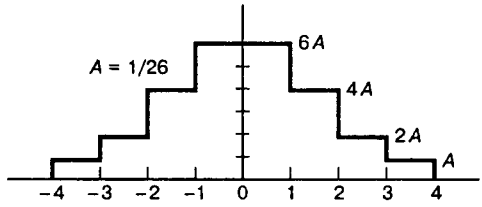


Рис. 313.1

13.6. 10-битовый преобразователь, использующий μ-закон, работает с полным диапазоном в ±5,0 В.

а) Если μ = 100, определите выходное отношение сигнала к шуму квантования для синусоиды в 5,0 В (полномасштабной).

б) Если μ = 100, определите выходное отношение сигнала к шуму квантования для синусоиды в 0,050 В ($\frac{1}{100}$ полного масштаба).

в) Повторите пп. а и б для μ = 250.

13.7. Записывающая система компакт-диска отображает каждый из двух стереосигналов с помощью 16-битового АЦП в $44{,}1\times 10^3$ выборок/с.

а) Определите выходное отношение сигнала к шуму для полномасштабной синусоиды.

б) Если записываемая музыка создана для коэффициента пиковой импульсной нагрузки (отношение максимального значения к среднеквадратическому), равного 20, определите среднее выходное отношение сигнала к шуму квантования.

в) Поток оцифрованных битов дополнен битами коррекции ошибок, битами подстановки (для извлечения сигнала ФАПЧ), полями битов изображения и управления. Эти дополнительные биты составляют 100% служебных издержек, т.е. для каждого бита, генерированного АЦП, сохраняется 2 бит. Определите выходную скорость передачи битов воспроизводящей системы проигрывания компакт-дисков.

г) На компакт-диск можно записать порядка часа музыки. Определите число бит, записанных на компакт-диск.

д) Для сравнения, хороший академический словарь может содержать 1 500 страниц, 2 колонки/страницу, 100 строк/колонку, 7 слов/строку, 6 букв/слово и 6 бит/букву. Определите число битов, требуемое для представления словаря, и оцените число подобных книг, которое может быть записано на компакт-диске.

13.8. 1-битовое устройство квантования дискретизирует входную синусоиду амплитуды A с равномерно распределенной фазой. Определите амплитуду x_0, выходной уровень 1-битового квантующего устройства, минимизирующего среднеквадратическую ошибку квантования.

13.9. Одношаговый линейный фильтр с предсказанием должен использоваться для дискретизации синусоиды постоянной амплитуды. Отношение частоты произведения выборки к час-

тоте синусоиды равно 10,0. Определите коэффициент предсказания фильтра. Определите отношение выходной мощности к входной для одноотводного предсказателя.

13.10. Двухотводный линейный фильтр с предсказанием работает в системе DPCM. Предсказание имеет вид $\hat{x}(n) = a_1 x(n-1) + a_2 x(n-2)$.

а) Определите величины a_1^{opt} и a_2^{opt}, минимизирующие среднеквадратическую ошибку предсказания.

б) Определите выражение для среднеквадратической ошибки предсказания.

в) Определите мощность ошибки предсказания, если коэффициент корреляции входного сигнала имеет следующий вид.

$$c(n) = \begin{cases} 1 - |n| & \text{для } n = -4, -3, -2, -1, 0, 1, 2, 3, 4 \\ 0 & \text{для других } n \end{cases}$$

г) Определите мощность ошибки предсказания, если коэффициент корреляции входного сигнала имеет вид $C(n) = \cos \theta_0 n$.

13.11. Одноконтурный сигма-дельта-модулятор работает с частотой, в 20 раз превышающей частоту Найквиста для сигнала с полосой частот 10 кГц. Преобразователь представляет собой 1-битовый АЦП.

а) Определите максимальное SNR для входного сигнала в 8,0 кГц.

б) Определите максимальное SNR для того же сигнала, если модулятор работает с частотой, в 50 раз превышающей частоту Найквиста.

в) Определите максимальное SNR для того же сигнала, если модулятор заменен на 2-нулевой модулятор, работающий с частотой, в 20 раз превышающей частоту Найквиста.

13.12. Создайте двоичный код Хаффмана для дискретного источника трех независимых символов A, B и C с вероятностями 0,9, 0,08 и 0,02. Определите среднюю длину кода для этого кода.

13.13. Создайте двоичный код расширения первого порядка (кодирование двух символов одновременно) для дискретного источника, описанного в задаче 13.12. Определите среднюю длину кода на символ для этого кода.

13.14. Входной алфавит (клавиатура текстового процессора) состоит из 100 символов.

а) Если нажатие клавиши кодируется с помощью кода фиксированной длины, определите требуемое число бит для кодирования.

б) Сделаем упрощающее предположение, состоящее в том, что 10 нажатий клавиш равновероятны и каждое происходит с вероятностью 0,05. Предположим также, что оставшиеся 90 нажатий клавиш равновероятны. Определите среднее число бит, требуемое для кодирования этого алфавита с использованием кода Хаффмана переменной длины.

13.15. Используйте модифицированный MKKTT факсимильный код Хаффмана для кодирования следующей последовательности единственной строки из 2 047 черных и белых пикселей. Определите отношение закодированных битов к входным.

11Б 1Ч 2Б 2Ч 4Б 4Ч 8Б 8Ч 16Б 16Ч 32Б 32Ч
664Б 64Ч 128Б 128Ч 256Б 256Ч 512Б 512Ч 1Б

13.16. JPEG квантует спектральные составляющие, полученные с помощью ДКП четного расширения обработанных данных. Чтобы показать относительные потери ДКП и БПФ, образуйте четное и скопированное расширения ряда {10 12 14 16 18 20 22 24}, чтобы получить {10 12 14 16 18 20 22 24 10 12 14 16 18 20 22 24} и {10 12 14 16 18 20 22 24 24 22 20 18 16 14 12 10}. Примените ДПФ к двум временным рядам и сравните относительный размер спектральных компонентов (отличных от постоянных составляющих). Теперь дополните спектр, полагая равными нулю все лепестки, кроме 5 спектральных. В четном расширении удерживаются лепестки {1 2 3 15 16}, в то время как в периодическом — {1 3 5 13 15}. Вычислите обратное ДПФ каждого и сравните относительный размер ошибки восстановления для двух преобразований.

13.17. JPEG использует зигзагообразную модель сканирования для обращения к спектральным составляющим ДКП, доставленным квантующим устройством. Альтернативной моделью сканирования будет растровое сканирование, сканирование последовательных строк,

обычно выполняемое при сканировании изображения. Сравните эффективность сканирования зигзагообразным методом с эффективностью растрового сканирования, если ненулевыми спектральными членами являются $S(0, 0) = 11001100$, $S(1, 0) = 10101$ и $S(0, 1) = 110001$. Используйте модифицированный код Хаффмана из табл. 13.1 для определения размеров групп нулей. Предположите, что следующая таблица определяет битовое присвоение на спектральный лепесток.

8	6	5	4	3	2	2	2
6	5	4	3	2	2	1	1
5	4	3	2	2	1	1	1
4	3	2	2	1	1	1	1
3	2	2	1	1	1	1	1
2	2	1	1	1	1	1	1
2	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1

13.18. ДКП преобразует блок 8×8 пикселей, содержащий 8-битовые слова, в блок 8×8 спектральных выборок, содержащих число бит, определенных в таблице квантования задачи 13.17. Предполагается, что не существует последовательностей нулей переменной длины, так что в выходе ДКП представлен каждый лепесток; вычислите коэффициент сжатия (отношение входных битов к выходным), приписанный ДКП. Вычислите коэффициент сжатия, предполагая, что количество значимых коэффициентов ДКП ограничено верхним треугольником таблицы квантования, состоящей из одного 8-битового слова, двух 6-битовых и трех 5-битовых, с оставшимися битами, которые описываются кодом для 101 нуля.

Вопросы для самопроверки

13.1. Почему сигналы подвергаются операциям *кодирования источника*, перед передачей или запоминанием (см. разделы 13.1 и 13.7)?

13.2. Какие свойства *непрерывного сигнала* позволяют представить его с помощью уменьшенного числа бит на выборку (см. разделы 13.1, 13.3 и 13.7)?

13.3. Какие свойства *дискретного сигнала* позволяют представить его с помощью уменьшенного числа бит на символ (см. раздел 13.1 и 13.7)?

13.4. Большинство квантующих устройств являются *равномерными* относительно размера шага. Существуют приложения, для которых требуются *неравномерные* квантующие устройства. Они иногда называются *компандирующими* квантующими устройствами. Зачем нужны подобные квантующие устройства (см. раздел 13.2.5)?

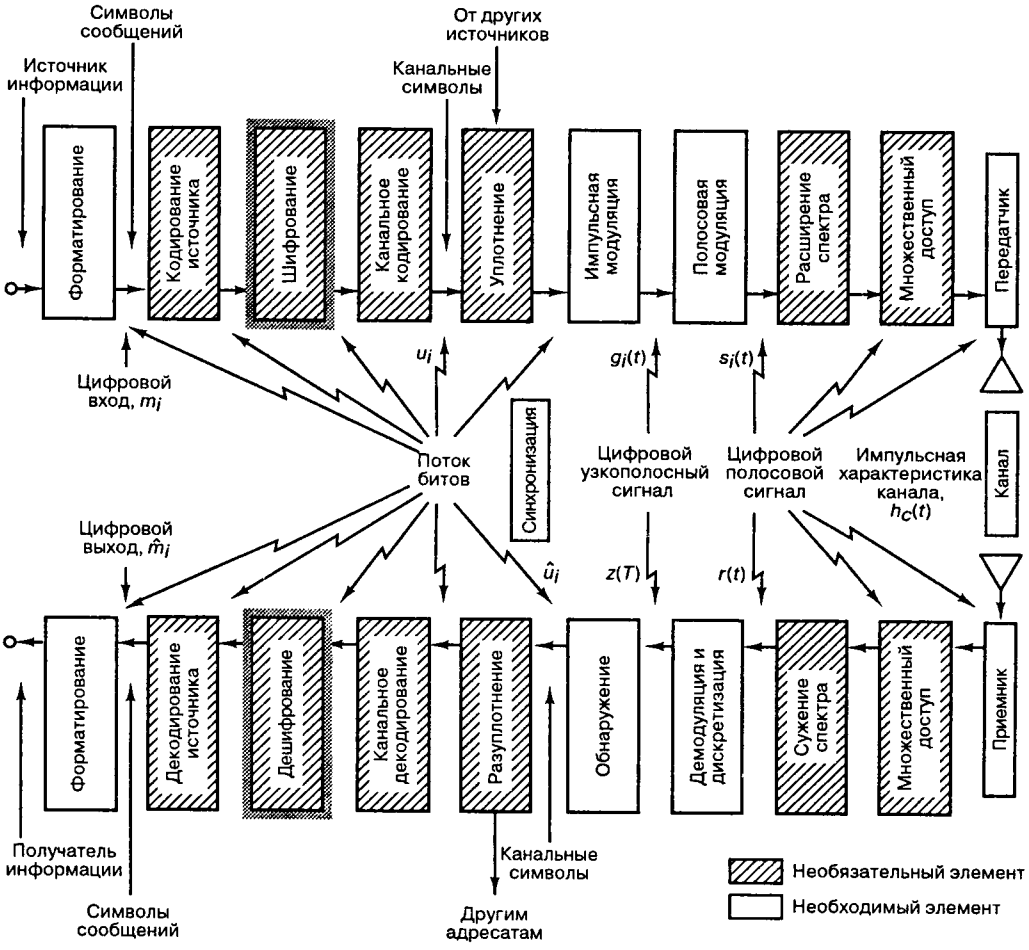
13.5. Аналого-цифровой преобразователь (analog-to-digital converter — ADC, АЦП) представляет выборочные данные сигнала с помощью такого числа бит на выборку, которое удовлетворяет требуемой точности. Большинство АЦП являются квантующими устройствами *без памяти*, что означает, что каждое квантование (преобразование) производится независимо от других преобразований. Как может использоваться память для ограничения числа бит на выборку (см. раздел 13.3)?

13.6. Кодирование источника уменьшает *избыточность* и отбрасывает *несущественное* содержимое. В чем состоит разница между избыточностью и несущественностью (см. раздел 3.7)?

13.7. Часто слышим такое высказывание, как "картина стоит тысячи слов". Действительно ли картина стоит тысячи слов (см. раздел 13.8.2)?

ГЛАВА 14

Шифрование и дешифрование



14.1. Модели, цели и ранние системы шифрования

14.1.1. Модель процесса шифрования и дешифрования

Желание общаться конфиденциально уходит своими корнями в далекое прошлое. История секретного общения богата уникальными изобретениями и красочными анекдотами [1]. Изучение путей передачи сообщений, которые не допускали бы постороннего вмешательства, называется *криптографией*. Термины *шифрование* и *кодирование* обозначают преобразования сообщений, выполняемые передатчиком, а термины *дешифрование* и *декодирование* — обратные преобразования, производимые приемником. Основными причинами использования криптосистем в общении являются (1) *обеспечение конфиденциальности*, т.е. предотвращение извлечения информации из канала посторонним лицом (подслушивание); (2) *аутентификация*, предотвращение внедрения в канал информации посторонними людьми (обманный доступ). Часто, как в случае электронной пересылки или договорных переговоров, важно обеспечить электронный эквивалент *письменной подписи*. Это необходимо для того, чтобы устранить какие-либо недоразумения между отправителем и получателем относительного того, какое сообщение было отправлено и было ли оно вообще отправлено.

На рис. 14.1 изображена модель криптографического канала. Сообщение, или открытый текст M, шифруется путем обратимого преобразования E_K, дающего шифрованный текст $C = E_K(M)$. Шифрованный текст пропускается через незащищенный, или *общедоступный канал*. После получения шифрованного сообщения C, его исходное значение восстанавливается с помощью операции дешифрования, описываемой обратным преобразованием $D_k = E_k^{-1}$, что выглядит следующим образом.

$$D_k(C) = E_k^{-1}\left[E_k(M)\right] = M \tag{14.1}$$

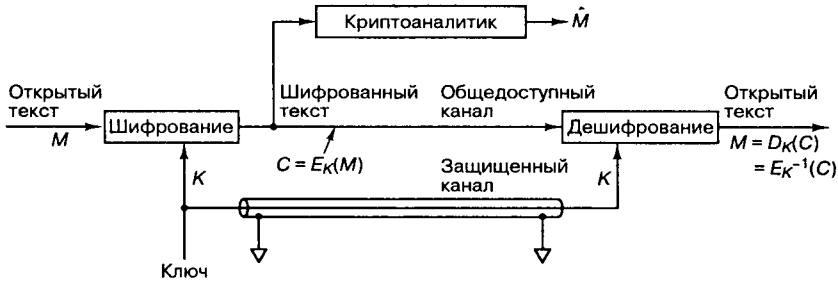


Рис. 14.1. Модель криптографического канала

Параметром K обозначается множество символов или характеристик, называемых *ключом*, определяющим конкретное шифрующее преобразование E_K из семейства криптографических преобразований. Первоначально защищенность криптосистем зависела от секретности всего процесса шифрования, но в конечном итоге были разработаны системы, для которых общая природа преобразования шифрования или алгоритма могла быть общеизвестна, а секретность системы зависела от специального ключа. Ключ использовался для шифрования нешифрованного сообщения, а также для дешифрования шифрованного сообщения. Здесь можно отметить аналогию с универсальным компьютером и компью-

терной программой. Компьютер, подобно криптосистеме, способен на множество преобразований, из которых компьютерная программа, подобно специальному ключу, выбирает одно. В большинстве криптосистем каждый, имеющий доступ к ключу, может как шифровать, так и дешифровать сообщения. Ключ передается разрешенным пользователям через секретный канал (в качестве примера может быть использован курьер для передачи из рук в руки важной ключевой информации); ключ, как правило, остается неизменным в течение значительного числа передач. Целью *криптоаналитика* (противника) является оценка открытого текста $\hat{M}$ посредством анализа шифрованного текста, полученного из общедоступного канала, без использования ключа.

Схемы шифрования можно разбить на две основные категории: *блочное* и *шифрование потока данных*, или просто *поточное*. При блочном шифровании нешифрованный текст делится на блоки фиксированного размера, после чего каждый блок шифруется независимо. Следовательно, одинаковые блоки открытого текста с помощью данного ключа будут преобразовываться в одинаковые блоки шифрованного текста (подобно блочному кодированию). При поточном шифровании (подобном сверточному кодированию) блоков фиксированного размера не существует. Каждый бит открытого текста m_i шифруется с помощью i-го элемента k_i последовательности символов (ключевого потока), генерируемой ключом. Процесс шифрования является *периодическим*, если ключевой поток начинает повторяться после p символов (причем p фиксированно); в противном случае он является непериодическим.

В общем случае схема шифрования существенно отличается от схемы канального кодирования. Например, при шифровании данные открытого текста не должны явно фигурировать в шифрованном тексте, а при канальном кодировании в *систематической форме* коды часто содержат неизмененные биты сообщения плюс биты четности (см. раздел 6.4.5). Существуют и другие отличия шифрования и канального кодирования. При блочном шифровании единственный бит ошибки на входе дешифратора может изменить значение многих выходных битов в блоке. Этот эффект, известный как *накопление ошибки* (error propagation), часто является желаемым криптографическим свойством, поскольку для несанкционированных пользователей он создает дополнительные сложности при расшифровке сообщений. В то же время при канальном кодировании такое свойство является нежелательным, поскольку хотелось бы, чтобы система исправила как можно больше ошибок и на выходную информацию входные ошибки относительно не влияли.

14.1.2. Задачи системы шифрования

Основные требования к системе шифрования можно сформулировать следующим образом.

1. Обеспечить *простые* и *недорогие средства* шифрования и дешифрования для разрешенных пользователей, обладающих соответствующим ключом.
2. Задачу криптоаналитика по производству оценки нешифрованного текста без помощи ключа сделать максимально *сложной* и *дорогой*.

Последовательно создаваемые криптосистемы делятся на *безусловно защищенные* или схемы, *защищенные по вычислениям*. Говорят, что система *безусловно защищена*, если информации, имеющейся у криптоаналитика, не достаточно для определения преобразований шифрования и дешифрования, независимо от того, какой вычислительной мощностью он располагает. Одна из таких систем, которая называется системой *разового заполне-*

ния, включает шифрование сообщения с помощью случайного ключа, который применяется только один раз. Ключ никогда не используется повторно; следовательно, криптоаналитик не получает информации, которая может использоваться для расшифровки последующих передач, использующих тот же ключ. Хотя такая система является безусловно защищенной (см. раздел 14.2.1), в общепринятой системе связи она применяется редко, поскольку для каждого нового сообщения необходимо распространить новый ключ, а это обычно затруднительно. Вообще, распределение ключей разрешенным пользователям является основной проблемой при использовании любой криптосистемы, даже если ключ применяется в течение продолжительного периода времени. Хотя и можно доказать, что некоторые системы являются безусловно защищенными, общей схемы доказательства защищенности произвольной криптосистемы в настоящее время не существует. Таким образом, в спецификациях большинства криптосистем формально указывается, что они *защищены по вычислениям* на *x* лет; это означает, что при обстоятельствах, благоприятных для криптоаналитика (т.е. при использовании самых современных компьютеров), защита системы может быть взломана за *x* лет, но никак не ранее.

14.1.3. Классические угрозы

Самая незначительная криптоаналитическая угроза — это *атака только шифрованного текста* (ciphertext-only attack). При использовании этого метода криптоанализа криптоаналитик может иметь *некоторую* информацию об общей системе и языке, используемом в сообщении, но единственными важными данными, имеющимися у него, является шифрованное сообщение, перехваченное из общедоступного канала.

Более серьезной угрозой для системы является *атака известного открытого текста* (known plaintext attack). Она включает в себя знание открытого текста *и* его шифрованного эквивалента. Жесткая структура большинства бизнес-форм и языков программирования часто дает оппоненту множество априорных знаний об элементах открытого сообщения. Вооруженный этим знанием и шифрованным сообщением, криптоаналитик может проводить криптоанализ с помощью известного открытого текста. Рассмотрим пример из области дипломатии: если шифрованное сообщение обязывает министра иностранных дел сделать определенное публичное заявление и он делает это, не перефразируя сообщение, криптоаналитик может получить как шифрованный текст, так *и* его точный перевод в открытую версию. Несмотря на то что атака известного открытого текста не всегда возможна, она используется достаточно часто, чтобы система не считалась защищенной, если она не проектировалась для противостояния такому типу атак [2].

Если криптоаналитик должен выбирать открытый текст для данного шифрованного сообщения, угроза называется *атакой выбранного открытого текста* (chosen plaintext attack). Во время Второй мировой войны такая атака использовалась Соединенными Штатами Америки для получения большей информации о японской криптосистеме. 20 мая 1942 года главнокомандующий Императорским Морским флотом адмирал Ямамото (Yamamoto) издал указ, детально излагающий тактику, которая должна была быть использована при атаке на острове Мидуэй. Этот указ был перехвачен подслушивающими постами союзников. К тому времени американцы узнали достаточно о японских кодах, чтобы дешифровать большинство сообщений. Однако все еще под сомнением были некоторые важные моменты, такие как *место* атаки. Они подозревали, что символы "AF" обозначали остров Мидуэй, но для того, чтобы убедиться, Джозеф Рошфор (Joseph Rochefort), глава военной разведки, решил использовать метод атаки выбранного открытого текста, чтобы обманным путем вынудить

японцев дать конкретное доказательство. По его приказу гарнизон острова Мидэуй выдал в эфир характерное открытое сообщение, в котором остров Мидуэй сообщал, что его завод по очистке воды вышел из строя. Американским криптоаналитикам пришлось подождать всего два дня, после чего они перехватили японское шифрованное сообщение, в котором говорилось, что на AF не хватает чистой воды [1].

14.1.4. Классические шифры

Одним из ранних примеров моноалфавитного шифра был *шифр Цезаря*, который использовался Юлием Цезарем во времена его Галльских походов. Каждая буква исходного текста заменяется новой, полученной путем *сдвига алфавита*. На рис. 14.2, *а* изображено такое шифрующее преобразование, состоящее из трех циклических сдвигов алфавита. Если использовать этот алфавит Цезаря, сообщение "now is the time" ("время пришло!") шифруется следующим образом.

Открытый текст:	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Шифрованный текст:	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C

а)

	1	2	3	4	5
1	A	B	C	D	E
2	F	G	H	I/J	K
3	L	M	N	O	P
4	Q	R	S	T	U
5	V	W	X	Y	Z

б)

Рис. 14.2. Примеры шифров: а) алфавит Цезаря со сдвигом 3; б) квадрат Полибиуса

Исходный текст:	N	O	W	I	S	T	H	E	T	I	M	E
Шифрованный текст:	Q	R	Z	L	V	W	K	H	W	L	P	H

Дешифрующий ключ — это просто число сдвигов алфавита; с выбором нового ключа код изменяется. Еще одна классическая система шифрования, изображенная на рис. 14.2, *б*, называется квадратом Полибиуса (Polybius square). Вначале объединяются буквы I и J и трактуются как один символ (в дешифрованном сообщении значение этой "двойной буквы" легко определяется из контекста). Получившиеся 25 символов алфавита размещаются в таблицу размером 5 × 5. Шифрование любой буквы производится с помощью выбора соответствующей пары чисел — строки и столбца (или столбца и строки). Ниже приведен пример шифрования того же сообщения "now is the time" с помощью квадрата Полибиуса.

Исходный текст:	N	O	W	I	S	T	H	E	T	I	M	E
Шифрованный текст:	33	43	25	42	34	44	32	51	44	42	23	51

Код изменяется путем перестановки букв в таблице 5 × 5.

Прогрессивный ключ Тритемиуса, который изображен на рис. 14.3, является примером *полиалфавитного шифра*. Строка, обозначенная как сдвиг 0, совпадает с обычным порядком букв в алфавите. Буквы в следующей строке сдвинуты на один символ влево с циклическим сдвигом оставшихся позиций. Каждая последующая строка получается с помощью

такого же сдвига алфавита на один символ влево относительно предыдущей строки. Это продолжается до тех пор, пока в результате циклических сдвигов алфавит не будет смещен на все возможные позиции. Один из методов использования такого алфавита заключается в выборе первого символа шифрованного сообщения из строки, полученной при сдвиге на 1 символ, второго символа — из строки, полученной при сдвиге на 2 символа, и т.д. Ниже приведен пример сообщения, зашифрованного подобным образом.

Открытый текст:	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
Сдвиг: 0	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	O	R	S	T	U	V	W	X	Y	Z	A
2	C	D	E	F	G	H	I	J	K	L	M	N	O	P	O	R	S	T	U	V	W	X	Y	Z	A	B
3	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
4	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
5	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
6	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
7	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
8	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
9	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
10	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
11	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
12	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
13	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
14	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
15	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
16	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
17	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
18	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
19	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
20	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
21	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
22	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
23	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
24	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
25	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

Рис. 14.3. Прогрессивный ключ Тритемиуса

Исходный текст:	N	O	W	I	S	T	H	E	T	I	M	E
Шифрованный текст:	O	Q	Z	M	X	Z	O	M	C	S	X	Q

Существует несколько интересных способов использования прогрессивного шифра Тритемиуса. В одном из них, называемом *методом ключа Вигнера* (Vigener key method), применяется ключевое слово (keyword). Этот ключ диктует выбор строк для шифрования и дешифрования каждого последующего символа в сообщении. Предположим, что в качестве ключа выбрано слово "TYPE"; тогда сообщение, зашифрованное с применением метода Вигнера, выглядит следующим образом.

Ключ:	T	Y	P	E	T	Y	P	E	T	Y	P	E
Исходный текст:	N	O	W	I	S	T	H	E	T	I	M	E
Шифрованный текст:	G	M	L	M	L	R	W	I	M	G	B	I

Здесь первая буква ключа (T) указывает, что в качестве строки для шифрования первой буквы открытого текста выбирается строка, начинающаяся с T (сдвиг 19). Следующей выбирается строка, начинающаяся с Y (сдвиг 24), и т. д. Разновидностью этого метода является так называемый *метод автоматического (явного) ключа Вигнера* (Vigener auto (plain) key method), когда в качестве *образующего ключа* используется единственная буква или слово. Этот ключ дает начальную строку или строки для шифрования первого или нескольких первых символов открытого текста аналогично предыдущему примеру. Затем в качестве ключа для выбора шифрующей строки используются *символы исходного текста*. В приведенном ниже примере в качестве образующего ключа использована буква "F".

Ключ:	F	N	O	W	I	S	T	H	E	T	I	M
Исходный текст:	N	O	W	I	S	T	H	E	T	I	M	E
Шифрованный текст:	S	B	K	E	A	L	A	L	X	B	U	Q

Метод автоматического ключа показывает, что в процесс шифрования может быть введена обратная связь. При использовании обратной связи выбор шифрованного текста определяется содержанием сообщения.

Последняя разновидность метода Вигнера — это *метод автоматического (шифрованного) ключа Вигнера* (Vigenere auto (cipher) key method), подобный простому методу ключа; в нем также используются образующий ключ и обратная связь. Отличие состоит в том, что после шифрования с помощью образующего ключа, каждый последующий символ ключа в последовательности берется не из символа исходного текста, а из *символа шифрованного текста*. Ниже приведен пример, который должен помочь понять принцип работы данного метода; как и ранее, в качестве начального ключа используется буква "F".

Ключ:	F	S	G	C	K	C	V	C	G	Z	H	T
Исходный текст:	N	O	W	I	S	T	H	E	T	I	M	E
Шифрованный текст:	S	G	C	K	C	V	C	G	Z	H	T	X

Хотя каждый символ ключа может быть найден из предшествующего ему символа шифрованного текста, функционально он зависит от *всех* предшествующих символов в сообщении и плюс основного ключа. Таким образом, имеется эффект рассеивания статистических свойств исходного текста вдоль шифрованного текста, что делает статистический анализ очень сложным для криптоаналитика. Слабым звеном описанного здесь примера шифрования с использованием ключа является то, что шифрованный текст содержит знаки ключа, которые будут публично выставлены через общедоступный канал "на всеобщее обозрение". Для того чтобы предотвратить такое публичное разоблачение, можно использовать вариации этого метода [3]. По нынешним стандартам схема шифрования Вигнера не является очень защищенной; основным вкладом Вигнера было открытие того, что неповторяющиеся ключевые последовательности можно создавать с использованием самих сообщений или функций от сообщений.

14.2. Секретность системы шифрования

14.2.1. Совершенная секретность

Рассмотрим систему шифрования с конечной областью сообщений $\{M\} = M_0, M_1, \ldots, M_{N-1}$ и конечной областью шифрованных текстов $\{C\} = C_0, C_1, \ldots, C_{U-1}$. Для любого M_i

априорная вероятность передачи сообщения M_i равна $P(M_i)$. Апостериорная вероятность принятия сообщения C_j при переданном сообщении M_i равна $P(M_i|C_j)$. Говорят, что *система шифрования имеет совершенную секретность*, если для любого сообщения M_i и любого шифрованного текста C_j апостериорная вероятность равна априорной.

$$P(M_i|C_j) = P(M_i) \tag{14.2}$$

Таким образом, для системы с совершенной секретностью характерно следующее: если криптоаналитик перехватил сообщение C_j, то дальнейшей информации, которая бы облегчила ему дешифровку сообщения, он не получит. Необходимое и достаточное условие совершенной секретности: для любого M_i и C_j

$$P(C_j|M_i) = P(C_j) \tag{14.3}$$

На рис. 14.4 изображен пример схемы совершенной секретности. В этом примере $\{M\} = M_0, M_1, M_2, M_3$; $\{C\} = C_0, C_1, C_2, C_3$; $\{K\} = K_0, K_1, K_2, K_3$; $N = U = 4$, $P(M_i) = P(C_j) = \frac{1}{4}$. Преобразование сообщения в шифрованный текст выполняется следующим образом.

$$C_s = T_{K_j}(M_i)$$
$$s = (i + j) \text{ по модулю } N \tag{14.4}$$

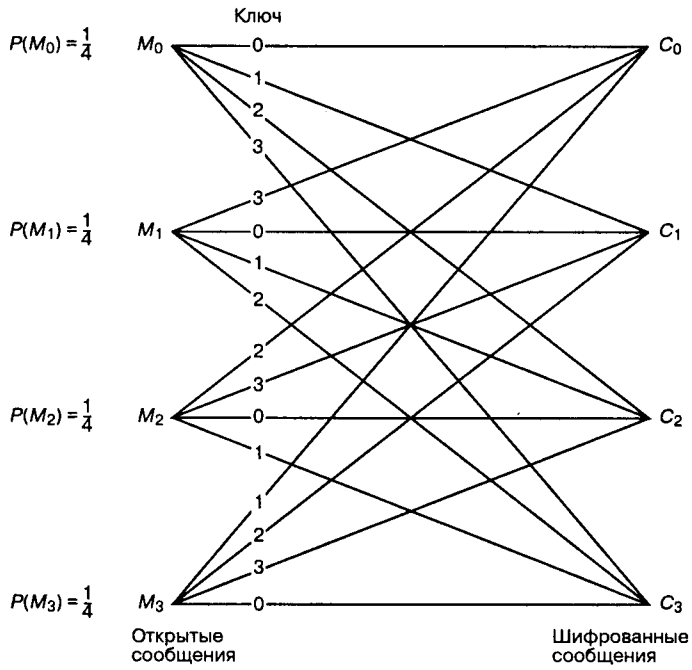


Рис. 14.4. Совершенная секретность

Здесь T_{K_j} определяет преобразование с помощью ключа K_j, а "x по модулю y" — это остаток от деления x на y. Таким образом, $s = 0, 1, 2, 3$. Криптоаналитик, перехвативший одно из шифрованных сообщений $C_s = C_0$, C_1, C_2 или C_3, не сможет определить, какой из четырех ключей использовался и, следовательно, какое из сообщений M_0, M_1, M_2 или M_3 является верным. Если в системе шифрования число сообщений, число ключей и число шиф-

рованных сообщений равны между собой, то система имеет совершенную секретность тогда и только тогда, когда выполняются следующие два условия.

1. Существует только один ключ, преобразующий каждое сообщение в каждый шифрованный текст.
2. Все ключи равновероятны.

Если эти условия не выполняются, то будет существовать некоторое сообщение M_i, при котором для данного C_j не существует ключа, который мог бы дешифровать C_j в M_i. Отсюда следует, что для некоторых i и j $P(M_i|C_j) = 0$. В этом случае криптоаналитик может исключить из рассмотрения определенные нешифрованные сообщения, упростив, таким образом, задачу. Вообще, совершенная секретность является очень желательным свойством, поскольку это означает, что система шифрования безусловно защищена. Должно быть очевидно, что в системах, передающих большое количество сообщений, для достижения совершенной секретности требуется распределить большое количество ключей, а это, в свою очередь, может привести к значительным практическим затруднениям, что делает такие системы нереализуемыми. В системе с совершенной секретностью число возможных ключей так же велико, как и число возможных сообщений, поэтому, если мы разрешим передавать сообщения неограниченной длины, совершенная секретность потребует бесконечного количества ключей.

Пример 14.1. Взлом системы шифрования, если область ключей меньше области сообщений

Рассмотрим шифрованный текст, состоящий из 29 символов.

G R O B O K B O D R O R O B Y O C Y P I O C D O B I O K B

Данный текст был получен с помощью шифра Цезаря (см. раздел 14.1.4); каждая буква получена сдвигом на K символов, где $1 \le K \le 25$. Покажите, как криптоаналитик может взломать этот код.

Решение

Поскольку количество возможных ключей (их 25) меньше количества возможных осмысленных сообщений из 29 символов (их огромное множество), совершенная секретность не может быть достигнута. В исходном полиалфавитном шифре, показанном на рис. 14.3, символ открытого текста заменяется буквой некоторой строки, причем номер строки постоянно возрастает. Следовательно, в процессе анализа шифрованного текста мы обращаем процесс: теперь буквы шифрованного текста заменяются буквами строк, причем номер строки постоянно уменьшается. Путем перебора всех ключей от 1 до 25 (рис. 14.5) можно легко рассмотреть все возможности. В результате, этот процесс приводит к единственному ключу ($K = 10$), дающему осмысленное сообщение (пробелы были добавлены вручную): WHERE ARE THE HEROES OF YESTERYEAR.

Пример 14.2. Совершенная секретность

Для создания шифра, имеющего совершенную секретность, можно несколько модифицировать область ключей, описанную в примере 14.1. В этой новой системе шифрования каждый символ сообщения шифруется с использованием *случайно выбранного* ключевого значения. Теперь ключ K задается последовательностью k_1, k_2, ..., k_{29}, где каждое k_i — это случайно выбранное целое число из интервала (1, 25), определяющее сдвиг, используемый для i-го символа. Таким образом, всего существует $(25)^{29}$ различных ключевых последовательностей. Значит, шифрованный текст из 29 символов, приведенный в примере 14.1, может соответствовать *любому* осмысленному сообщению из 29 символов. Например, шифрованный текст мог соответствовать следующему открытому тексту (пробелы были добавлены вручную).

ENGLISH AND FRENCH ARE SPOKEN HERE

Данный текст получен с помощью ключа 2, 4, 8, 16, 6, 18, 20, ... Стоит отметить, что большинство возможных наборов из 29 символов можно исключить, поскольку они не являются осмысленными сообщениями. Совершенная секретность данного кода — результат того, что перехват шифрованного текста не дает никакой дополнительной информации об открытом сообщении.

Ключ	Текст																											
0	G	R	O	B	O	K	B	O	D	R	O	R	O	B	Y	O	C	Y	P	I	O	C	D	O	B	I	O	K	B
1	F	Q	N	A	N	J	A	N	C	Q	N	Q	N	A	X	N	B	X	O	H	N	B	C	N	A	H	N	J	A
2	E	P	M	Z	M	I	Z	M	B	P	M	P	M	Z	W	M	A	W	N	G	M	A	B	M	Z	G	M	I	Z
3	D	O	L	Y	L	H	Y	L	A	O	L	O	L	Y	V	L	Z	V	M	F	L	Z	A	L	Y	F	L	H	Y
4	C	N	K	X	K	G	X	K	Z	N	K	N	K	X	U	K	Y	U	L	E	K	Y	Z	K	X	E	K	G	X
5	B	M	J	W	J	F	W	J	Y	M	J	M	J	W	T	J	X	T	K	D	J	X	Y	J	W	D	J	F	W
6	A	L	I	V	I	E	V	I	X	L	I	L	I	V	S	I	W	S	J	C	I	W	X	I	V	C	I	E	V
7	Z	K	H	U	H	D	U	H	W	K	H	K	H	U	R	H	V	R	I	B	H	V	W	H	U	B	H	D	U
8	Y	J	G	T	G	C	T	G	V	J	G	J	G	T	Q	G	U	Q	H	A	G	U	V	G	T	A	G	C	T
9	X	I	F	S	F	B	S	F	U	I	F	I	F	S	P	F	T	P	G	Z	F	T	U	F	S	Z	F	B	S
10	W	H	E	R	E	A	R	E	T	H	E	H	E	R	O	E	S	O	F	Y	E	S	T	E	R	Y	E	A	R
11	V	G	D	Q	D	Z	Q	D	S	G	D	G	D	Q	N	D	R	N	E	X	D	R	S	D	Q	X	D	Z	Q
12	U	F	C	P	C	Y	P	C	R	F	C	F	C	P	M	C	Q	M	D	W	C	Q	R	C	P	W	C	Y	P
13	T	E	B	O	B	X	O	B	Q	E	B	E	B	O	L	B	P	L	C	V	B	P	Q	B	O	V	B	X	O
14	S	D	A	N	A	W	N	A	P	D	A	D	A	N	K	A	O	K	B	U	A	O	P	A	N	U	A	W	N
15	R	C	Z	M	Z	V	M	Z	O	C	Z	C	Z	M	J	Z	N	J	A	T	Z	N	O	Z	M	T	Z	V	M
16	Q	B	Y	L	Y	U	L	Y	N	B	Y	B	Y	L	I	Y	M	I	Z	S	Y	M	N	Y	L	S	Y	U	L
17	P	A	X	K	X	T	K	X	M	A	X	A	X	K	H	X	L	H	Y	R	X	L	M	X	K	R	X	T	K
18	O	Z	W	J	W	S	J	W	L	Z	W	Z	W	J	G	W	K	G	X	Q	W	K	L	W	J	Q	W	S	J
19	N	Y	V	I	V	R	I	V	K	Y	V	Y	V	I	F	V	J	F	W	P	V	J	K	V	I	P	V	R	I
20	M	X	U	H	U	Q	H	U	J	X	U	X	U	H	E	U	I	E	V	O	U	I	J	U	H	O	U	Q	H
21	L	W	T	G	T	P	G	T	I	W	T	W	T	G	D	T	H	D	U	N	T	H	I	T	G	N	T	P	G
22	K	V	S	F	S	O	F	S	H	V	S	V	S	F	C	S	G	C	T	M	S	G	H	S	F	M	S	O	F
23	J	U	R	E	R	N	E	R	G	U	R	U	R	E	B	R	F	B	S	L	R	F	G	R	E	L	R	N	E
24	I	T	Q	D	Q	M	D	Q	F	T	Q	T	Q	D	A	Q	E	A	R	K	Q	E	F	Q	D	K	Q	M	D
25	H	S	P	C	P	L	C	P	E	S	P	S	P	C	Z	P	D	Z	Q	J	P	D	E	P	C	J	P	L	C

Рис. 14.5. Пример взлома системы шифрования, если область ключей меньше области сообщений

14.2.2. Энтропия и неопределенность

Как обсуждалось в главе 9, объем информации в сообщении связан с вероятностью появления сообщения. Сообщения вероятности 0 либо 1 не содержат информации, поскольку можно с известной долей определенности предсказать их появление. Чем больше неопределенности существует в предсказании появления сообщения, тем больше оно содержит информации. Следовательно, если все сообщения множества равновероятны, мы *не* можем быть уверенными в возможности предсказания появления конкретного сообщения, и неопределенность информационного содержания сообщения является максимальной.

Энтропия $H(K)$ определяется как средний объем информации на сообщение. Она может рассматриваться как мера того, насколько в выбор сообщения X вовлечен *случай*. Она записывается как следующее суммирование по всем возможным сообщениям.

$$H(X) = -\sum_{X} P(X) \log_2 P(X) = \sum_{X} P(X) \log_2 \frac{1}{P(X)} \qquad (14.5)$$

Если, как выше, логарифм берется по основанию 2, $H(X)$ представляет собой *математическое ожидание числа битов* в *оптимально закодированном* сообщении X. Это все еще не та мера, которую хотел бы иметь криптоаналитик. Им будут перехвачены некоторые шифрованные тексты, и он захочет узнать, насколько достоверно он может предсказать сообщение (или ключ) при условии, что был отправлен именно этот конкретный шифрованный текст. Неопределенность, определенная как условная энтропия X при данном Y, является для криптоаналитика более полезной мерой при попытке взлома шифра. Она задается с помощью следующей формулы.

$$H(X \mid Y) = -\sum_{X,Y} P(X,Y)\log_2 P(X,Y) = \\ = \sum_{Y} P(Y) \sum_{X} P(X \mid Y)\log_2 \frac{1}{P(X \mid Y)} \tag{14.6}$$

Неопределенность может рассматриваться как неуверенность в том, что отправлено было сообщение X, при условии получения Y. Желательным для криптоаналитика является приближение $H(X|Y)$ к нулю при увеличении объема перехваченного шифрованного текста Y.

Пример 14.3. Энтропия и неопределенность

Рассмотрим выборочное множество сообщений, состоящее из восьми равновероятных сообщений $\{X\} = X_1, X_2, ..., X_8$.

а) Найдите энтропию, связанную с сообщением из множества $\{X\}$.

б) Дано другое множество равновероятных сообщений $\{Y\} = Y_1, Y_2$. Пусть появление каждого сообщения Y сужает возможный выбор X следующим образом.

При наличии Y_1 возможны только X_1, X_2, X_3 или X_4

При наличии Y_2 возможны только X_5, X_6, X_7 или X_8

Найдите неопределенность сообщения X, обусловленную сообщением Y.

Решение

а) $P(X) = \frac{1}{8}$

$H(X) = 8\left[\frac{1}{8}\log_2 8\right] = 3$ бит/сообщение

б) $P(Y) = \frac{1}{2}$. Для каждого Y, $P(X|Y) = \frac{1}{4}$ для четырех сообщений из множества $\{X\}$ и $P(X|Y) = 0$ для оставшихся четырех. Используя уравнение (14.6), получим следующее.

$$H(X|Y) = 2\left[\left(\tfrac{1}{2}\right)4\left(\tfrac{1}{4}\log_2 4\right)\right] = 2 \text{ бит/сообщение}$$

Видно, что знание Y сводит неопределенность X с 3 бит/сообщение до 2 бит/сообщение.

14.2.3. Интенсивность и избыточность языка

Истинная интенсивность языка определяется как среднее число *информационных битов*, содержащихся в каждом символе, и для сообщения длиной N выражается следующим образом.

$$r = \frac{H(X)}{N} \tag{14.7}$$

Здесь $H(X)$ — энтропия сообщения, или число битов в *оптимально закодированном* сообщении. Для письменного английского языка при больших N оценки r дают значения

между 1,0 и 1,5 бит/символ [4]. *Абсолютная интенсивность* или максимальная энтропия языка определяется как максимальное число информационных битов, содержащихся в каждом символе, в предположении, что все возможные последовательности символов одинаково вероятны. Абсолютная интенсивность задается следующим образом.

$$r' = \log_2 L \tag{14.8}$$

Здесь L — число знаков в языке. Для английского алфавита $r' = \log_2 26 = 4{,}7$ бит/символ. Истинная интенсивность английского языка, конечно, гораздо меньше его абсолютной интенсивности, поскольку, как и большинство языков, английский очень избыточен и структурирован.

Избыточность языка определяется через его истинную и абсолютную интенсивности.

$$D = r' - r \tag{14.9}$$

Для английского языка, где $r' = 4{,}7$ бит/символ и $r = 1{,}5$ бит/символ, $D = 3{,}2$, а отношение $D/r' = 0{,}68$ — это мера избыточности языка.

14.2.4. Расстояние единственности и идеальная секретность

Ранее утверждалось, что если допускаются сообщения неограниченной длины, то совершенная секретность требует бесконечного количества ключей. При конечном размере ключа его неопределенность $H(K|C)$ обычно приближается к нулю, откуда следует, что ключ может быть определен единственным образом, а система шифрования может быть взломана. *Расстояние единственности* (unicity distance) определяется как наименьшая длина шифрованного текста N, при которой неопределенность ключа $H(K|C)$ близка к нулю. Следовательно, расстояние единственности — это количество шифрованного текста, необходимое для того, чтобы однозначно определить ключ и таким образом взломать систему шифрования. Шеннон (Shennon) [5] описал систему с *идеальной секретностью* как систему, в которой $H(K|C)$ не стремится к нулю, если количество шифрованного текста стремится к бесконечности. Иными словами, ключ не может быть определен, независимо от того, сколько шифрованного текста перехвачено. Термин "идеальная секретность" описывает систему, которая не достигает совершенной секретности, но, тем не менее, не поддается взлому (безусловно защищенная система), поскольку она не дает достаточно информации для определения ключа.

Большинство систем шифрования слишком сложны для определения вероятностей, необходимых для вычисления расстояния единственности. В то же время расстояние единственности иногда можно аппроксимировать, что было показано Шенноном [5] и Хэллманом (Hellman) [6]. Следуя Хэллману, предположим, что каждый открытый текст и шифрованное сообщение получены с помощью конечного алфавита из L символов. Таким образом, всего существует $2^{r'N}$ возможных сообщений длиной N, где r' — абсолютная интенсивность языка. Всю область сообщений можно разделить на два класса — осмысленные сообщения M_1 и бессмысленные сообщения M_2. Тогда имеем

$$\text{число осмысленных сообщений } 2^{r'N} \tag{14.10}$$

$$\text{число бессмысленных сообщений } 2^{r'N} - 2^{rN}, \tag{14.11}$$

где r — истинная интенсивность языка, а априорные вероятности классов сообщений описываются следующими выражениями.

$$P(M_1) = \frac{1}{2^{rN}} = 2^{-rN} \quad M_1 \text{ — осмысленное} \tag{14.12}$$

$$P(M_2) = 0 \quad M_2 \text{ — бессмысленное} \tag{14.13}$$

Предположим, что существует $2^{H(K)}$ возможных ключа (размер алфавита ключей), где $H(K)$ — энтропия ключа (бит в ключе). Предположим, что все ключи равновероятны.

$$P(K) = \frac{1}{2^{H(K)}} = 2^{-H(K)} \tag{14.14}$$

Определение расстояния единственности основано на модели *случайного шифра*, которая утверждает, что для каждого ключа K и шифрованного текста C операция дешифрования $D_K(C)$ дает независимую случайную переменную, распределенную по всем возможным 2^{rN} сообщениям (как осмысленным, так и бессмысленным). Следовательно, для данных K и C операция $D_K(C)$ может с равной вероятностью давать любое из открытых сообщений.

При данном шифровании, описываемом как $C_i = E_{K_i}(M_i)$, *неверное решение* F возникает всегда, когда шифрование с помощью другого ключа K_j может давать C_i из того же сообщения M_i или из некоторого другого сообщения M_j.

$$C_i = E_{K_i}(M_i) = E_{K_j}(M_i) = E_{K_j}(M_j) \tag{14.15}$$

Криптоаналитик, перехвативший C_i, не сможет выбрать верный ключ и, следовательно, не сможет взломать систему шифрования. Мы не рассматриваем операции дешифрования, которые дают *бессмысленные* сообщения, так как они могут легко отбрасываться.

Для каждого верного решения конкретного шифрованного текста существует $2^{H(K)-1}$ неверных ключа, каждый из которых имеет ту же вероятность $P(F)$ получения неверного решения. Так как все осмысленные открытые сообщения предполагаются равновероятными, вероятность неверного решения равна вероятности получения осмысленного сообщения.

$$P(F) = \frac{2^{rN}}{2^{r'N}} = 2^{(r-r')N} = 2^{-DN} \tag{14.16}$$

Здесь $D = r' - r$ — избыточность языка. Тогда ожидаемое число неверных решений $\overline{F}$ равно следующему.

$$\overline{F} = \left[2^{H(K)} - 1\right]P(F) = \left[2^{H(K)} - 1\right]2^{-DN} \approx 2^{H(K)-DN} \tag{14.17}$$

Поскольку $\overline{F}$ быстро убывает с увеличением N, то

$$\log_2 \overline{F} = H(K) - DN = 0 \tag{14.18}$$

является точкой, где число неверных решений достаточно мало; так что шифр может быть взломан. Следовательно, получаемое расстояние единственности описывается следующим выражением.

$$N = \frac{H(K)}{D} \tag{14.19}$$

Из уравнения (14.17) следует, что если $H(K)$ значительно больше DN, то будет множество осмысленных расшифровок, и, следовательно, существует малая вероятность вы-

деления криптоаналитиком верного сообщения из возможных осмысленных. Приблизительно, DN — это число уравнений для ключа, а $H(K)$ — число неизвестных. Если число уравнений меньше числа неизвестных битов ключа, единственное решение невозможно; говорят, что система на поддается взлому. Если число уравнений больше числа неизвестных, возможно единственное решение, и система не может больше считаться не поддающейся взлому (хотя она все еще может относиться к защищенным по вычислениям).

Стоит отметить, что доминирование бессмысленных дешифровок позволяет взламывать криптограммы. Уравнение (14.19) показывает значение использования *сжатия данных* до шифрования. Сжатие данных устраняет избыточность языка, таким образом увеличивая расстояние единственности. Совершенное сжатие данных даст $D = 0$ и $N = \infty$ для любого размера ключа.

Пример 14.4. Расстояние единственности

Вычислите расстояние единственности для системы шифрования, использующей письменный английский язык, ключ которой задается последовательностью k_1, k_2, ..., k_{29}, где каждое k_i — случайное целое из интервала (1, 25), которое определяет номер сдвига (рис. 14.3) для i-го символа. Предположим, что все возможные ключевые последовательности равновероятны.

Решение

Существует $(25)^{29}$ возможных равновероятных ключевых последовательностей. Следовательно, используя равенства (14.5), (14.8) и (14.19), получаем следующее.

$$\text{Энтропия ключа: } H(K) = \log_2 (25)^{29} = 135 \text{ бит}$$

$$\text{Абсолютная интенсивность английского языка: } r' = \log_2 26 = 4{,}7 \text{ бит/символ}$$

$$\text{Предполагаемая истинная интенсивность английского языка: } r = 1{,}5 \text{ бит/символ}$$

$$\text{Избыточность: } D = r' - r = 3{,}2 \text{ бит/символ}$$

$$N = \frac{H(K)}{D} = \frac{135}{3.2} \approx 43 \text{ символа}$$

В примере 14.2 совершенная секретность сообщения из 29 символов иллюстрировалась с использованием того же типа ключевой последовательности, что и в данном примере, где показано, что если имеющийся шифрованный текст состоит из 43 символов (откуда следует, что некоторая часть ключевой последовательности должна использоваться дважды), то возможно единственное решение. В то же время не определена вычислительная сложность отыскания решения. Даже если оценить теоретическое количество шифрованного текста, необходимое для взлома шифра, практически это может оказаться невозможным.

14.3. Практическая защищенность

Для последовательностей шифрованного текста, размер которых больше расстояния единственности, любая система уравнений (определяющая ключ) может быть решена путем простого перебора всех возможных ключей, пока не будет получено единственное решение. Однако это совершенно непрактично, за исключением применения очень короткого ключа. Например, для ключа, полученного путем перестановки английского алфавита, существует $26! \approx 4 \times 10^{26}$ возможных перестановок (в криптографическом смысле это считается малым). Будем считать, что в результате изнурительных поисков мы нашли правильный ключ, перебрав при-

близительно половину возможных комбинаций. Если допустить, что каждая проверка потребует для вычисления 1 мкс, то полное время поиска превысит 10^{12} лет. Следовательно, если криптоаналитик хочет иметь некоторую надежду на успех, то о "лобовых" методах перебора следует забыть и применять какую-ту иную технологию (например, статистический анализ).

14.3.1. Смешение и диффузия

При расшифровке многих систем шифрования может применяться статистический анализ, использующий частоту появления отдельных символов и их комбинаций. Шеннон [5] предложил две концепции шифрования, усложняющие задачу криптоаналитика. Он назвал эти преобразования "смешение" (confusion) и "диффузия" (diffusion). *Смешение* — это подстановки, которые делают взаимосвязь между ключом и шифрованным текстом как можно более сложной. Это усложняет применение статистического анализа, сужающего поиск практического подмножества области ключей. В результате смешения дешифрование даже очень короткой последовательности шифрованного текста требует большого числа ключей. *Диффузия* — это преобразования, сглаживающие статистические различия между символами и их комбинациями. Примером диффузии 26-буквенного алфавита является преобразование последовательности сообщений $M = M_0, M_1, \ldots$ в новую последовательность сообщений $Y = Y_0, Y_1, \ldots$ с помощью следующего соотношения.

$$Y_n = \sum_{i=0}^{s-1} M_{n+i} \text{ по модулю } 26 \qquad (14.20)$$

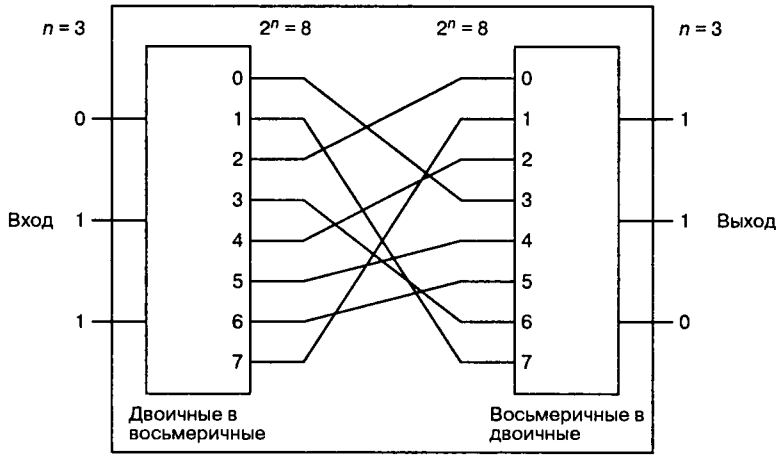
Здесь каждый символ в последовательности рассматривается как число по модулю 26, s — некоторое выбранное целое число и $n = 1, 2, \ldots$. Новое сообщение Y будет иметь ту же избыточность, что и исходное сообщение M, но частота появления всех букв в Y будет более равномерной, чем в M. В результате, чтобы статистический анализ принес криптоаналитику какую-либо пользу, ему необходимо перехватить большую последовательность шифрованного текста.

14.3.2. Подстановка

Технология шифрования с помощью подстановки, например использование шифра Цезаря и прогрессивного ключа шифрования.Тритемиуса, широко используется в головоломках. Такие простые подстановочные шифры дают малую защищенность. Чтобы к подстановочной технологии можно было применить концепцию *смешения*, требуется более сложное соотношение. На рис. 14.6 изображен пример создания большей подстановочной сложности с помощью использования нелинейного преобразования. В общем случае n входных битов сначала представляются как один из 2^n различных символов (на приведенном рисунке $n = 2$). Затем множество из 2^nсимволов перемешивается так, чтобы каждый символ заменялся другим символом множества. После этого символ снова превращается в n-битовый.

Можно легко показать, что существует $(2^n)!$ различные подстановки или связанные с ними возможные модели. Задача криптоаналитика становится вычислительно невозможной для больших n. Пусть $n = 128$, тогда $2^n = 10^{38}$ и $(2^n)!$ пред-

ставляет собой астрономическое число. Видим, что для $n = 128$ это преобразование с помощью блока подстановки (substitution block, S-блок) является сложным (запутывающим). Впрочем, хотя S-блок с $n = 128$ можно считать идеальным, его реализация является невозможной, поскольку она потребует блока с $2n = 10^{38}$ контактами.



Вход	000	001	010	011	100	101	110	111
Выход	011	111	000	110	010	100	101	001

Рис. 14.6. Блок подстановки

Чтобы убедиться, что S-блок, приведенный на рис. 14.6, представляет собой *нелинейное преобразование*, достаточно использовать теорему о суперпозиции, которая формулируется ниже. Предположим, что

$$C = Ta + Tb$$
$$C' = T(a + b), \tag{14.21}$$

где a и b — входные элементы, C и C' — выходные элементы, а T — преобразование. Тогда

Если T линейно, $C = C'$ для всех входных элементов.

Если T нелинейно, $C \neq C'$.

Предположим, $a = 001$ и $b = 010$; тогда, используя преобразование T, показанное на рис. 14.6, получим следующее.

$$C = T(001) \oplus T(010) = 111 \oplus 000 = 111$$
$$C' = T(001 \oplus 010) = T(011) = 110$$

Здесь символ $\oplus$ обозначает сложение по модулю 2. Поскольку $C \neq C'$, S-блок является нелинейным.

14.3.3. Перестановка

При перестановке (транспозиции) буквы исходного открытого текста в сообщении не заменяются другими буквами алфавита, как в классических шифрах, а

просто переставляются. Например, слово "THINK" после перестановки может выглядеть как шифрованный текст HKTNI. На рис. 14.7 приведен пример бинарной перестановки данных (линейная операция). Видно, что входные данные просто перемешиваются или переставляются. Преобразование выполняется с помощью блока перестановки (permutation block, *P*-блок). Технология, используемая сама по себе, имеет один основной недостаток: она уязвима по отношению к обманным сообщениям. Обманное сообщение изображено на рис. 14.7. Подача на вход единственной 1 (при остальных 0) позволяет обнаружить одну из внутренних связей. Если криптоаналитику необходимо выполнить криптоанализ такой системы с помощью атаки открытого текста, он отправит последовательность таких обманных сообщений, при каждой передаче смещая единственную 1 на одну позицию. Таким образом, обнаруживаются все связи входа и выхода. Данный пример показывает, почему защищенность системы не должна зависеть от ее архитектуры.

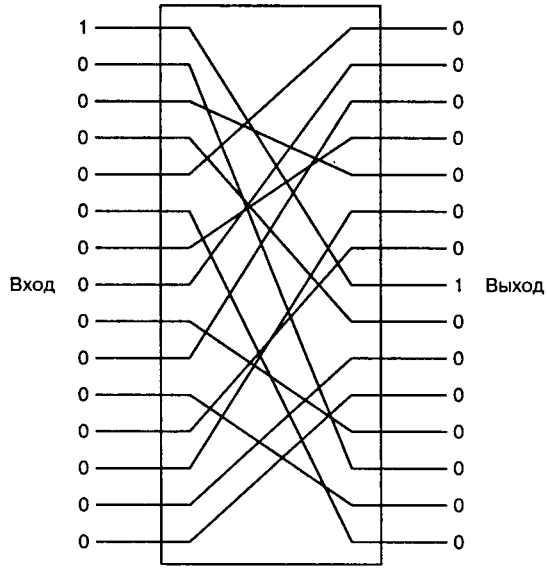


Рис. 14.7. Блок перестановки

14.3.4. Продукционный шифр

Для преобразований, включающих значительное число n-символьных сообщений, желательным является применение обеих описанных выше схем (S-блока и P-блока). Шеннон [5] предложил использовать *продукционный шифр*, или комбинацию преобразований S- и P-блоков, которые вместе могут дать более мощную систему шифрования, чем каждый из них в отдельности. Этот подход, выборочно использующий преобразования замещения и перестановки, был использован IBM в системе LUCIFER [7, 8] и стал основой национального стандарта шифрования данных (Data Encryption Standard — DES) [9]. На рис. 14.8 изображены такие комбинации P- и S-блоков. Дешифрование выполняется обратным прогоном данных, при котором используются преобразования, обратные к преобразованию каждого S-блока. Систему, изображенную на рис. 14.8, реализовать довольно труд-

но, поскольку все S-блоки являются различными, случайно генерируемый ключ неприменим и система не дает возможности повторить одну и ту же последовательность операций. Поэтому в системе LUCIFER [8] использовались два различных типа S-блоков, S_1 и S_0, которые могли быть общедоступными. Пример такой системы изображен на рис. 14.9. Входные данные преобразуются с помощью последовательности S- и P-блоков, определяемой ключом. В приведенном примере ключ размером 25 бит определяет, какой из двух блоков (S_1 или S_0) следует выбрать на каждой из 25 позиций схемы. Таким образом, подробности аппарата шифрования могут быть открыты, поскольку защищенность системы обеспечивается ключом.

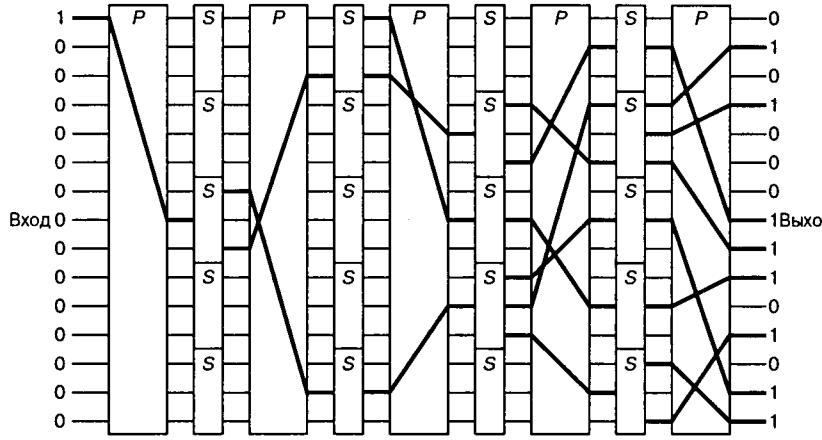


Рис. 14.8. Продукционная система шифрования

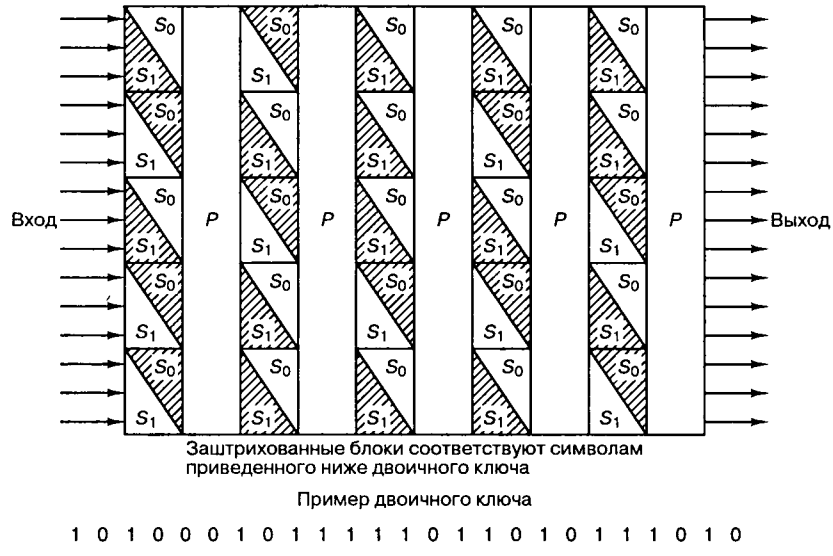


Рис. 14.9. Индивидуальные возможности, определяемые ключом

Итеративная структура продукционной системы шифрования (рис. 14.9) является типичной для большинства реальных блочных шифров. Сообщения делятся на последовательные блоки по n бит, каждый из которых шифруется одним и тем же ключом. n-битовый блок представляет один из 2^n различных символов, допускающих $(2^n)!$ различные схемы подстановки. Следовательно, чтобы реализация схемы была разумной, подстановочная часть шифрования выполняется параллельно на небольших сегментах блока. Пример подобной схемы рассмотрен в следующем разделе.

14.3.5. Стандарт шифрования данных

В 1977 году Национальное бюро стандартов США (National Bureau of Standards) приняло модифицированную систему LUCIFER в качестве Национального стандарта шифрования данных (Data Encryption Standard — DES) [9]. Как показано на рис. 14.10, с точки зрения системы ввода-вывода DES может считаться блочной системой шифрования с алфавитом в 2^{64} символа. Входной блок из 64 бит, который является в этом алфавите символом открытого текста, заменяется новым символом шифрованного текста. На рис. 14.11 в виде блочной диаграммы показаны функции системы. Алгоритм шифрования начинается с начальной перестановки 64 бит открытого текста, описанной в таблице начальной перестановки (табл. 14.1). Таблица начальной перестановки читается слева направо и сверху вниз, так что после перестановки биты $x_1, x_2, \ldots, x_{64}$ превращаются в $x_{58}, x_{50}, \ldots, x_7$. После этой начальной перестановки начинается основная часть алгоритма шифрования, состоящая из 16 итераций, которые используют стандартный блок, показанный на рис. 14.12. Для преобразования 64 бит входных данных в 64 бит выходных, определенных как 32 бит левой половины и 32 бит правой, стандартный блок использует 48 бит ключа. Выход каждого стандартного блока становится входом следующего стандартного блока. Входные 32 бит правой половины (R_{i-1}) без изменений подаются на выход и становятся 32 бит левой половины (L_i). Эти R_{i-1} бит с помощью таблицы расширения (табл. 14.2) также *расширяются* и преобразуются в 48 бит, после чего суммируются по модулю 2 с 48 бит ключа. Как и в случае таблицы начальной перестановки, таблица расширения читается слева направо и сверху вниз.

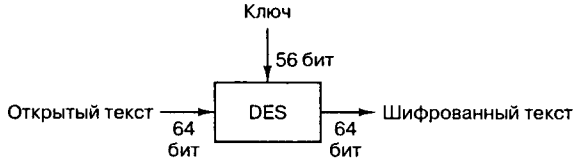


Рис. 14.10. Стандарт шифрования данных (DES) в виде блочной системы шифрования

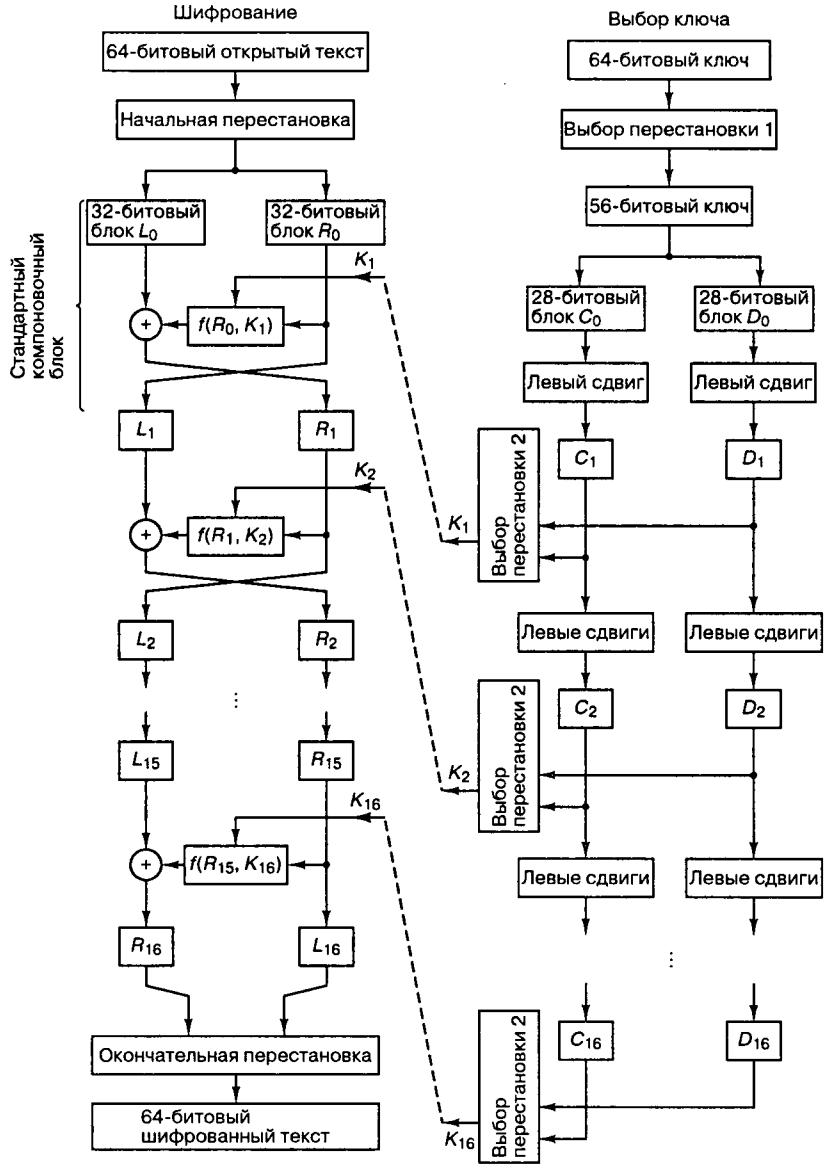


Рис. 14.11. Стандарт шифрования данных

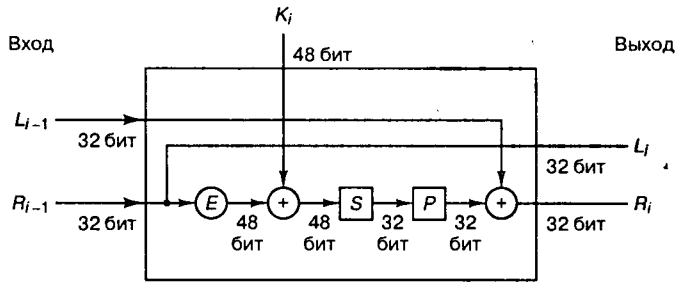


Рис. 14.12. Стандартный компоновочный блок

Данная таблица отображает биты

Таблица 14.1. Начальная перестановка

58	50	42	34	26	18	10	2
60	52	44	36	28	20	12	4
62	54	46	38	30	22	14	6
64	56	48	40	32	24	16	8
57	49	41	33	25	17	9	1
59	51	43	35	27	19	11	3
61	53	45	37	29	21	13	5
63	55	47	39	31	23	15	7

Таблица 14.2. Таблица выбора бит

32	1	2	3	4	5
4	5	6	7	8	9
8	9	10	11	12	13
12	13	14	15	16	17
16	17	18	19	20	21
20	21	12	23	24	25
24	25	26	27	28	29
28	29	30	31	32	1

$$R_{i-1} = x_1, x_2, \ldots, x_{32}$$

в биты

$$(R_{i-1})_E = x_{32}, \ldots, x_1, x_2, \ldots, x_{32}, x_1. \tag{14.22}$$

Отметим, что биты, обозначенные в первом и последнем столбцах таблицы расширения, — это те битовые разряды, которые дважды использовались для расширения из 32 до 48 бит.

Далее $(R_{i-1})_E$ суммируется по модулю 2 с i-м ключом, выбор которого описывается позднее, а результат разделяется на восемь 6-битовых блоков.

$$B_1, B_2, \ldots, B_8$$

Иными словами,

$$(R_{i-1})_E \oplus K_i = B_1, B_2, \ldots, B_8 \tag{14.23}$$

Каждый из восьми 6-битовых блоков B_j используется как вход функции S-блока, возвращающей 4-битовый блок $S_j(B_j)$. Таким образом, входные 48 бит с помощью функции S-блока преобразуются в 32 бит. Функция отображения S-блока S_j определена в табл. 14.3. Преобразование $B_j = b_1, b_2, b_3, b_4, b_5, b_6$ выполняется следующим образом. Нужная строка — это b_1b_6, а нужный столбец — $b_2b_3b_4b_5$. Например, если $b_1 = 110001$, то преобразование S_1 возвращает значение из строки 3, столбца 8, т.е. число 5 (в двоичной записи 0101). 32-битовый блок, полученный на выходе S-блока, переставляется с использованием таблицы перестановки (табл. 14.4). Как и другие таблицы, P-таблица читается слева направо и сверху вниз, так что в результате перестановки битов $x_1, x_2, ..., x_{32}$ получаем $x_{16}, x_7, ..., x_{25}$. 32-битовый выход P-таблицы суммируется по модулю 2 с 32 бит левой половины (L_{i-1}), образуя выходные 32 бит правой половины (R_i).

Таблица 14.3. Функции выбора S-блока

	Столбец																
Строка	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
0	14	4	13	1	2	15	11	8	3	10	6	12	5	9	0	7	
1	0	15	7	4	14	2	13	1	10	6	12	11	9	5	3	8	
2	4	1	14	8	13	6	2	11	15	12	9	7	3	10	5	0	S_1
3	15	12	8	2	4	9	1	7	5	11	3	14	10	0	6	13	
0	15	1	8	14	6	11	3	4	9	7	2	13	12	0	5	10	
1	3	13	4	7	15	2	8	14	12	0	1	10	6	9	11	5	
2	0	14	7	11	10	4	13	1	5	8	12	6	9	3	2	15	S_2
3	13	8	10	1	3	15	4	2	11	6	7	12	0	5	14	9	
0	10	0	9	14	6	3	15	5	1	13	12	7	11	4	2	8	
1	13	7	0	9	3	4	6	10	2	8	5	14	12	11	15	1	
2	13	6	4	9	8	15	3	0	11	1	2	12	5	10	14	7	S_3
3	1	10	13	0	6	9	8	7	4	15	14	3	11	5	2	12	
0	7	13	14	3	0	6	9	10	1	2	8	5	11	12	4	15	
1	13	8	11	5	6	15	0	3	4	7	2	12	1	10	14	9	
2	10	6	9	0	12	11	7	13	15	1	3	14	5	2	8	4	S_4
3	3	15	0	6	10	1	13	8	9	4	5	11	12	7	2	14	
0	2	12	4	1	7	10	11	6	8	5	3	15	13	0	14	9	
1	14	11	2	12	4	7	13	1	5	0	15	10	3	9	8	6	
2	4	2	1	11	10	13	7	8	15	9	12	5	6	3	0	14	S_5
3	11	8	12	7	1	14	2	13	6	15	0	9	10	4	5	3	
0	12	1	10	15	9	2	6	8	0	13	3	4	14	7	5	11	
1	10	15	4	2	7	12	9	5	6	1	13	14	0	11	3	8	
2	9	14	15	5	2	8	12	3	7	0	4	10	1	13	11	6	S_6
3	4	3	2	12	9	5	15	0	11	14	1	7	6	0	8	13	
0	4	11	2	14	15	0	8	13	3	12	9	7	5	10	6	1	
1	13	0	11	7	4	9	1	10	14	3	5	12	2	15	8	6	

Окончание табл. 14.3

	Столбец																
Строка	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
2	1	4	11	13	12	3	7	14	10	15	6	8	0	5	9	2	S_7
3	6	11	13	8	1	4	10	7	9	5	0	15	14	2	3	12	
0	13	2	8	4	6	15	11	1	10	9	3	14	5	0	12	7	
1	1	15	13	8	10	3	7	4	12	5	6	11	0	14	9	2	
2	7	11	4	1	9	12	14	2	0	6	10	13	15	3	5	8	S_8
3	2	1	14	7	4	10	8	13	15	12	9	0	3	5	6	11	

Таблица 14.4. Таблица перестановки

16	7	20	21
29	12	28	17
1	15	23	26
5	18	31	10
2	8	24	14
32	27	3	9
19	13	30	6
22	11	4	25

Алгоритм стандартного блока может быть представлен следующим образом.

$$L_i = R_{i-1} \tag{14.24}$$

$$R_i = L_{i-1} \oplus f(R_{i-1}, K_i) \tag{14.25}$$

Здесь $f(R_{i-1}, K_i)$ обозначает функциональное соотношение, включающее описанные выше расширение, преобразование в S-блоке и перестановку. После 16 итераций в таких стандартных блоках данные размещаются согласно окончательной обратной перестановке, описанной в табл. 14.5, где, как и ранее, выходные биты читаются слева направо и сверху вниз.

Таблица 14.5. Окончательная перестановка

40	8	48	16	56	24	64	32
39	7	47	15	55	23	63	31
38	6	46	14	54	22	62	30
37	5	45	13	53	21	61	29
36	4	44	12	52	20	60	28
35	3	43	11	51	19	59	27
34	2	42	10	50	18	58	26
33	1	41	9	49	17	57	25

Для дешифрования применяется тот же алгоритм, но ключевая последовательность, используемая в стандартном блоке, берется в обратном порядке. Отметим, что

значение $f(R_{i-1}, K_i)$, которое может быть также выражено через выход i-го блока как $f(L_i, K_i)$, делает процесс дешифрования возможным.

14.3.5.1. Выбор ключа

Выбор ключа также происходит в течение 16 итераций, как показано в соответствующей части рис. 14.11. Входной ключ состоит из 64-битового блока с 8 бит четности в разрядах 8, 16, ..., 64. Перестановочный выбор 1 отбрасывает биты четности и переставляет оставшиеся 56 бит согласно табл. 14.6. Выход данной процедуры делится пополам на два элемента — C и D, каждый из которых состоит из 28 бит. Выбор ключа проходит за 16 итераций, проводимых для создания различных множеств 48 ключевых бит для каждой итерации шифрования. Блоки C и D последовательно сдвигаются согласно следующим выражениям.

Таблица 14.6. Круговая перестановка

57	49	41	33	25	17	9
1	58	50	42	14	26	18
10	2	59	51	43	35	27
19	11	3	60	52	44	36
63	55	47	39	31	23	15
7	62	54	46	38	30	22
14	6	61	53	45	37	29
21	13	5	28	20	12	4

$$C_i = LS_i(C_{i-1}) \quad \text{и} \quad D_i = LS_i(D_{i-1}) \tag{14.26}$$

Здесь LS_i — левый циклический сдвиг на число позиций, показанных в табл. 14.7. Затем последовательность C_i, D_i переставляется согласно перестановочному выбору 2, показанному в табл. 14.8. Результатом является ключевая последовательность K_i, которая используется в i-й итерации алгоритма шифрования.

Таблица 14.7. Ключевая последовательность сдвигов влево

Итерация i	Количество сдвигов влево
1	1
2	1
3	2
4	2
5	2
6	2
7	2
8	2
9	1
10	2
11	2
12	2

Окончание табл. 14.7

Итерация *i*	Количество сдвигов влево
13	2
14	2
15	2
16	1

Таблица 14.8. Ключевая перестановка 2

14	17	11	24	1	5
3	28	15	6	21	10
23	19	12	4	26	8
16	7	27	20	13	2
41	52	31	37	47	55
30	40	51	45	33	48
44	49	39	56	34	53
46	42	50	36	29	32

DES может реализовываться подобно блочной системе шифрования (см. рис. 14.11), что иногда называют методом *шифровальной книги*. Основным недостатком этого метода является то, что (при использовании одного ключа) данный блок входного открытого текста будет всегда давать тот же выходной шифрованный блок. Еще один способ шифрования, называемый способом *шифрования с обратной связью*, приводит к шифрованию отдельных битов, а не символов, что дает поточное шифрование [3]. В системе шифрования с обратной связью (описанной ниже) шифрование сегмента открытого текста зависит не только от ключа и текущих данных, но и от некоторых предшествующих данных.

С конца 1970-х широко обсуждались два спорных момента, связанных с DES [10]. Первый касается длины ключа. Некоторые исследователи считали, что 56 бит не достаточно, чтобы исключить взлом путем перебора. Второй момент касается внутренней структуры *S*-блоков, которые никогда не выпускались IBM. Агентство национальной безопасности США, которое было привлечено к тестированию алгоритма DES, потребовало, чтобы эта информация не обсуждалась публично. Критики опасаются, что АНБ участвовало в проектировании этих схем и теперь способно "проникать" в любое сообщение, шифрованное согласно DES [10]. В настоящее время стандарт DES больше не является приемлемым выбором, обеспечивающим надежное шифрование. Поиск 56-битового ключа с помощью недорогих компьютерных методов является делом нескольких дней [11]. (Некоторые альтернативные алгоритмы обсуждаются в разделе 14.6.)

14.4. Поточное шифрование

Ранее мы определили *разовое заполнение* как систему шифрования со случайным одноразовым ключом, который обеспечивает безусловную защищенность. Реализовать разовое поточное заполнение можно с использованием действительно случайного потока ключей (ключевая последовательность никогда не повторяется). Таким образом, совершенная секретность может достигаться для бесконечного числа сообщений, так как каждое сообщение шифруется с помощью разных частей случайного ключевого

потока. Развитие схем поточного шифрования — это попытка имитации схем одномоментного заполнения. Большой упор делается на генерации ключевых потоков, которые должны выглядеть случайными. Реализовать такие последовательности можно с помощью соответствующих алгоритмов. Названная технология поточного шифрования использует псевдослучайные последовательности; их название отражает тот факт, что они выглядят случайными для случайного наблюдателя. Статистические свойства двоичных псевдослучайных последовательностей подобны получаемым при случайном подбрасывании симметричной монеты. В то же время, разумеется, эти последовательности являются детерминистическими (см. раздел 12.2). Данные технологии популярны, поскольку алгоритмы шифрования и дешифрования воплощаются с использованием регистров сдвига с обратной связью. На первый взгляд может показаться, что поточный псевдослучайный ключ может обеспечивать ту же защищенность, что и метод одномоментного заполнения, поскольку период последовательности, порожденной линейным регистром сдвига, составляет 2^{n-1} бит, где n — количество разрядов в регистре. Если псевдослучайная последовательность воплощается с помощью 50-разрядного регистра и дискретности в 1 МГц, последовательность будет повторяться каждые $2^{50} - 1$ микросекунды, или каждые 35 лет. В эпоху больших интегральных схем совсем несложно реализовать схему с 100 разрядами. В этом случае последовательность будет повторяться каждые 4×10^{16} лет. Следовательно, можно предположить, что поскольку псевдослучайная последовательность не повторяется в течение такого длительного периода, она может казаться действительно случайной и давать совершенную секретность. Но все же существует одно важное отличие псевдослучайной последовательности от действительно случайной последовательности, используемой в методе одномоментного заполнения. Псевдослучайная последовательность генерируется алгоритмом. Таким образом, если известен алгоритм, то известна и сама последовательность. В разделе 14.4.2 будет показано, что из-за этой особенности схема шифрования, которая использует линейный регистр сдвига с обратной связью, слишком уязвима к *атаке известного открытого текста*.

14.4.1. Пример генерирования ключа с использованием линейного регистра сдвига с обратной связью

В технологии поточного шифрования для генерации псевдослучайной ключевой последовательности обычно используются регистры сдвига. Регистр сдвига может быть превращен в генератор псевдослучайной последовательности путем введения контура обратной связи, который вычисляет новый элемент для первого разряда, основываясь на предыдущих n элементах. Говорят, что регистр является линейным, если линейна операция, производимая в контуре обратной связи. В разделе 12.2 мы уже рассматривали пример генератора псевдослучайной последовательности. На рис. 14.13 этот генератор приведен повторно. В данном случае разряды регистра удобно нумеровать так, как показано на рис. 14.13, где $n = 4$, а выходы разрядов 1 и 2 суммируются по модулю 2 (линейная операция) и передаются обратно на разряд 4. Если начальное состояние разрядов (x_4, x_3, x_2, x_1) — это 1000, то следующие состояния будут выглядеть как 1000, 0100, 1001, 1100 и т.д. Выходная последовательность составлена из битов, снимаемых с крайнего правого разряда регистра, т.е. 111101011001000, где крайний правый бит в последовательности является самым ранним, а крайний левый — наиболее поздним. При данном произвольном n-разрядном линейном регистре сдвига с обратной связью выходная последовательность в конечном счете периодична.

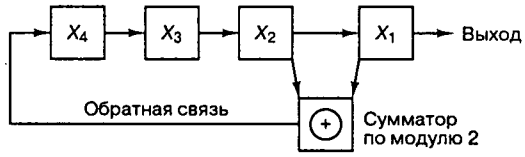


Рис. 14.13. Пример линейного регистра сдвига с обратной связью

14.4.2. Слабые места линейных регистров сдвига с обратной связью

Схема шифрования, в которой для порождения ключевого потока применяются линейные регистры сдвига с обратной связью (linear feedback shift register — LFSR), является очень уязвимой по отношению к атакам. Чтобы определить отводы обратной связи, начальное состояние регистра и всю последовательность кода, криптоаналитику требуется всего $2n$ бит открытого текста и соответствующий им шифрованный текст. Как правило, $2n$ намного меньше периода $2^n - 1$. Проиллюстрируем эту уязвимость с помощью примера регистра, изображенного на рис. 14.13. Пусть криптоаналитику, который ничего не знает о внутренних связях регистра, удалось получить $2n = 8$ бит шифрованного текста и их открытый эквивалент.

Открытый текст: 01010101

Шифрованный текст: 00001100

Здесь крайний правый бит получен первым, а крайний левый — последним.

Чтобы получить фрагмент ключевого потока 01011001 (рис. 14.14), криптоаналитик складывает обе последовательности по модулю 2. Ключевой поток показывает содержание регистров в различные моменты времени. Крайние правые четыре ключевых бита показывают содержание регистра сдвига в момент t_1. Если последовательно "сдвигать" эту четверку на один символ влево, то получим содержимое регистра в моменты t_2, t_3, t_4. Используя линейную структуру регистра сдвига, можно записать следующее.

$$g_4x_4 + g_3x_3 + g_2x_2 + g_1x_1 = x_5 \qquad (14.27)$$

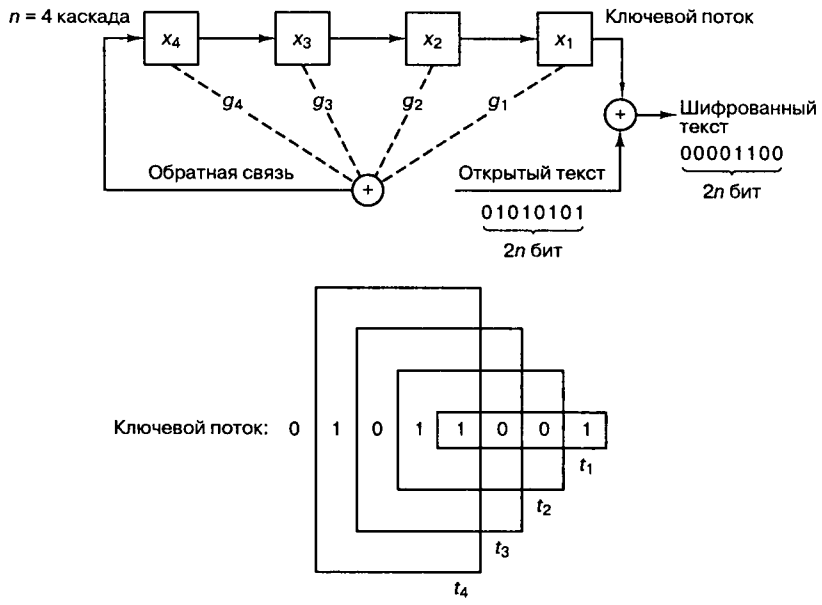


Рис. 14.14. Пример уязвимости линейного регистра сдвига с обратной связью

Здесь x_5 — цифра, которая через контур обратной связи подана обратно на вход, а g_i (= 1 или 0) определяет i-е соединение обратной связи. Таким образом, изучая содержание регистра в четыре момента времени, изображенных на рис. 14.14, можно написать следующие четыре уравнения с четырьмя неизвестными.

$$\begin{aligned} g_4(1) + g_3(0) + g_2(0) + g_1(1) &= 1 \\ g_4(1) + g_3(1) + g_2(0) + g_1(0) &= 0 \\ g_4(0) + g_3(1) + g_2(1) + g_1(0) &= 1 \\ g_4(1) + g_3(0) + g_2(1) + g_1(1) &= 0 \end{aligned} \tag{14.28}$$

Решение уравнений (14.28), соответствующих регистру, изображенному на рис. 14.13, является $g_1 = 1$, $g_2 = 1$, $g_3 = 0$, $g_4 = 0$. Таким образом, криптоаналитик узнал связи регистра, а также его начальное состояние в момент t_1. Следовательно, он может узнать последовательность в любой момент времени [3]. Обобщив этот пример на любой регистр сдвига с n разрядами, можно переписать уравнение (14.27) следующим образом.

$$x_{n+1} = \sum_{i=1}^{n} g_i x_i \tag{14.29}$$

Уравнение (14.29) можно записать в матричной форме.

$$\mathbf{x} = \mathbf{X}\mathbf{g}, \tag{14.30}$$

где

$$\mathbf{x} = \begin{bmatrix} x_{n+1} \\ x_{n+2} \\ \cdot \\ \cdot \\ \cdot \\ x_{2n} \end{bmatrix} \quad \mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \\ \cdot \\ \cdot \\ \cdot \\ g_n \end{bmatrix}$$

и

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ x_2 & x_3 & \cdots & x_{n+1} \\ \vdots & \vdots & & \vdots \\ x_n & x_{n+1} & \cdots & x_{2n-1} \end{bmatrix}.$$

Можно показать [3], что столбцы **X** линейно независимы; таким образом, матрица **X** невырождена (ее определитель отличен от нуля) и имеет обратную. Следовательно,

$$\mathbf{g} = \mathbf{X}^{-1}\mathbf{x} \tag{14.31}$$

Обращение матрицы требует порядка n^3 операций и, таким образом, легко выполняется на компьютере для любого разумного значения n. Например, если $n = 100$, то $n^3 = 10^6$, и компьютеру со скоростью работы одна операция за 1 мкс для обращения матрицы понадобится 1 с. Слабость регистра сдвига с обратной связью обусловлена линейностью уравнения (14.31). Использование *нелинейной обратной связи* в регистре сдвига делает задачу криптоаналитика гораздо сложнее, если не вычислительно трудноосуществимой.

14.4.3. Синхронные и самосинхронизирующиеся системы поточного шифрования

Системы поточного шифрования можно разделить на *синхронные* и *самосинхронизирующиеся*. В первых ключевой поток генерируется независимо от сообщения; так что потеря символа во время передачи неизбежно требует повторной синхронизации передачи и генераторов ключей приемника. Синхронный поточный шифр изображен на рис. 14.15. Начальное состояние генератора ключа инициализируется с помощью известного входа I_0. Шифрованный текст получается путем сложения по модулю 2 i-го символа ключа k_i и i-го символа сообщения m_i. Такие синхронные шифры обычно создаются для *смешения* (см. раздел 14.3.1), но не *диффузии*. Иными словами, шифрование символа не распространяется вдоль некоторого блока сообщения. По этой причине синхронные поточные шифры не имеют *накопления ошибки*.

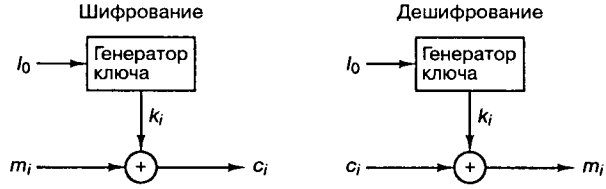


Рис. 14.15. Синхронный поточный шифр

При *самосинхронизирующемся* поточном шифре каждый ключевой символ определяется из фиксированного числа n предшествующих символов шифрованного текста (отсюда и название *обратная связь по шифру*). В таких системах происходит следующее: если символ шифрованного текста теряется во время передачи, ошибка накапливается для n символов, но после получения n верных символов шифрованного текста система восстанавливается.

В разделе 14.1.4 приводился пример обратной связи для шифрования с помощью автоматического ключа Вигнера. Показывалось, что преимуществом такой системы является: (1) генерация неповторяющегося ключа и (2) диффузия статистик открытого сообщения в шифрованном тексте. В то же время был и недостаток — ключ проявлялся в шифрованном тексте. Этой проблемы можно избежать, если при получении ключа пропустить символы шифрованного текста через нелинейный блок шифрования. На рис. 14.16 изображен регистр сдвига генератора ключа, работающий в режиме обратной связи по шифру. Каждый выходной символ шифрованного текста c_i (образованный путем сложения по модулю 2 символа сообщения m_i и символа ключа k_i) подается обратно на вход регистра сдвига. Как и ранее, инициализация происходит с помощью известного входа I_0. При каждой итерации выход регистра сдвига используется как вход (нелинейного) блочного алгоритма шифрования E_B. Символ младшего разряда на выходе E_B становится следующим символом ключа k_{i+1}, который используется в следующем символе сообщения m_{i+1}. Поскольку после нескольких первых итераций вход алгоритма зависит только от шифрованного текста, система является самосинхронизирующейся.

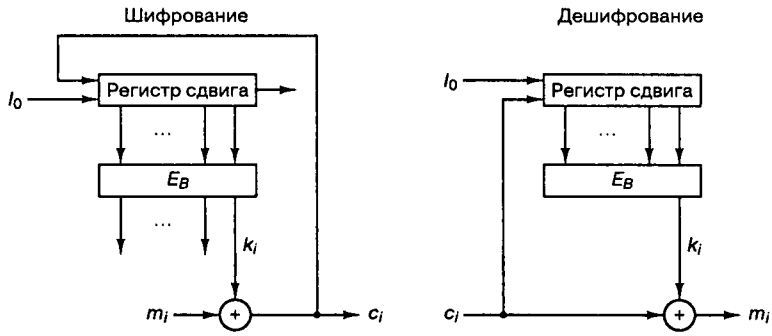


Рис. 14.16. Шифрование в режиме обратной связи

14.5. Криптосистемы с открытыми ключами

Понятие систем с открытыми ключами было введено в 1976 году Диффи (Diffie) и Хэллманом (Hellman) [12]. В общепринятых криптосистемах алгоритм шифрования может быть обнаружен, поскольку защищенность системы зависит от сохранности ключа. Один и тот же ключ применяется как для шифрования, так и для дешифрования. Криптосистемы с открытыми ключами используют *два разных ключа*: один — для шифрования, другой — для дешифрования. В таких криптосистемах общедоступными (без потери защищенности системы) могут быть не только алгоритм шифрования, но и ключ, применяемый для шифрования. Фактически это общедоступный каталог, подобный телефонному каталогу, который содержит ключи шифрования всех абонентов. Держатся в секрете только ключи дешифрования. Пример такой системы приведен на рис. 14.17. Перечислим важные особенности криптосистемы с открытым ключом.

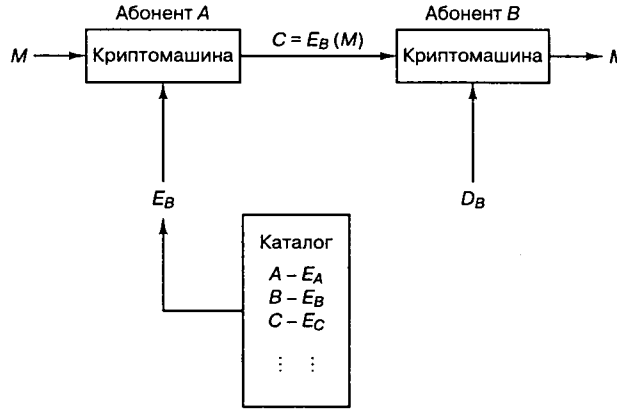


Рис. 14.17. Криптосистема с открытым ключом

1. Алгоритм шифрования E_K и алгоритм дешифрования D_K являются обратимыми преобразованиями открытого текста M или шифрованного текста C, определяемыми ключом K.
2. Для каждого ключа K алгоритмы E_K и D_K легко вычисляемы.
3. Для каждого ключа K определение D_K из E_K вычислительно трудноосуществимо.

Такая система обычно способна обеспечивать защищенность переговоров между пользователями, которые никогда ранее не встречались или не общались. Например, как показано на рис. 14.17, пользователь A может послать сообщение пользователю B, найдя ключ шифрования пользователя B в каталоге и используя алгоритм шифрования E_B. Получив таким образом шифрованный текст $C = E_B(M)$, он передает его через общедоступный канал. Пользователь B — это единственный человек, который может дешифровать сообщение C, чтобы в результате получилось $M = D_B(C)$, с помощью своего алгоритма дешифрования D_B.

14.5.1. Проверка подлинности подписи с использованием криптосистемы с открытым ключом

На рис. 14.18 изображено применение криптосистемы с открытым ключом для проверки подлинности подписи. Пользователь A "подписывает" свое сообщение, используя свой алгоритм дешифрования D_A, что дает $S = D_A(M) = E_A^{-1}(M)$. Затем для шифрования S он воспользуется алгоритмом шифрования E_B пользователя B и в результате получит сообщение $C = E_B(S) = E_B[E_A^{-1}(M)]$, которое он передает через общедоступный канал. Когда пользователь B получает сообщение C, он сначала дешифрует его с помощью собственного алгоритма дешифрования D_B, что дает $D_B(C) = E_A^{-1}(M)$. Затем он использует алгоритм шифрования пользователя A, в результате чего получает $E_A[E_A^{-1}(M)] = M$.

Если в результате получается вразумительное сообщение, оно точно было послано пользователем A, поскольку больше никто не знает секретного кода шифрования пользователя A, с помощью которого выполняется преобразование $S = D_A(M)$. Отметим, что сообщение S зависит и от сообщения, и от подписи, а это означает, что не только B может быть уверен, что сообщения действительно приходят от A, но и A уверен, что никто, кроме B, не сможет прочесть это сообщение.

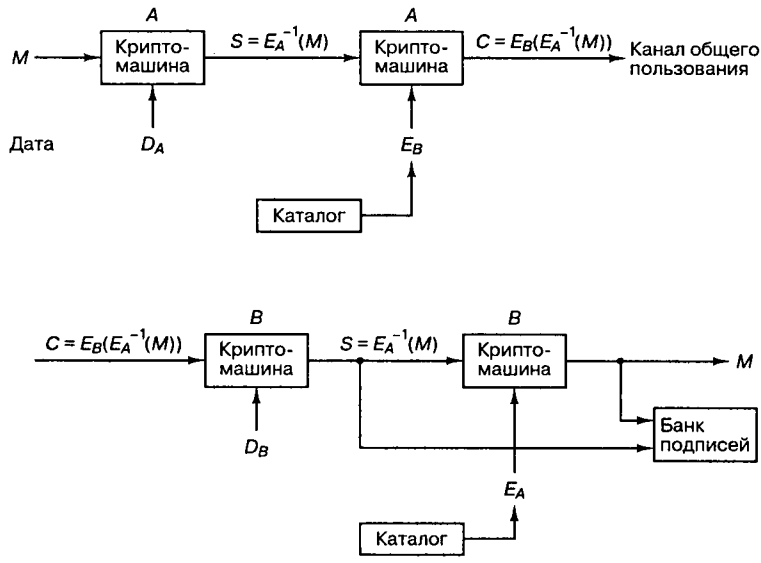


Рис. 14.18. Проверка подлинности подписи с использованием криптосистемы с открытым ключом

14.5.2. Односторонняя функция с "лазейкой"

Криптосистемы с открытым ключом основаны на понятии односторонних функций с "лазейками". Определим *одностороннюю функцию* как легко вычисляемую, для которой невозможно вычислить обратную. Рассмотрим, например, функцию $y = x^5 + 12x^3 + 107x + 123$. Должно быть очевидно, что при данном x легко вычислить y, но при данном y относительно сложно вычислить x. *Односторонняя функция с "лазейкой"* — это односторонняя функция, для которой легко вычислить обратную, если известны некоторые особенности, используемые для создания функции. Как и лазейка, такие функции легко проходимы в одном направлении. Обратный процесс без специальной информации занимает невероятно много времени. Понятие "лазейки" будет применено в разделе 14.5.5, когда будет обсуждаться схема Меркла-Хэллмана (Merkle-Hellman).

14.5.3. Схема RSA

Сообщения в схеме Ривеста-Шамира-Адельмана (Rivest-Shamir-Adelman — RSA) сначала представляются как целые числа из интервала $(0, n-1)$. Каждый пользователь выбирает собственное значение n и пару положительных целых чисел e и d описанным ниже способом. Пользователь помещает свой ключ шифрования, числовую пару (n, e), в общедоступный каталог. Ключ дешифрования состоит из числовой пары (n, d), в которой d держится в секрете. Шифрование сообщения M и дешифрование шифрованного текста C определяются следующим образом.

$$\begin{aligned} &\text{Шифрование: } C = E(M) = (M)^e \text{ по модулю } n \\ &\text{Дешифрование: } M = D(C) = (C)^d \text{ по модулю } n \end{aligned} \tag{14.32}$$

Это легко вычислить. Результатом каждой операции являются целые числа из интервала $(0, n-1)$. В схеме RSA n получается в результате перемножения *двух больших простых чисел* p и q.

$$n = pq \tag{14.33}$$

Несмотря на то что n общедоступно, p и q являются скрытыми из-за большой сложности в разложении n на множители. Затем определяется функция, называемая *функцией Эйлера.*

$$\phi(n) = (p-1)(q-1) \tag{14.34}$$

Параметр $\phi(n)$ имеет интересное свойство [12]: для любого целого X из интервала $(0, n-1)$ и любого целого k имеет место следующее соотношение.

$$X = X^{k\phi(n)+1} \text{ по модулю } n \tag{14.35}$$

Следовательно, если все остальные арифметические действия выполняются по модулю n, арифметические действия в степени выполняются по модулю $\phi(n)$. Затем случайным образом выбирается большое целое число d, являющееся взаимно простым с $\phi(n)$; это означает, что $\phi(n)$ и d не должны иметь общих делителей, отличных от 1. Это записывается следующим образом.

$$\text{НОД}\,[\phi(n), d] = 1 \tag{14.36}$$

В данном случае НОД означает "наибольший общий делитель". Этому условию будет удовлетворять любое простое число, большее наибольшего из (p, q). Далее находится целое e, $0 < e < \phi(n)$,

$$ed \text{ по модулю } \phi(n) = 1, \tag{14.37}$$

что, вследствие равенства (14.35), равносильно выбору e и d, которые удовлетворяют следующему условию.

$$X = X^{ed} \text{ по модулю } n \tag{14.38}$$

Следовательно,

$$E[D(X)] = D[E(X)] = X \tag{14.39}$$

и возможно корректное дешифрование. Один из возможных способов взлома шифра при данном ключе (n, e) — это разложить n на множители p и q, вычислить $\phi(n) = (p-1)(q-1)$ и вычислить d из равенства (14.37). Все это, за исключением разложения n на множители, представляет собой простые действия.

Схема RSA основывается на том, что два больших простых целых числа p и q легко выбрать и перемножить, но гораздо сложнее разложить на множители результат. Следовательно, произведение, как часть ключа шифрования, может быть сделано общедоступным, в то время как множители, которые могут "разоблачить" ключ дешифрования, соответствующий ключу шифрования, остаются скрытыми. Если длина каждого множителя составляет порядка 100 разрядов, умножение может быть выполнено в доли секунды, а изнурительное разложение на множители результата может потребовать миллиарды лет [2].

14.5.3.1. Использование схемы RSA

Используя пример из работы [13], положим $p = 47$, $q = 59$. Следовательно, $n = pq = 1773$ и $\phi(n) = (p-1)(q-1) = 2668$. Параметр d выбирается взаимно простым с $\phi(n)$. Например, выберем $d = 157$. Затем вычислим значение e следующим образом (подробности приведены в следующем разделе).

$$ed \text{ по модулю } \phi(n) = 1$$

$$157e \text{ по модулю } 2688 = 1$$

Следовательно, $e = 17$. Рассмотрим пример открытого текста.

ITS ALL GREEK TO ME

Если заменить каждую букву двухразрядным числом из интервала (01, 26), соответствующим ее позиции в алфавите, и закодировать пробел как 00, открытое сообщение можно записать следующим образом.

0920 1900 0112 1200 0718 0505 1100 2015 0013 0500

Каждый символ выражается целым числом из интервала $(0, n-1)$. Поэтому в данном примере шифрование может быть представлено в виде блоков по четыре разряда, так как это максимальное число разрядов, которое всегда дает число, меньшее $n-1 = 2772$. Первые четыре разряда (0920) открытого текста шифруются следующим образом.

$$C = (M)^e \text{ по модулю } n = (920)^{17} \text{ по модулю } 2773 = 948$$

Продолжая этот процесс для оставшихся разрядов открытого текста, получим следующее.

$$C = 0948 \quad 2342 \quad 1084 \quad 1444 \quad 2663 \quad 2390 \quad 0778 \quad 0774 \quad 0229 \quad 1655$$

Открытый текст восстанавливается с помощью ключа дешифрования.

$$M = (C)^{157} \text{ по модулю } 2773$$

14.5.3.2. Как вычислить e

Для вычисления e используется разновидность алгоритма Евклида вычисления НОД $\phi(n)$ и d. Сначала вычисляем последовательность значений $x_0, x_1, x_2, \ldots$, где $x_0 = \phi(n)$, $x_1 = d$, а $x_{i+1} = x_{i-1}$ по модулю x_i, пока не будет получено $x_k = 0$. Тогда НОД$(x_0, x_1) = x_{k-1}$. Для каждого x_i вычисляются числа a_i и b_i, при которых $x_i = a_i x_0 + b_i x_i$. Если $x_{k-1} = 1$, то b_{k-1} — мультипликативное обратное к x_1 по модулю x_0. Если b_{k-1} — отрицательное число, решением является $b_{k-1} + \phi(n)$.

Пример 14.5. Вычисление e с помощью d и $\phi(n)$

Для предыдущего примера, в котором $p = 47$, $q = 59$, $n = 2773$ и d выбрано равным 157, примените алгоритм Евклида для проверки, что $e = 17$.

Решение

i	x_i	a_i	b_i	y_i
0	2668	1	0	
1	157	0	1	16
2	156	1	-16	1
3	1	-1	17	

Здесь

$$y_i = \left\lfloor \frac{x_{i-1}}{x_i} \right\rfloor$$

$$x_{i+1} = x_{i-1} - y_i x_i \,.$$

$$a_{i+1} = a_{i-1} - y_i a_i$$

$$b_{i+1} = b_{i-1} - y_i b_i$$

Следовательно,

$$e = b_3 = 17.$$

14.5.4. Задача о рюкзаке

Классическая задача о рюкзаке изображена на рис. 14.19. Рюкзак наполнен множеством предметов с указанием их веса в граммах. Зная вес наполненного рюкзака (шкала весов градуирована так, что вес пустого рюкзака вычитается), нужно определить содержимое рюкзака. В этом простом примере решение легко найти методом проб и ошибок. Однако если в заданном множестве не 10, а 100 возможных единиц, задача может стать вычислительно неосуществимой.

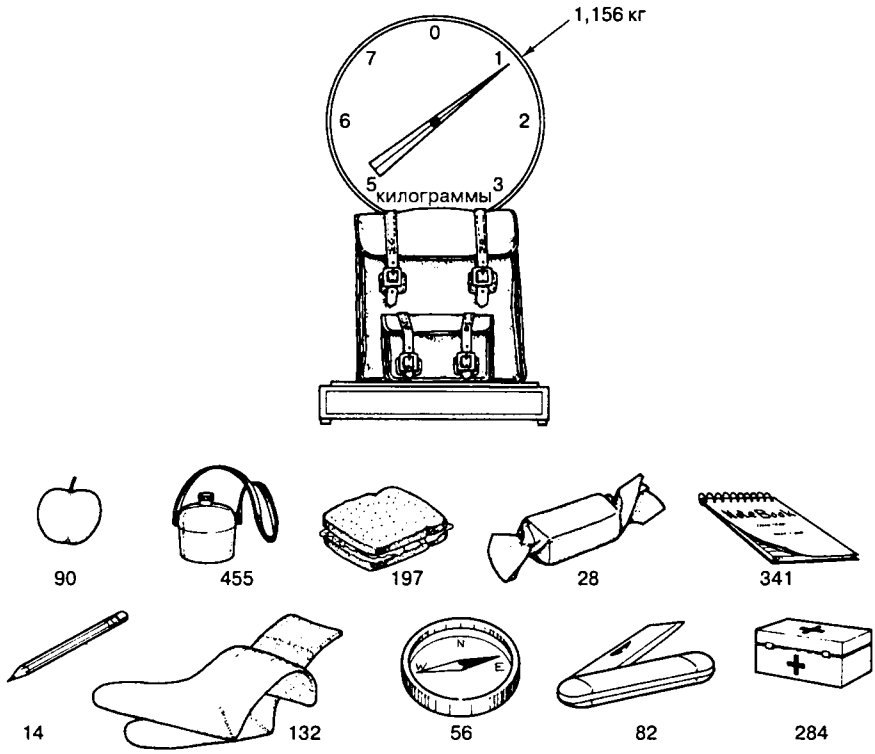


Рис. 14.19. Задача о рюкзаке

Опишем задачу о рюкзаке через вектор рюкзака и вектор данных. Вектор рюкзака представляет собой n-кортеж разных целых чисел (аналогично множеству разных предметов содержимого рюкзака).

$$\mathbf{a} = a_1, a_2, \ldots, a_n$$

Вектор данных — это n-кортеж двоичных символов.

$$\mathbf{x} = x_1, x_2, \ldots, x_n$$

Рюкзак S — это сумма подмножества компонентов вектора рюкзака.

$$S = \sum_{i=1}^{n} a_i x_i = \mathbf{ax}\ , \qquad \text{где } x_i = 0,\ 1 \tag{14.40}$$

Задачу о рюкзаке можно сформулировать следующим образом: при данном S и известном **a** определите **x**.

Пример 14.6. Пример рюкзака

Дано **a** = 1, 2, 4, 8, 16, 32 и S = **ax** =26. Найдите **x**.

Решение

Видно, что в этом примере **x** — это *двоичное* представление S. Преобразование из десятичного в двоичное окажется более знакомым, если представить **a** как 2^0, 2^1, 2^2, 2^3, 2^4, 2^5, 2^6. Вектор данных **x** находится легко, поскольку **a** в этом примере является *быстровозрастающим*; это означает, что каждый компонент набора n-кортежа **a** больше суммы предыдущих компонентов. Другими словами,

$$a_i > \sum_{j=1}^{i-1} a_i \qquad i = 2, 3, \ldots, n \tag{14.41}$$

Если **a** является быстровозрастающим, то первый элемент **x** — $x_n = 1$, если $S \geq a_n$ (в противном случае, $x_n = 0$); следующий элемент находится согласно соотношению

$$x_i = \begin{cases} 1, & \text{если } S - \sum_{j=i+1}^{n} x_j a_j \geq a_i \\ 0, & \text{в других случаях} \end{cases}, \tag{14.42}$$

где $i = n - 1, n - 2, \ldots, 1$. С помощью равенства (14.42) легко вычисляется **x** = 010110.

Пример 14.7. Пример рюкзака

Дано **a** = 171, 197, 459, 1191, 2410, 4517 и S = **ax** = 3798. Найдите **x**.

Решение

Как и в примере 14.6, **a** является быстровозрастающим. Поэтому с помощью равенства (14.42) можно вычислить **x**.

$$\mathbf{x} = 010110$$

14.5.5. Криптосистема с открытым ключом, основанная на "лазейке" в рюкзаке

Эта схема, известная также как схема Меркла-Хэллмана [15], основана на образовании вектора рюкзака, который не является быстровозрастающим. Следовательно, задача не является легкоразрешимой. При этом данная задача о рюкзаке обязательно включает *лазейку*, позволяющую разрешенным пользователям решить задачу.

Сначала образуем быстровозрастающий n-кортеж. Затем выберем простое число M, при котором имеет место следующее неравенство.

$$M > \sum_{i=1}^{n} a_i' \tag{14.43}$$

Выберем также случайное число W ($1 < W < M$) и сформируем W^{-1}, удовлетворяющее следующему соотношению.

$$WW^{-1} \text{ по модулю } M = 1 \tag{14.44}$$

Вектор $\mathbf{a}'$ и числа M, W и W^{-1} удерживаются скрытыми. Затем из элементов $\mathbf{a}'$ формируем $\mathbf{a}$.

$$a_i = Wa_j' \text{ по модулю } M \tag{14.45}$$

Формирование $\mathbf{a}$ с использованием равенства (14.45) — это создание вектора рюкзака с *лазейкой*. Если нужно передать вектор $\mathbf{x}$, то вначале $\mathbf{x}$ умножается на $\mathbf{a}$, что дает число S, которое передается через общедоступный канал. С помощью равенства (14.45) S можно записать следующим образом.

$$S=\mathbf{ax}=\sum_{i=1}^{n} a_i x_i = \sum_{i=1}^{n} (Wa_i' \text{ по модулю } M)x_i \tag{14.46}$$

Разрешенный пользователь получает S и, используя равенство (14.44), превращает его в S'.

$$\begin{aligned} S' = W^{-1}S \text{ по модулю } M &= W^{-1}\sum_{i=1}^{n} (Wa_i' \text{ по модулю } M)x_i \text{ по модулю } M = \\ &= \sum_{i=1}^{n} (W^{-1}Wa_i' \text{ по модулю } M)x_i \text{ по модулю } M = \\ &= \sum_{i=1}^{n} a_i' x_i \text{ по модулю } M = \\ &= \sum_{i=1}^{n} a_i' x_i \end{aligned} \tag{14.47}$$

Поскольку разрешенный пользователь знает засекреченный быстровозрастающий вектор $\mathbf{a}'$, для отыскания $\mathbf{x}$ он может использовать S'.

14.5.5.1. Использование схемы Меркла-Хэллмана

Предположим, пользователь A желает создать общедоступную и конфиденциальную функции шифрования. Сначала он рассматривает быстровозрастающий вектор $\mathbf{a}' = (171, 197, 459, 1191, 2410, 4517)$.

$$\sum_{i=1}^{6} a_i' = 8945$$

Затем он выбирает простое число M, большее 8945, случайное число W, такое, что $1 \le W < M$, и вычисляет W^{-1}, при котором $WW^{-1} = 1$ по модулю M.

$$\left.\begin{array}{l} \text{Пусть } M = 9109 \\ \text{Пусть } W = 2251 \\ \text{тогда } W^{-1} = 1388 \end{array}\right\} \text{ скрыты}$$

После этого он образует вектор, который оставляет "лазейку" в рюкзаке.

$$a_i = a'_i\ 2251 \text{ по модулю } 9109$$
$$\mathbf{a} = 2343, 6215, 3892, 2895, 5055, 2123$$

Пользователь *A* делает общедоступным вектор **a**, который, очевидно, не является быстровозрастающим. Предположим, что пользователь *B* желает послать сообщение пользователю *A*.

Если **x** = 010110 — сообщение, которое нужно передать, то пользователь *B* создает следующее число.

$$S = \mathbf{ax} = 14\ 165 \text{ и передает его пользователю } A$$

Пользователь *A* получает *S* и превращает его в *S'*.

$$S' = \mathbf{a}\hat{\mathbf{x}} = W^{-1}S \text{ по модулю } M =$$
$$= 1388 \cdot 14\ 165 \text{ по модулю } 9109 =$$
$$= 3798$$

Используя $S' = 3798$ и быстровозрастающий вектор **a'**, пользователь *A* легко находит **x**.

Схема Меркла-Хэллмана сейчас считается взломанной [16], поэтому для реализации криптосистем с открытыми ключами используется алгоритм RSA (равно как и другие рассмотренные позднее).

14.6. Pretty Good Privacy

PGP (Pretty Good Privacy, *буквально*: "весьма хорошая секретность") — это программа обеспечения секретности, которая была создана Филиппом Циммерманом (Philip Zimmerman) [17] и опубликована в 1991 году как бесплатное программное обеспечение. Затем она "де-факто" стала стандартом для электронной почты и шифрования файлов. PGP версии 2.6 (наиболее широко используемая) оставалась неизменной вплоть до появления версии 5.0 (совместимой с версией 2.6). В табл. 14.9 приведены алгоритмы, используемые в версиях 2.6, 5.0 и более поздних.

Таблица 14.9. Сравнение PGP 2.6 и PGP 5.0

Функция	PGP версии 2.6 Используемый алгоритм [17]	PGP версии 5.0 и более поздних Используемый алгоритм [18]
Шифрование сообщения с использованием алгоритма частного ключа с помощью ключа частного сеанса	IDEA	"Тройной" DES, CAST или IDEA
Шифрование ключа частного сеанса с помощью алгоритма частного ключа	RSA	RSA или алгоритм Диффи-Хэллмана (вариант Элгемала)
Цифровая подпись	RSA	RSA и DSS[1] (от NIST[2])
Хэш-функция, используемая при создании профиля сообщения для цифровых подписей	MD5	SHA-1

[1] Digital Signature Standard — Стандарт цифровой подписи, разработанный NIST.

[2] National Institute of Standards and Technology — Национальный институт стандартов и технологий США; отдел Министерства торговли США.

Как показано в табл. 14.9, PGP использует множество алгоритмов шифрования, включающих как схемы частного ключа, так и схемы открытого ключа. При шифровании сообщения применяется алгоритм частного ключа (для каждого сеанса генерируется новый ключ сеанса). В качестве алгоритмов частного ключа, предлагаемых PGP, представлены Международный алгоритм шифрования данных (International Desalination and Environmental Association — IDEA), "тройной" DES и алгоритм CAST (названный в честь авторов Карлайла Адамса (Carlisle Adams) и Стэффорда Тевереса (Stafford Tavares) [19]). Для шифрования ключа каждого сеанса используется алгоритм открытого ключа. В качестве алгоритмов, использующих открытые ключи, PGP предлагает алгоритм RSA, описанный в разделе 14.5.3, и алгоритм Диффи-Хэллмана (Diffie-Hellman).

Алгоритмы с открытыми ключами применяются также для создания цифровых подписей. PGP версии 5.0 использует алгоритм цифровой подписи (Digital Signature Algorithm — DSA), заданный в стандартах цифровой подписи (Digital Signature Standard — DSS) института NIST. PGP версии 2.6 в своих цифровых подписях использует алгоритм RSA. Если имеющийся канал не защищен от изменений ключа, он более безопасен для использования алгоритма с ключом общего доступа. Для защищенного канала предпочтительно шифрование с частным ключом, поскольку это, как правило, дает лучшее быстродействие по сравнению с системами, использующими открытые ключи.

Технология шифрования сообщения, применяемая PGP версии 2.6, изображена на рис. 14.20. Перед шифрованием открытый текст сжимается с помощью ZIP-алгоритма. Система PGP использует ZIP-метод, описанный Жаном-Лупом Гейли (Jean-Loup Gaily), Марком Элдером (Mark Alder) и Ричардом Б. Уэльсом (Richard B. Wales) [18]. Если сжатый текст короче несжатого, то шифроваться будет сжатый текст, в противном случае будет шифроваться несжатый.

Небольшие файлы (приблизительно 30 символов для файлов ASCII) не выигрывают от сжатия. К тому же, PGP распознает файлы, ранее сжатые с помощью распространенных технологий сжатия, таких как PKZIP, и не будет пытаться сжать их. Сжатие данных устраняет избыточные строки символов и приводит к более равномерному распределению символов. С помощью сжатия получаем более короткий файл для шифрования и дешифрования (что сокращает время, необходимое для шифрования, дешифрования и передачи файла). Сжатие также создает препятствия некоторым криптоаналитическим атакам, использующим избыточность. Необходимо отметить, что сжатие файла должно *предшествовать* шифрованию (а не наоборот). Почему стоит следовать этому правилу? *Хороший* алгоритм шифрования дает шифрованный текст с практически статистически равномерным распределением символов. Следовательно, если алгоритм сжатия данных следует после такого шифрования, он не будет давать никакого сжатия вообще. Если некоторый шифрованный текст может быть сжат, то алгоритм шифрования, с помощью которого получен шифрованный текст, был неудачным. Алгоритм сжатия не должен обнаруживать избыточные фрагменты в тексте, зашифрованном с помощью хорошего алгоритма.

Как показано на рис. 14.20, PGP начинает шифрование файла с создания 128-битового ключа сеанса, используя генератор псевдослучайных чисел. Затем с помощью этого случайного ключа сеанса шифруется сжатый файл открытого текста, для чего применяется алгоритм частного ключа IDEA.

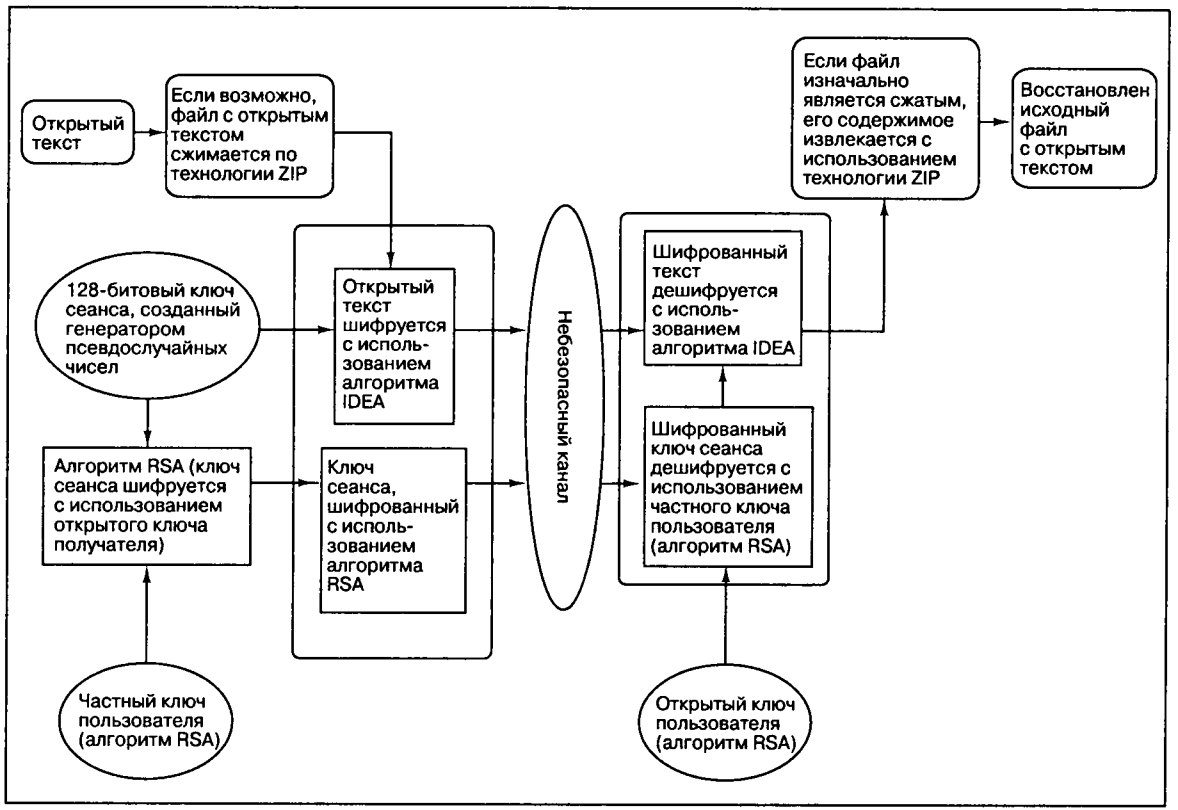


Рис. 14.20. Метод PGP

После этого случайный ключ сеанса шифруется с помощью алгоритма открытого ключа RSA; при этом используется *открытый ключ получателя*. Ключ сеанса, зашифрованный с помощью алгоритма RSA, и файл, зашифрованный с использованием алгоритма IDEA, посылаются получателю. Когда получателю нужно прочесть файл, вначале, с помощью алгоритма RSA, дешифруется зашифрованный ключ сеанса. При этом используется *частный ключ получателя*. Затем дешифруется собственно шифрованный файл, при этом применяется дешифрованный ключ сеанса и алгоритм IDEA. После разархивации получатель может читать расшифрованный файл.

14.6.1. "Тройной" DES, CAST и IDEA

Как показано в табл. 14.9, PGP предлагает три блочных шифра для шифрования сообщения — "тройной" DES, CAST и IDEA. Все три шифра оперируют 64-битовыми блоками открытого и шифрованного текстов. Размер ключа "тройного" DES составляет 168 бит, в то время как CAST и IDEA используют ключи длиной 128 бит.

14.6.1.1. Описание "тройного" DES

Стандарт шифрования данных (Data Encryption Standard — DES), описанный в разделе 14.3.5, использовался с конца 1970-х годов. Однако у многих вызывала беспокойство его защищенность, так как в нем применялся ключ относительно малого размера (56 бит). При использовании "тройного" алгоритма DES, шифруемое сообщение трижды пропускается через алгоритм DES (вторая операция проводится в режиме дешифрования). Каждая операция производится с помощью разных 56-битовых ключей. Как показано на рис. 14.21, это равносильно использованию ключа длиной 168 бит.

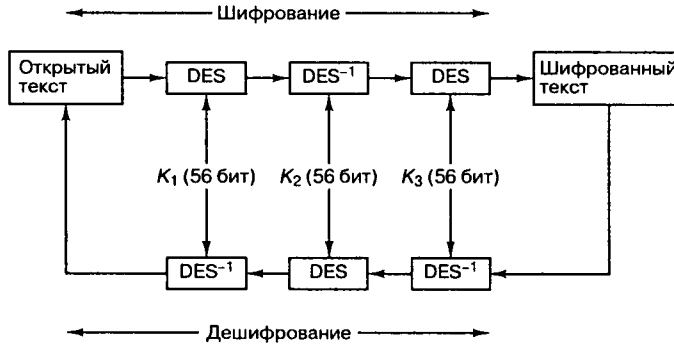


Рис. 14.21. Шифрование/дешифрование с помощью "тройного" алгоритма DES

14.6.1.2. Описание CAST

CAST — это семейство блочных шифров, разработанных Адамсом (Adams) и Тевересом (Tavares) [19]. PGP 5.0 использует версию CAST, известную как CAST5 или CAST-128. В этой версии размер блока составляет 64 бит, а длина ключа — 128 бит. Алгоритм CAST использует шесть *S*-блоков с 8-битовым входом и 32-битовым выходом. Для сравнения, DES применяет восемь *S*-блоков с 6-битовым входом и 4-битовым выходом. *S*-блоки в CAST-128 были созданы для обеспечения существенно нелинейных преобразований, которые делают этот алгоритм практически не поддающимся криптоанализу [11].

14.6.1.3. Описание IDEA

Международный алгоритм шифрования данных (International Data Encryption Algorithm — IDEA) представляет собой блочный шифр, разработанный Ксуэйя Лаи (Xuejia Lai) и Джеймсом Мэсси (James Massey) [19]. Это 64-битовый итерационный блочный шифр (включающий восемь итераций или циклов) с 128-битовым ключом. Защищенность IDEA зависит от использования трех типов арифметических операций над 16-битовыми символами: сложение по модулю 2^{16}, умножение по модулю $2^{16} + 1$ и побитовое исключающее ИЛИ. Для итерационных операций шифрования и дешифрования используется 128-битовый ключ. Как показано в табл. 14.10, начальный ключ K_0 делится на восемь 16-битовых подключей $Z_x^{(R)}$, где x — номер подключа цикла R. Шесть из этих подключей используются в цикле 1, а оставшиеся два — в цикле 2. Затем K_0 циклически сдвигается на 25 бит влево, в результате чего образуется ключ K_1, который, в свою очередь, делится на восемь подключей. Первые 4 из этих подключей используются в цикле 2, а последние четыре — в цикле 3. Процесс продолжается, как показано в табл. 14.10, в результате чего в общей сложности появляется 52 подключа.

Таблица 14.10. Образование подключей в алгоритме IDEA

128-битовый ключ (делится на восемь 16-битовых подключей)	Строка битов, из которой выводятся ключи
$Z_1^1Z_2^1Z_3^1Z_4^1Z_5^1Z_6^1Z_1^2Z_2^2$	K_0 = исходный 128-битовый ключ
$Z_3^2Z_4^2Z_5^2Z_6^2Z_1^3Z_2^3Z_3^3Z_4^3$	K_1 = сдвиг K_0 на 25 бит
$Z_5^3Z_6^3Z_1^4Z_2^4Z_3^4Z_4^4Z_5^4Z_6^4$	K_0 = сдвиг K_1 на 25 бит
$Z_1^5Z_2^5Z_3^5Z_4^5Z_5^5Z_6^5Z_1^6Z_2^6$	K_0 = сдвиг K_2 на 25 бит
$Z_3^6Z_4^6Z_5^6Z_6^6Z_1^7Z_2^7Z_3^7Z_4^7$	K_0 = сдвиг K_3 на 25 бит
$Z_5^7Z_6^7Z_1^8Z_2^8Z_3^8Z_4^8Z_5^8Z_6^8$	K_0 = сдвиг K_4 на 25 бит
$Z_1^{out}Z_2^{out}Z_3^{out}Z_4^{out}$	Первые 64 бит K_6, где K_6 = сдвиг K_5 на 25 бит

Маршрут подключа для каждого цикла показан в табл. 14.11 как для цикла шифрования, так и дешифрования. Дешифрование проводится так же, как и шифрование. Подключи дешифрования вычисляются из подключей шифрования, как показано в табл. 14.11, из которой видно, что подключи дешифрования являются либо аддитивными, либо мультипликативными, обратными к подключам шифрования.

Таблица 14.11. Эволюция подключа алгоритма IDEA

Цикл	Набор подключей шифрования	Набор ключей дешифрования
1	$Z_1^1Z_2^1Z_3^1Z_4^1Z_5^1Z_6^1$	$(Z_1^{out})^{-1} - Z_2^{out} - Z_3^{out}(Z_4^{out})^{-1}Z_5^8Z_6^8$
2	$Z_1^2Z_2^2Z_3^2Z_4^2Z_5^2Z_6^2$	$(Z_1^8)^{-1} - Z_2^8 - Z_3^8(Z_4^8)^{-1}Z_5^7Z_6^7$
3	$Z_1^3Z_2^3Z_3^3Z_4^3Z_5^3Z_6^3$	$(Z_1^7)^{-1} - Z_2^7 - Z_3^7(Z_4^7)^{-1}Z_5^5Z_6^6$
4	$Z_1^4Z_2^4Z_3^4Z_4^4Z_5^4Z_6^4$	$(Z_1^6)^{-1} - Z_2^6 - Z_3^6(Z_4^6)^{-1}Z_5^5Z_6^5$
5	$Z_1^5Z_2^5Z_3^5Z_4^5Z_5^5Z_6^5$	$(Z_1^5)^{-1} - Z_2^5 - Z_3^5(Z_4^5)^{-1}Z_5^5Z_6^5$
6	$Z_1^6Z_2^6Z_3^6Z_4^6Z_5^6Z_6^6$	$(Z_1^4)^{-1} - Z_2^4 - Z_3^4(Z_4^4)^{-1}Z_5^4Z_6^4$
7	$Z_1^7Z_2^7Z_3^7Z_4^7Z_5^7Z_6^7$	$(Z_1^3)^{-1} - Z_2^3 - Z_3^3(Z_4^3)^{-1}Z_5^3Z_6^3$
8	$Z_1^8Z_2^8Z_3^8Z_4^8Z_5^8Z_6^8$	$(Z_1^2)^{-1} - Z_2^2 - Z_3^2(Z_4^2)^{-1}Z_5^2Z_6^2$
Выходное преобразование	$Z_1^{out}Z_2^{out}Z_3^{out}Z_4^{out}$	$(Z_1^1)^{-1} - Z_2^1 - Z_3^1(Z_4^1)^{-1}$

Сообщение делится на 64-битовые блоки данных. Затем эти блоки делятся на четыре 16-битовых подблока: M_1, M_2, M_3 и M_4. Последовательность таких четырех подблоков становится входом первого цикла алгоритма IDEA. Эти данные используются для всех восьми циклов. Как показано в табл. 14.11, в каждом цикле применяются разные множества из шести подключей. После завершения цикла второй и третий 16-битовые блоки данных переставляются. После завершения восьмого цикла четыре подблока дают окончательное выходное преобразование. Для упрощения записи в представлении $Z_x^{(R)}$ в табл. 14.10 и 14.11 опущены круглые скобки.

Каждый цикл состоит из шагов, показанных в табл. 14.12. Окончательные значения, полученные на шагах 11–14, образуют выход цикла. Два внутренних 16-битовых подблока данных переставляются (за исключением последнего цикла), затем эти четыре подблока составляют вход следующего цикла. Этот метод в общей сложности включает 8 циклов. После восьмого цикла окончательное выходное преобразование имеет следующий вид.

1. $M_1 \times Z_1^{out}$ (первый подключ выходного преобразования)
2. $M_2 \times Z_2^{out}$
3. $M_3 \times Z_3^{out}$
4. $M_4 \times Z_4^{out}$

Таблица 14.12. Шаги каждого цикла алгоритма IDEA

1. $M_1 \times Z_1^{(R)}$.
2. $M_2 \times Z_2^{(R)}$.
3. $M_3 \times Z_3^{(R)}$.
4. $M_4 \times Z_4^{(R)}$.
5. К результатам шагов (1) и (3) применяется операция XOR[3].
6. К результатам шагов (2) и (4) — операция XOR.
7. Результат шага (5) умножается на $Z_5^{(R)}$.
8. Складываются результаты шагов (6) и (7).
9. Результат шага (8) умножается на $Z_6^{(R)}$.
10. Складываются результаты шагов (7) и (9).
11. К результатам шагов (1) и (9) применяется операция XOR.
12. К результатам шагов (3) и (9) — операция XOR.
13. К результатам шагов (2) и (10) — операция XOR.
14. К результатам шагов (4) и (10) — операция XOR.

Пример 14.8. Первый цикл шифра IDEA

Пусть сообщение (слово "HI") сначала нужно записать в шестнадцатеричной форме. Начнем с ASCII-кода, представленного на рис. 2.3, на котором бит 1 представляет собой самый младший разряд. Затем добавим равный нулю восьмой бит старшего разряда, который обычно используется

[3] Операция XOR (исключающее ИЛИ) определяется следующим образом: 0 XOR 0 = 0, 0 XOR 1 = 1, 1 XOR 0 = 1, 1 XOR 1 = 0.

для проверки четности, и выполним необходимое преобразование, взяв по четыре бита (порядок — от старшего разряда до младшего). Таким образом, буква "H" в сообщении преобразуется в 0048, а буква "I" — в 0049. Для этого примера выберем 128-битовый ключ K_0, выраженный восемью группами *подключей* из 4-разрядных шестнадцатеричных чисел: K_0 = 0008 0007 0006 0005 0004 0003 0002 0001, где крайний правый подключ представляет самый младший разряд. Используя этот ключ и шифр IDEA, найдите выход цикла 1.

Решение

Сначала сообщение делится на 64-битовые блоки данных. Каждый из этих блоков затем делится на подблоки M_i, где i = 1, ..., 46, каждый из которых содержит 16-битовые или 4-значные шестнадцатеричные цифры. В этом примере длина сообщения "HI" равна всего 16 бит; следовательно, (используя шестнадцатеричное обозначение) M_1 = 4849 и $M_2 = M_3 = M_4$ = 0000. Сложение производится по модулю 2^{16}, а умножение — по модулю $2^{16} + 1$. 128-битовый ключ, определенный для первого цикла, делится на восемь 16-битовых подключей, начиная с младшей группы шестнадцатеричных кодов: $Z_1^{(1)}$ = 0001, $Z_2^{(1)}$ = 0002, $Z_3^{(1)}$ = 0003, $Z_4^{(1)}$ = 0004, $Z_5^{(1)}$ = 0005, $Z_6^{(1)}$ = 0006, $Z_1^{(2)}$ = 0007 и $Z_2^{(2)}$ = 0008.
Шаги, обозначенные в табл. 14.11, дают следующие результаты.

1. $M_1 \times Z_1$ = 4849 × 0001 = 4849.
2. $M_2 \times Z_2$ = 0000 + 0002 = 0002.
3. $M_3 \times Z_3$ = 0000 + 0003 = 0003.
4. $M_4 \times Z_4$ = 0000 × 0004 = 0000.
5. К результатам шагов (1) и (3) применяется операция XOR, в результате чего получится следующее: 4849 XOR 0003 = 484A.

```
  0100 1000 0100 1001  (4849 из шестнадцатеричной системы переведено в двоичную)
XOR 0000 0000 0000 0011  (0003 из шестнадцатеричной системы переведено в двоичную)
---------------------

  0100 1000 0100 1010
```

Обратное преобразование в шестнадцатеричную систему дает следующее: 484A (где A — шестнадцатеричное обозначение двоичного числа 1010).

6. К результатам шагов (2) и (4) применяется операция XOR: 0002 XOR 0000 = 0002.
7. Результат шага (5) умножается на Z_5: 484A × 0005 = 6971.
8. Результаты шагов (6) и (7) складываются: 0002 + 6971 = 6973.
9. Результат шага (8) умножается на Z_6: 6973 × 0006 = 78B0.
10. Результаты шагов (7) и (9) складываются: 6971 + 78B0 = E221.
11. К результатам шагов (1) и (9) применяется операция XOR: 4849 XOR 78B0 = 30F9.
12. К результатам шагов (3) и (9) применяется операция XOR: 0003 XOR 78B0 = 78B3
13. К результатам шагов (2) и (10) применяется операция XOR: 0002 XOR E221 = E223.
14. К результатам шагов (2) и (10) применяется операция XOR: 0000 XOR E221 = E221.

Выход цикла 1 (результат шагов 11–14): 30F9 78B3 E223 E221. Перед началом цикла 2 переставляются два внутренних слова выхода цикла 1. Затем производится еще семь циклов и выполняется окончательное выходное преобразование.

14.6.2. Алгоритмы Диффи-Хэллмана (вариант Элгемала) и RSA

Для шифрования ключа сеанса PGP предлагает на выбор два алгоритма ключа шифрования общего доступа, RSA и протокол Диффи-Хэллмана (Diffie-Hellman) (вариант Элгемала (Elgamal)). Для алгоритмов RSA и Диффи-Хэллмана допустимый размер ключа составляет от 1024 до 4096 бит. Ключ размером 1024 бит считается безопасным для большинства сеансов обмена информацией. Защищенность алгоритма RSA (см. раздел 14.5.3) основана на сложности разложения на множители больших чисел.

Протокол Диффи-Хэллмана был разработан Вайтфилдом Диффи (Whitefield Diffie), Мартином Е. Хэллманом (Martin E. Hellman) и Ральфом С. Мерклем (Raph C Merkle) в

1976 году [19, 20] для обмена информацией по незащищенному каналу с помощью открытого ключа. Данный протокол основан на сложности задачи нахождения дискретного логарифма для конечных полей [21]. Он предполагает, что вычислить g^{ab}, зная только g^a и g^b, практически невозможно. Патент №4 200 770 (США), срок которого истек в 1997 году, содержит протокол Диффи-Хэллмана и его разновидности, такие как вариант Элгемала. Данный вариант, разработанный Тахером Элгемалом (Taher Elgamal), расширяет протокол Диффи-Хэллмана на шифрование сообщений. В PGP вариант Элгемала алгоритма Диффи-Хэллмана применяется для шифрования ключа сеанса.

14.6.2.1. Описание алгоритма Диффи-Хэллмана, вариант Элгемала

Протокол имеет два системных параметра n и g, которые являются общедоступными. Параметр n — это большое простое число, а параметр g — целое число, меньшее n, которое обладает следующим свойством: для любого числа p, лежащего между 1 и $n-1$ включительно, существует степень k числа g, при которой $g^k = p \bmod n$. Ниже описывается схема шифрования Элгемала [19, 21], позволяющая пользователю B посылать сообщение пользователю A.

- Пользователь A случайным образом выбирает большое целое число a (это частный ключ пользователя A).
- Открытый ключ пользователя A вычисляется следующим образом: $y = g^a \bmod n$.
- Пользователь B желает послать пользователю A сообщение M. Сначала пользователь B генерирует случайное число k, меньшее n.
- Пользователь B вычисляет следующие величины:
  $y_1 = g^k \bmod n$
  $y_2 = M \times (y^k \bmod n)$ (напомним, что y — это открытый ключ пользователя A).
- Пользователь B посылает пользователю A шифрованный текст (y_1, y_2).
- После получения шифрованного текста (y_1, y_2) пользователь A вычисляет открытое сообщение M.

$$M = \frac{y_2}{y_1^a \bmod n}$$

Пример 14.9. Применение алгоритма Диффи-Хэллмана (вариант Элгемала) для шифрования сообщения

Пусть общедоступными системными параметрами являются $n = 11$ и $g = 7$. Предположим, что пользователь A в качестве частного ключа выбрал $a = 2$. Покажите, как вычисляется открытый ключ пользователя A. Покажите также, как пользователь B будет шифровать сообщение $M = 13$, которое должно быть отправлено пользователю A, и как пользователь A последовательно дешифрует полученный шифрованный текст.

Решение

Открытый ключ пользователя A ($y = g^a \bmod n$) вычисляется следующим образом: $y = 7^2 \bmod 11 = 5$. Пользователь B желает послать пользователю A сообщение $M = 13$. В данном примере пусть пользователь B в качестве случайного значения k (меньшего $n = 11$) выбирает $k = 1$. Далее пользователь B вычисляет шифрованную пару.

$$y_1 = g^k \bmod n = 7^1 \bmod 11 = 7$$

$$y_2 = M \times (y^k \bmod n) = 13 \times (5^1 \bmod 11) = 13 \times 5 = 65$$

Пользователь A получает шифрованный текст (7, 65) и вычисляет сообщение M.

$$M = \frac{y_2}{y_1^a \bmod n} = \frac{65}{7^2 \bmod 11} = \frac{65}{5} = 13$$

14.6.3. Шифрование сообщения в системе PGP

Алгоритмы с частным ключом, применяемые PGP для шифрования сообщения, были представлены в разделе 14.6.1. Алгоритмы с открытым ключом, используемые PGP для шифрования ключа частного сеанса, были представлены в разделе 14.6.2. Чтобы проиллюстрировать технологию шифрования PGP, изображенную на рис. 14.20, рассмотрим следующий пример, объединяющий алгоритмы двух типов.

Пример 14.10. Использование алгоритмов RSA и IDEA для шифрования в PGP

Для шифрования ключа сеанса используем алгоритм RSA с открытым ключом с параметрами из раздела 14.5.3.1: $n = pq = 2773$, ключ шифрования $e = 17$, а ключ дешифрования $d = 157$. Ключом шифрования является открытый ключ получателя, а ключом дешифрования — частный ключ получателя. Используем ключ сеанса K_0 = 0008 0007 0006 0005 0004 0003 0002 0001 и шифрованный текст 30F9 78B3 E223 E2216, представляющий сообщение "HI", из примера 14.8 (все величины представлены в шестнадцатеричной записи). (Отметим, что шифрованный текст был создан с использованием только одного цикла алгоритма IDEA. В реальной системе производится 8 циклов плюс выходное преобразование.) Зашифруйте ключ сеанса и покажите, какое сообщение должно передаваться.

Решение

Следуя описанию, приведенному в разделе 14.5.3.1, ключ сеанса будет шифроваться с помощью алгоритма RSA с открытым ключом получателя 17. Для удобства вычисления при помощи простого калькулятора преобразуем сначала ключ сеанса в группы, составленные из величин в десятичной записи. Согласно требованиям алгоритма RSA, значения, приписанные каждой группе, не должны превышать $n - 1 = 2772$. Следовательно, выразим 128-битовый ключ в терминах 4-разрядных групп, где самая старшая (самая левая) группа будет представлять 7 бит и 11 групп будут представлять 11 бит каждая. Преобразование чисел из шестнадцатеричных в десятеричные можно рассматривать как двухэтапный процесс: (1) преобразование в двоичную систему и (2) переход к основанию 10. В результате получаем K_0 = 0000 0032 0000 1792 0048 0001 0512 0064 0001 1024 0064 0001. Напомним, из уравнения (14.32) следует, что $C = (M)^e$ по модулю n, где M — одна из 4-разрядных групп K_0. Левые крайние четыре группы шифруются следующим образом.

$$C_{12} = (0000)^{17} \bmod 2773 = 0.$$

$$C_{11} = (0032)^{17} \bmod 2773 = 2227.$$

$$C_{10} = (0000)^{17} \bmod 2773 = 0.$$

$$C_9 = (1792)^{17} \bmod 2773 = 2704.$$

Эффективным способом модульного возведения в степень является использование алгоритма "Возведение в квадрат и умножение" (Square-and-Multiply — SM). Этот алгоритм [21] сводит число необходимых модульных умножений с $e - 1$ почти до $2l$, где l — число бит в двоичном представлении. Покажем использование алгоритма SM, шифруя одну из десятичных групп ключа сеанса (одиннадцатую группу справа M_{11} = 0032), где $n = 2773$ и $e = 17$. Для применения этого алгоритма сначала запишем число e в его двоичном представлении ($17_{10} = 10001_2$).

Вычисления даны в табл. 14.13. Используется математика по модулю n, в этом примере $n = 2773$. Второй столбец содержит двоичный код, где старший бит находится в строке 1. Каждая двоичная величина в этом столбце используется для контроля результата в столбце 3. Начальное значение, расположенное в столбце 3, строка 0, всегда равно 1. Далее результат в каждой строке столбца 3 зависит от бита в соответствующей строке столбца 2. Если этот бит равен "1", то результат предыдущей строки возводится в квадрат и умножается на открытый текст (для этого примера — 32). Если строка во втором столбце содержит "0", то значение соответствующей строки в столбце 3 равно квадрату значения в предыдущей строке. Окончательным значением является зашифрованный текст ($C = 2227$). Повторение этого метода для каждой из двенадцати десятичных групп, составляющих K_0, дает шифрованный текст ключа сеанса: C = 0000 2227 0000 2704 0753 0001 1278 0272 0001 1405 0272 0001. Этот ключ сеанса (здесь он представлен в десятичной форме), зашифрованный с помощью алгоритма RSA, вместе с зашифрованным с помощью IDEA сообщением вида 30F9 78B3 E223 E221 (здесь оно представлено в шестнадцатеричной форме) может теперь передаваться через незащищенный канал.

Таблица 14.13. Алгоритм SM с открытым текстом = 32

Номер строки	Двоичное представление e (первым идет старший разряд)	Модульное умножение (модуль 2773)
0		1
1	1	$1^2 \times 32 = 32$
2	0	$32^2 = 1024$
3	0	$1024^2 = 328$
4	0	$328^2 = 1728$
5	1	$1728^2 \times 32 = 2227$

14.6.4. Аутентификация с помощью PGP и создание подписи

Алгоритмы с открытыми ключами могут использоваться для проверки подлинности (аутентификации) или "подписания" сообщения. Как показано на рис. 14.18, отправитель может шифровать документ с помощью своего частного ключа (к которому никто больше не имеет доступа), а затем с помощью открытого ключа получателя. Получатель должен сначала использовать свой частный ключ для дешифрования сообщения. Затем должно последовать второе дешифрование, при котором используется открытый ключ отправителя. С помощью этой технологии засекречивается сообщение, а также обеспечивается проверка подлинности отправителя.

Поскольку алгоритмы с открытыми ключами работают достаточно медленно, PGP допускает разные методы проверки подлинности отправителя. Вместо трудоемкого процесса шифрования всего открытого сообщения, PGP предлагает шифрование профиля сообщения (message digest) фиксированной длины, созданного с помощью односторонней хэш-функции. Шифрование профиля сообщения производится посредством алгоритма открытого ключа. Этот метод, называемый *цифровой подписью*, изображен на рис. 14.22. Цифровая подпись используется для проверки подлинности как *отправителя*, так и *сообщения*. Проверка подлинности сообщения обеспечивает проверку того, что сообщение не было некоторым образом изменено. Данная технология основана на том, что если сообщение было изменено (т.е. было постороннее вмешательство), его профиль будет другим.

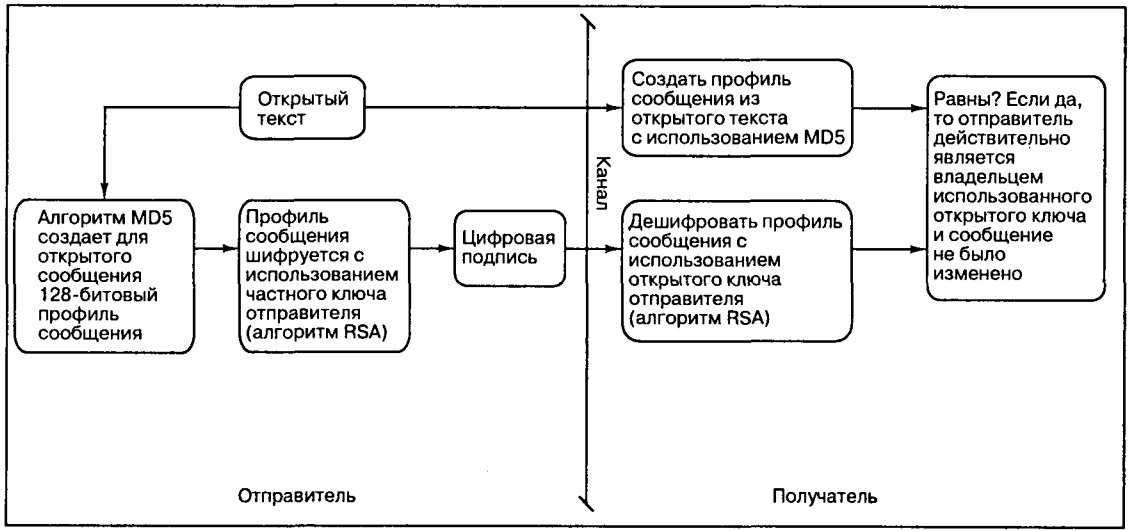


Рис. 14.22. Технология создания подписи, используемая PGP

PGP версии 2.6 использует алгоритм MD5 (Message Digest 5) для создания 128-битового профиля сообщения (или значения хэш-функции) открытого текста. Затем значение хэш-функции шифруется с помощью частного ключа отправителя и посылается с открытым текстом. Когда получатель принимает сообщение, он сначала дешифрует профиль сообщения, используя открытый ключ отправителя. Затем получатель действует на открытый текст хэш-функцией и сравнивает два профиля сообщения. Если они совпадают, подпись подлинная. На рис. 14.22 сообщение отправляется без шифрования (как открытый текст); впрочем, оно может быть зашифровано с помощью метода, изображенного на рис. 14.20.

14.6.4.1. MD5 и SHA-1

MD5 и SHA-1 являются хэш-функциями. Вообще, хэш-функция $H(x)$ принимает аргумент и возвращает строку h фиксированного размера, называемую значением хэш-функции (или профилем сообщения). Криптографическая хэш-функция обладает следующими свойствами.

1. Длина выхода фиксированна.
2. Значение хэш-функции относительно просто вычисляется.
3. Функция является односторонней; другими словами, ее трудно обратить. Для данного значения h вычислительно неосуществимо найти аргумент функции x.
4. Функция является бесконфликтной; таковой называется функция, для которой два разных аргумента не могут порождать одно и то же значение.

Алгоритм MD-5, используемый PGP версии 2.6, создает 128-битовый профиль сообщения. За четыре цикла данный алгоритм разбивает текст на 512-битовые блоки. В каждом цикле используются разные нелинейные функции, включающие логические операторы И, ИЛИ, НЕ или исключающее ИЛИ. За цикл каждая функция применяется 16 раз. Кроме того, в каждом цикле используются сдвиги битов и скалярное сложение [19]. Ганс Доббертин (Hans Dobbertin) [18] определил, что в MD-5 возможны конфликты. В силу этих потенциальных недостатков PGP рекомендует Стандарт цифровой подписи (Digital Signature Standard — DSS), который использует алгоритм SHA-1 (Secure Hash Algorithm-1). Данный алгоритм (SHA-1) берет сообщение, длиной меньше 2^{64} бит, и создает 160-битовый профиль сообщения. Алгоритм SHA-1 подобен MD-5 тем, что в каждом из 4 циклов используются различные нелинейные функции. В SHA-1 каждая функция применяется 20 раз в течение цикла. Кроме того, в SHA-1 используются разные скалярные сложения и сдвиги битов. Алгоритм имеет более медленное действие, чем MD-5, но больший профиль сообщения (160 бит в отличие от 128 бит) делает его более защищенным от криптоаналитических атак по методу грубой силы [19]. Метод грубой силы — это попытка подобрать профиль сообщения путем перебора входных комбинаций.

14.6.4.2. Стандарт цифровой подписи и алгоритм RSA

При создании цифровых подписей PGP версии 2.6 использует алгоритм RSA для шифрования значения, производимого хэш-функцией MD-5. Однако в версиях 5.0 и более поздних применяется стандарт цифровой подписи (DSS) института NIST [22]. Данный стандарт требует использования хэш-функции SHA-1. Значение этой функции затем шифруется с помощью алгоритма цифрового стандарта DSA (Digital Standard Algorithm). Подобно протоколу Диффи-Хэллмана, DSA основан на задаче взятия дискретного логарифма. (Подробно об алгоритме DSA рассказано в работе [22]).

14.7. Резюме

В этой главе представлены основные модели криптографического процесса и рассмотрены его цели. Здесь описаны некоторые ранние системы шифрования и рассмотрена математическая теория секретного общения, учрежденная Шенноном. Описана также система, которая может представлять совершенную секретность, и показано, что такие системы могут быть реализованы, но их использование не является приемлемым там, где требуется интенсивное общение. Кроме того, в данной главе рассмотрены системы с практической защищенностью, использующие технологии Шеннона (известные как смешение и диффузия), которые позволяют предотвращать статистические попытки криптоаналитиков.

Результаты работы Шеннона были воплощены IBM в системе LUCIFER, которая позднее переросла в Стандарт шифрования данных (Data Encryption Standard — DES) Национального бюро стандартов (National Bureau of Standards). Здесь подробно описан алгоритм DES. Рассмотрено также применение в системах поточного шифрования линейных регистров сдвига с обратной связью. Продемонстрирована внутренняя уязвимость регистров, использующих генератор ключей.

В данной главе описаны криптосистемы с открытыми ключами и рассмотрены две схемы — Ривеста-Шамира-Адельмана (RSA), основанная на использовании произведения двух больших простых чисел, и Меркла-Хэллмана, основанная на классической задаче о рюкзаке. В заключение была описана схема PGP, разработанная Филиппом Циммерманом (опубликована в 1991 году). PGP использует преимущества обеих систем — системы с частным ключом и системы с открытым ключом. Доказано, что применение этой системы представляет собой важный метод шифрования файлов, используемый для пересылки данных по электронной почте.

Литература

1. Kahn D. *The Codebreakers.* Macmillan Publishing Company, New York, 1967.
2. Diffie W. and Hellman M. E. *Privacy and Authentication: An Introduction to Cryptography.* Proc. IEEE, vol. 67, n. 3, March, 1979, pp. 397–427.
3. Beker H. and Piper F. *Cipher Systems.* John Wiley & Sons, Inc., New York, 1982.
4. Denning D. E. R. *Cryptography and Data Security.* Addison-Wesley Publishing Company, Reading, Mass, 1982.
5. Shannon C. E. *Communication Theory of Secrecy Systems.* Bell Syst. Tech. J., vol. 28, October, 1949, pp. 656–715.
6. Hellman M. E. *An Extension of the Shannon Theory Approach to Cryptography.* IEEE Trans. Inf. Theory, vol. IT23, May, 1978, pp. 289–294.
7. Smith J. L. *The Design of Lucifer, a Cryptographic Device for Data Communications.* IBM Research Rep. RC-3326, 1971.
8. Feistel H. *Cryptography and Computer Privacy.* Sci. Am., vol. 228, n. 5, May, 1973, pp. 15–23.
9. National Bureau of Standards. *Data Encryption Standard.* Federal Information Processing Standard (FIPS), Publication n. 46, January, 1977.
10. United States Senate Select Committee in Intelligence. *Unclassified Summary: Involvement of NSA in the Development of the Data Encryption Standard.* IEEE Commun. Soc. Mag., vol. 16, n. 6, November, 1978, pp. 53–55.
11. Stallings W. *Cryptography and Network Security.* Second Addition, Prentice Hall, Upper Saddle River, NJ. 1998. (Столлингс В. *Криптография и защита сетей. Принципы и практика*, 2-е издание. М.: — Издательский дом "Вильямс", 2001. — 672 с.)

12. Diffie W. and Hellman M. E. *New Directions in Cryptography.* IEEE Trans. Inf. Theory, vol. 1722, November, 1976, pp. 644–654.
13. Rivest R. L., Shamir A. and Adelman L. *On Digital Signature and Public Key Cryptosystems.* Commun. ACM. Vol. 21, February, 1978, pp.120–126.
14. Knuth D. E. *The Art of Computer Programming,* Vol. 2, *Seminumerical Algorithms.* 2nd ed., Addison-Wesley Publishing Company, Reading, Mass, 1981. (Кнут Д. *Искусство программирования, т. 2. Получисленные алгоритмы,* 3-е издание. — М.: Издательский дом "Вильямс", 2000. — 832 с.)
15. Mercal R. C. and Hellman M. E. *Hiding Information and Signatures in Trap-Door Knapsacks.* IEEE, Trans. Inf. Theory, vol. IT24, September, 1978, pp. 525–530.
16. Shamir A. *A Polynomial Time Algorithm for Breaking the Basic Merkle-Hellman Cryptosystems.* IEEE 23rd Ann. Symp. Found. Comput. Sci., 1982, pp. 145–153.
17. Zimmerman P. *The Official PGP User's Guide.* MIT Press, Cambridge, 1995.
18. *PGP Freeware User's Guide, Version 6.5.* Network Associates, Inc., 1999.
19. Schneier B. *Applied Cryptography.* John Wiley & Sons, New York, 1996.
20. Hellman M. E., Martin, Bailey, Diffie, W. and Merkle R. C. *United States Patent 4,200,700: Cryptographic Apparatus and Method.* United States Patent and Trademark Office, Washington, DC, 1980.
21. Stinson, Douglas. *Cryptography Theory and Practice.* CRC Press, Boca Raton, FL, 1995.
22. *Digital Signature Standard* (Federal Information Processing Standards Publication 186-1). Government Printing Office, Springfield, VA, December, 15, 1998.

Задачи

14.1. Пусть X — целая переменная, представленная 64 бит. Вероятность попадания X в интервал $(0, 2^{16} - 1)$ равна 1/2, вероятность попадания X в интервал $(2^{16}, 2^{32} - 1)$ — 1/4, а вероятность попадания X в интервал $(2^{32}, 2^{64} - 1)$ — 1/4. Внутри каждого интервала значения равновероятны. Вычислите энтропию X.

14.2. Существует множество равновероятных сообщений о погоде: солнечно (С), пасмурно (П), небольшой дождь (Д), ливень (Л). При наличии дополнительной информации о времени дня (утро или день) вероятности изменяются следующим образом.

$$\text{Утро: } P(\text{С}) = \tfrac{1}{8},\ P(\text{П}) = \tfrac{1}{8},\ P(\text{Д}) = \tfrac{3}{8},\ P(\text{Л}) = \tfrac{3}{8}$$

$$\text{День: } P(\text{С}) = \tfrac{3}{8},\ P(\text{П}) = \tfrac{3}{8},\ P(\text{Д}) = \tfrac{1}{8},\ P(\text{Л}) = \tfrac{1}{8}$$

а) Найдите энтропию сообщения о погоде.

б) Найдите энтропию сообщения при указании времени дня.

14.3. Гавайский алфавит состоит только из 12 букв — гласные a, e, i, o, u и согласные h, k, l, m, n, p, w. Предположим, что каждая гласная встречается с вероятностью 0,116, а каждая согласная — с вероятностью 0,06. Предположим также, что среднее число *бит информации,* попадающих на каждую букву, такое же, как и в английском языке. Вычислите расстояние единственности для зашифрованного гавайского сообщения, если ключевая последовательность состоит из случайной перестановки 12 букв алфавита.

14.4. Оцените расстояние единственности англоязычной системы шифрования, которая использует ключевую последовательность, составленную из 10 случайных символов алфавита.

а) Каждый ключевой символ может представлять собой одну из 26 букв алфавита (повторения допускаются).

б) Ключевые символы не могут повторяться.

14.5. Решите задачу 14.4, когда ключевая последовательность составлена из десяти целых чисел, случайно выбранных из множества 0–999.

14.6. а) Найдите расстояние единственности для системы DES, которая шифрует 64-битовые блоки (восемь символов алфавита) с помощью 56-битового ключа.

б) Как отразится на расстоянии единственности увеличение ключа до 128 бит?

14.7. На рис. 14.8 и 14.9 чередуются *P*- и *S*-блоки. Является ли это более безопасным, нежели если бы сначала были сгруппированы все *P*-блоки, а затем все *S*-блоки? Ответ аргументируйте.

14.8. Каким будет выход первой итерации алгоритма DES, если и открытый текст, и ключ составлены из нулевых последовательностей?

14.9. Рассмотрим открытое 10-битовое сообщение в виде последовательности 0101101001 и соответствующую ему последовательность шифрованного текста 0111011010, где крайний правый бит является самым ранним. Опишите пятиразрядный линейный регистр сдвига с обратной связью, производящий ключевую последовательность, и укажите начальное состояние регистра. Имеет ли выходная последовательность максимальную длину?

14.10. Используя параметры примера 14.5 и следуя алгоритму RSA, вычислите ключ шифрования e, если в качестве ключа дешифрования выбрано число 151.

14.11. Даны e и d, такие, что ed по модулю $\phi(n) = 1$, и сообщение, которое зашифровано как целое число M из интервала $(0, n-1)$, такое что НОД$(M, n) = 1$. Докажите, что $(M^e$ по модулю $n)^d$ по модулю $n = M$.

14.12. Используйте схему RSA для шифрования сообщения $M = 3$. В качестве простых чисел возьмите $p = 5$ и $q = 7$. Ключ дешифрования $d = 11$. Вычислите значение ключа шифрования e.

14.13. Используется схема RSA.

а) Пусть простыми числами являются $p = 7$ и $q = 11$. Перечислите пять допустимых значений ключа дешифрования d.

б) Пусть простыми числами являются $p = 13$ и $q = 31$, а ключ дешифрования $d = 37$. Найдите ключ шифрования e и опишите его использование для шифрования слова "DIGITAL".

14.14. Используйте схему Меркла-Хэллмана с открытым ключом и быстровозрастающим вектором $\mathbf{a}' = 1, 3, 5, 10, 20$. Воспользуйтесь следующими дополнительными параметрами: большое простое число $M = 51$, случайное число $W = 37$.

а) Найдите небыстровозрастающий вектор $\mathbf{a}$, который следует сделать общедоступным, и зашифруйте вектор данных 11011.

б) Покажите этапы дешифрования текста разрешенным получателем.

14.15. С помощью протокола Диффи-Хэллмана (вариант Элгемала) зашифруйте сообщение $M = 7$. Параметры системы: $n = 17$ и $g = 3$. Частный ключ получателя: $a = 4$. Определите открытый ключ получателя. Для шифрования сообщения со случайно выбранным k используйте $k = 2$. Проверьте точность данного шифрования с помощью частного ключа получателя.

14.16. Найдите шестнадцатеричное значение сообщения "no" после одного цикла алгоритма IDEA. Ключ сеанса (шестнадцатеричная запись) = 0002 0003 0002 0003 0002 0003 0002 0003, где крайняя правая 4-разрядная группа представляет подключ Z_1. Пусть каждый символ ASCII для сообщения "no" представлен 16-битовым подблоком данных, где "n" = 006E и "o" = 006F.

14.17. В примере 14.10 ключ сеанса для алгоритма IDEA шифруется с использованием алгоритма RSA. Результирующим ключом сеанса шифрования (в десятичной записи) был 0000 2227 0000 2704 0753 0001 1278 0272 0001 1405 0272 0001, где наименее значимой (крайней правой) группой является 1. Используя ключ дешифрования, дешифруйте группу 11 этого ключа сеанса с помощью алгоритма SM (см. пример 14.10).

Вопросы для самопроверки

14.1. В чем состоят два основных требования, предъявляемые к полезным *криптосистемам* (см. раздел 14.1.2)?

14.2. Шеннон предложил две концепции шифрования — *смешение* (confusion) и *диффузия* (diffusion). Объясните значение этих терминов (см. раздел 14.3.1).

14.3. Объясните, почему при необходимости *высокого уровня секретности* не должен использоваться линейный регистр сдвига с обратной связью (см. раздел 14.4.2)?

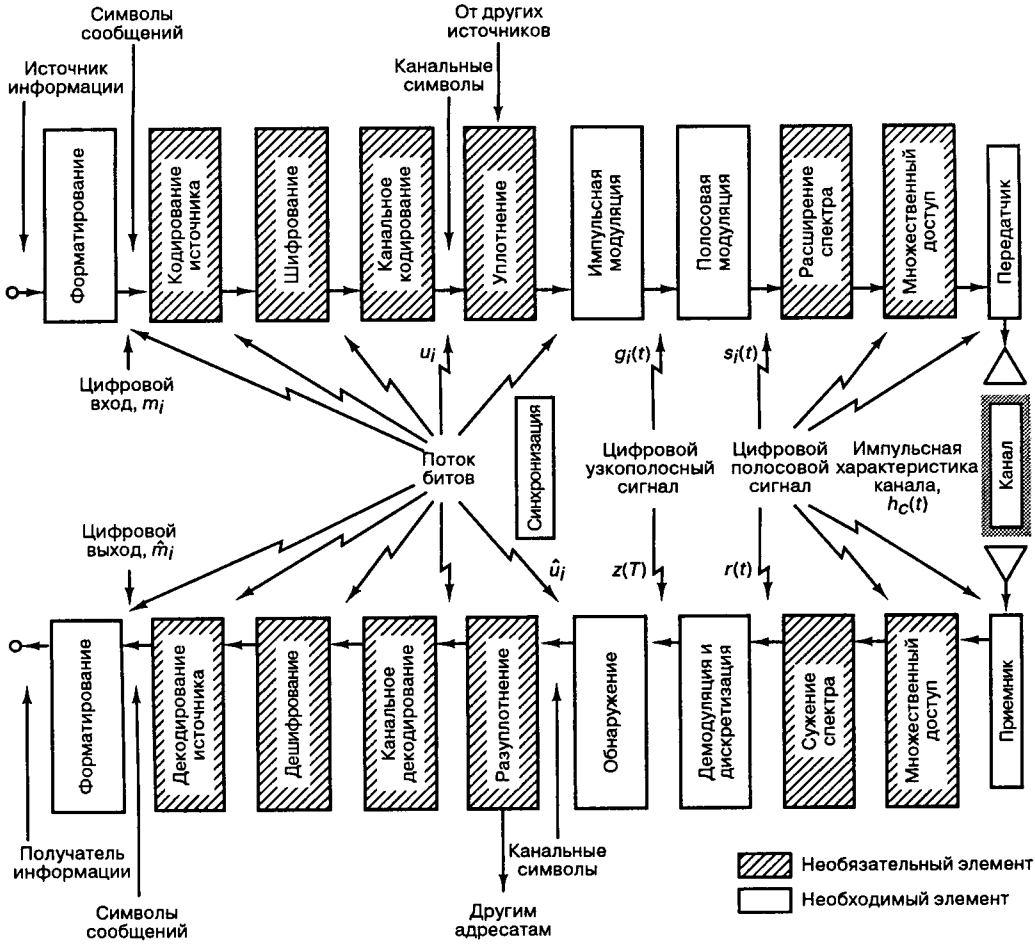
14.4. Объясните основное отличие между общепринятыми криптосистемами и *криптосистемами с открытым ключом* (см. раздел 14.5).

14.5. Опишите шаги шифрования сообщения при использовании *стандарта шифрования данных* (DES). Насколько отличаются эти операции при "тройном" DES (см. разделы 14.3.5 и 14.6.1.1)?

14.6. Опишите этапы шифрования сообщения с помощью PGP версии 2.6 (см. раздел 14.6.1.3).

ГЛАВА 15

Каналы с замираниями



В 1950–60-е годы впервые были смоделированы механизмы, приводящие к замиранию в каналах связи; они преимущественно применялись к тропосферной связи, охватывающей широкий диапазон частот. Примерами каналов, в которых наблюдаются явления замирания, могут служить диапазон высоких частот (high-frequency — HF) (3–300 МГц), используемый для передач через ионосферу, и диапазон ультравысоких частот (ultra-high-frequency — UHF) (300 МГц–3 ГГц) с диапазоном сверхвысоких частот (super-high-frequency — SHF) (3–30 ГГц), используемые при передаче сигналов через тропосферу. Несмотря на то что эффекты замирания в каналах радиосвязи с подвижными объектами несколько отличаются от встречающихся в ионосферных и тропосферных каналах, ранние модели все же вполне приемлемы для описания эффектов замирания в системах мобильной цифровой связи. В этой главе особое внимание уделяется так называемому *релеевскому замиранию* (Rayleigh fading) преимущественно в диапазоне УВЧ, которое воздействует на такие мобильные системы связи, как сотовые и персональные (personal communication systems — PCS). Кроме того, особое внимание уделяется основным проявлениям замирания, типам ухудшения характеристик и методам борьбы с ухудшением характеристик. Рассматриваются два примера характерных методов борьбы: использование эквалайзера Витерби, реализованного в системе GSM (Global System for Mobile — глобальная система мобильной связи), и RAKE-приемника (RAKE receiver), применяемого в системах CDMA, разработанных согласно требованиям стандарта Interim Standard-95 (IS-95).

15.1. Сложности связи по каналу с замираниями

При анализе характеристик систем связи отправной точкой является описание основных характеристик в классическом (идеальном) канале с белым аддитивным гауссовым шумом (additive white Gaussian noise — AWGN) со статистически независимыми гауссовыми шумовыми выборками, искажающими информационные выборки, и отсутствием межсимвольной интерференции (intersymbol interference — ISI). Основным источником ухудшения характеристик является тепловой шум, генерируемый в приемнике. Другим источником потерь являются естественные и искусственные источники шума и помех, воздействие которых на принимающую антенну можно качественно описать через параметр, называемый *температурой антенны* (см. раздел 5.5.5). Тепловой шум имеет, как правило, плоскую спектральную плотность мощности по всей полосе сигнала и гауссову функцию плотности вероятности напряжения с нулевым средним. В системах мобильной связи внешние шумы и помехи часто оказываются более значительными, чем тепловой шум приемника. При моделировании реальных систем следующим шагом является введение полосовых фильтров. Обычно фильтрация в передатчике служит для удовлетворения некоторых условий к спектральным составляющим. Фильтрация в приемнике часто является результатом применения согласованного фильтра, о чем говорилось в разделе 3.2.2. Из-за ограниченности полосы частот и фазовых искажений в фильтрах для снижения ISI, вызываемой фильтром, может потребоваться специальная обработка сигнала и его выравнивание.

Если характеристики радиоканала не заданы, то обычно подразумевается, что сигнал затухает с расстоянием так же, как при распространении в идеальном свободном пространстве. В модели свободного пространства область между антеннами передатчика и приемника предполагается свободной от объектов, которые могли бы поглощать или отражать энергию на радиочастотах. Предполагается также, что внутри этой области атмосфера ведет себя как совершенно однородная непоглощающая среда. Кроме того, считает-

ся, что земля находится бесконечно далеко от распространяемого сигнала (или, что равносильно, имеет пренебрежимо малый коэффициент отражения). По существу, в этой идеализированной модели свободного пространства ослабление между передатчиком и приемником радиочастотной энергии происходит по закону обратных квадратов. Мощность приемника, выраженная через переданную мощность, ослабляется в $L_s(d)$ раз, причем данный параметр называется *потерями в тракте* (path loss), или *потерями в свободном пространстве* (free space loss). Если антенна приемника изотропна, то этот коэффициент выражается следующим образом (см. раздел 5.3.1.1).

$$L_s(d) = \left(\frac{4\pi d}{\lambda}\right)^2 \tag{15.1}$$

Здесь d — это расстояние между передатчиком и приемником, а λ — длина волны распространяемого сигнала. При таком идеальном распространении мощность полученного сигнала весьма предсказуема. Для большинства реальных каналов, в которых распространение происходит в атмосфере и вблизи поверхности земли, модель распространения в свободном пространстве неадекватно описывает поведение канала и не позволяет предсказывать характеристики системы. В системах мобильной радиосвязи сигнал может передаваться от передатчика к приемнику по множеству отражательных путей. Это явление, называемое *многолучевым распространением* (multipath propagation), может вызывать флуктуации амплитуды, фазы и угла прибытия полученного сигнала, что определило название *замирание вследствие многолучевого распространения* (multipath fading). Другое название — *сцинтилляция* (scintillation) — которое происходит из радиоастрономии, используется для описания замирания, вызванного физическими изменениями в среде распространения, такими как изменение электронной плотности слоев ионосферы, которые отражают высокие частоты радиосигналов. Как замирание, так и сцинтилляция относится к случайным флуктуациям сигнала; основное отличие заключается в том, что явление сцинтилляции объясняется механизмами, существенными на расстояниях, намного меньших длины волны (например, движение электронов). Прямое моделирование и проектирование систем, включающих методы борьбы с замиранием, обычно сложнее разработки систем, где единственным источником ухудшения рабочих характеристик считается шум AWGN.

15.2. Описание распространения радиоволн в мобильной связи

На рис. 15.1 представлен обзор проявления эффектов замирания в каналах. Он начинается с двух типов эффектов замирания, характерных для мобильной связи: крупномасштабное и мелкомасштабное замирание. Крупномасштабное замирание отражает среднее ослабление мощности сигнала или потери в тракте вследствие распространения на большое расстояние. На рис. 15.1 проявления крупномасштабного замирания показаны в блоках 1–3. На это явление влияют выступающие наземные элементы (например холмы, леса, рекламные щиты, группы строений и т.д.) между передатчиком и приемником. Часто говорят, что приемник "затеняется" этими выступами. Статистика крупномасштабного замирания позволяет приблизительно рассчитать потери в тракте как функцию расстояния. Это часто описывается через средние потери в тракте (степенной закон n-го порядка) и логарифмически нормально распределенные

отклонения от среднего. Мелкомасштабное замирание — это значительные изменения амплитуды и фазы сигнала, которые на практике могут быть результатом небольших изменений (порядка половины длины волны) расстояния между передатчиком и приемником. Как указано на рис. 15.1 (блоки 4–6), мелкомасштабное замирание проявляется двумя способами — расширение сигнала во времени (или дисперсия сигнала) и нестационарное поведение канала. В мобильной радиосвязи параметры каналов изменяются во времени, поскольку движение передатчика и/или приемника приводит в результате к изменению пути распространения. Скорость изменения таких условий распространения определяет скорость замирания (скорость изменения ухудшения характеристик вследствие замирания). Мелкомасштабное замирание называется *релеевским* , если имеется большое число многократно отражающихся путей и нет компонента сигнала вдоль луча обзора; огибающая такого полученного сигнала статистически описывается с помощью релеевской функции плотности вероятности. Если преобладает незамирающий компонент сигнала, такой как путь распространения вдоль луча обзора, огибающая мелкомасштабного замирания описывается функцией плотности вероятности Райса [1]. Иными словами, статистики мелкомасштабного замирания всегда распределены по Релею, если путь распространения вдоль луча обзора блокирован, в противном случае имеем распределение Райса. Мобильный радиоприемник, который перемещается по большому пространству, должен иметь возможность обрабатывать сигналы, подвергнувшиеся замиранию обоих типов (мелкомасштабное, наложенное на крупномасштабное).

Крупномасштабное замирание (ослабление или потери в тракте) можно рассматривать как пространственное усреднение мелкомасштабных флуктуаций сигнала. Оно вычисляется, как правило, путем усреднения полученного сигнала по интервалу, превышающему 10–30 длин волн, чтобы отделить мелкомасштабные (главным образом релеевские) флуктуации от крупномасштабных эффектов затенения (обычно с логарифмически нормальным распределением). Существует три основных механизма, воздействующих на распространение сигнала в системах мобильной связи [1].

- *Отражение* (reflection) происходит тогда, когда распространяющаяся электромагнитная волна сталкивается с гладкой поверхностью, размер которой гораздо больше длины волны радиочастотного сигнала (λ).
- *Дифракция* (diffraction) встречается тогда, когда путь распространения между передатчиком и приемником преграждается плотным телом, размеры которого велики по сравнению с λ, что вызывает появление вторичных волн, образующихся позади преграждающего тела. Дифракция — это явление, которое является причиной того, что распространение радиочастотной энергии от передатчика к приемнику происходит в обход пути прямой видимости между ними. Ее часто называют *затенением* (shadowing), поскольку дифрагированное поле может достичь приемника, даже если оно затенено непроницаемой преградой.
- *Рассеяние* (scattering) встречается тогда, когда радиоволна сталкивается с любой неровной поверхностью или с поверхностью, размеры которой равны порядка λ или меньше, что приводит к распространению (рассеянию) или отражению энергии во всех направлениях. В городской местности обычные препятствия, вызывающие рассеивание сигнала, — это фонарные столбы, уличные знаки и листья. Название *рассеивающий элемент* (scatterer) применимо к любым препятствиям на пути распространения, которые являются причиной отражения или рассеяния сигнала.

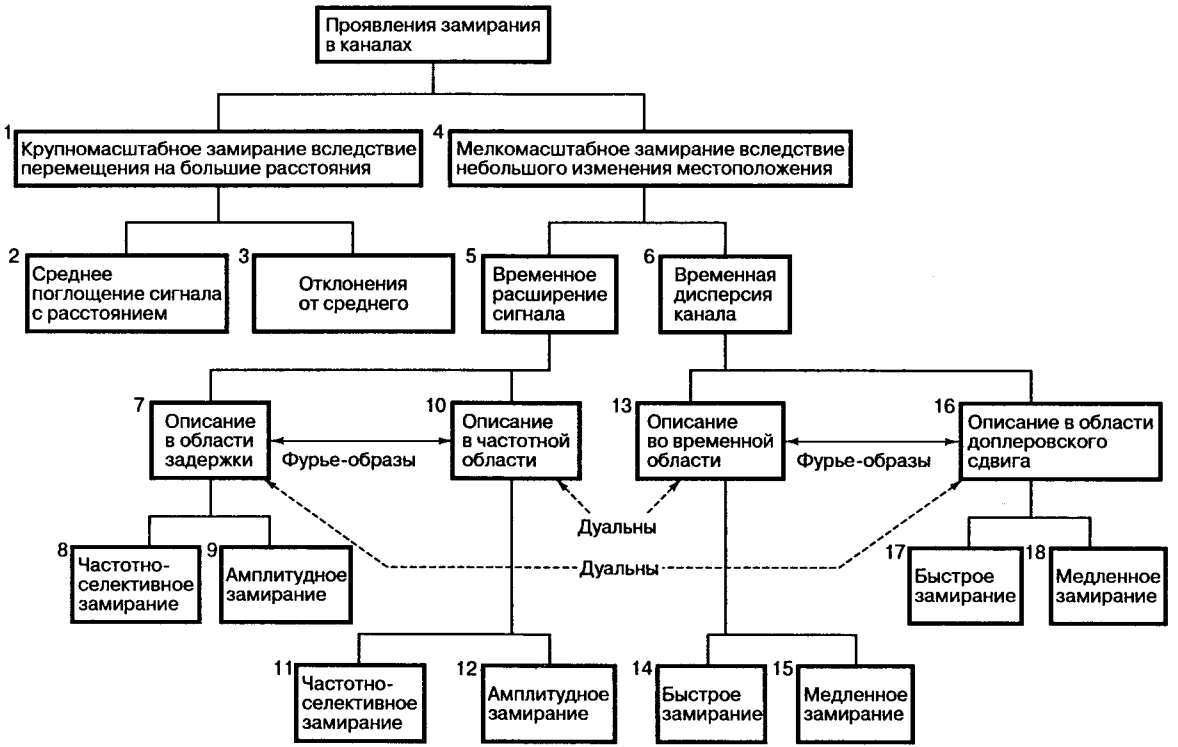


Рис. 15.1. Проявление замирания в канале

Рис. 15.1 можно использовать как оглавление следующих разделов. Два проявления мелкомасштабного замирания, временное расширение сигнала (дисперсия сигнала) и нестационарное поведение канала, будут исследованы в двух областях: временной и частотной, как указано в блоках 7, 10, 13 и 16 (рис. 15.1). При дисперсии сигнала типы ухудшений характеристик, возникающих вследствие замирания, разделены на частотно-селективные или частотно-неселективные (амплитудные), как показано в блоках 8, 9, 11 и 12. При переменном во времени поведении типы ухудшений характеристик, возникающих вследствие замирания, разделены на быстрые и медленные, как показано в блоках 14, 15, 17 и 18. Пометки "Фурье-образы" и "дуальны" будут объяснены позже.

Удобной (но не совсем точной) иллюстрацией является рис. 15.2, показывающий различные вклады, которые должны рассматриваться при оценке потерь в тракте при анализе бюджета линии связи для мобильной радиосвязи [2]: (1) средние потери в тракте в результате крупномасштабного замирания как функция расстояния, (2) резерв крупномасштабного замирания в расчете на (почти) наихудший вариант отклонения от средних потерь в тракте (обычно 6–10 дБ) и (3) резерв релеевского или мелкомасштабного замирания в расчете на (почти) наихудший вариант (обычно 20–30 дБ). На рис. 15.2 примечание "≈ 1–2%" указывает предложенную область (вероятность) под хвостом каждой функции распределения вероятности, используемую как задачу разработки. Таким образом, величина указанного резерва предназначена для обеспечения достаточной мощности полученного сигнала для приблизительно 98–99% возможных значений замирания (крупно- и мелкомасштабного).

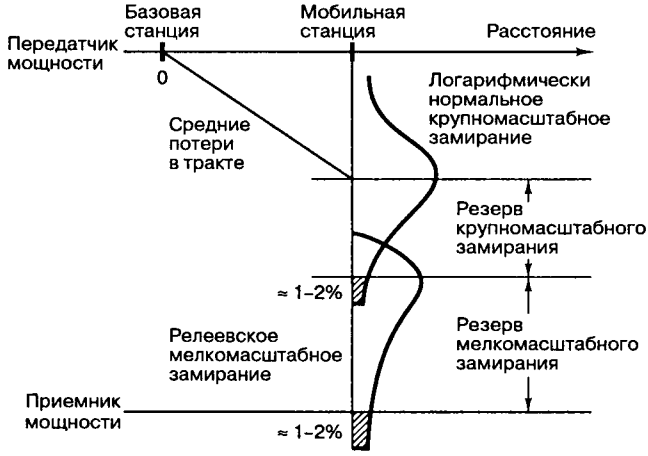


Рис. 15.2. Замирание в канале через бюджет линии связи. (Источник: Greenwood D. and Hanzo L. "Characterization of Mobile Radio Channels". Mobile Radio Communications, edited by R. Steele, Chapter 2, Pentech Press, London, 1994.)

С помощью комплексной формы записи переданный сигнал можно представить следующим образом.

$$s(t) = \mathrm{Re}\left\{g(t)e^{2\pi i f_c t}\right\} \tag{15.2}$$

В данном случае $\mathrm{Re}\{\cdot\}$ — действительная часть величины $\{\cdot\}$, а f_c — несущая частота. Узкополосный сигнал $g(t)$ называется комплексной огибающей $s(t)$ (см. раздел 6.4) и может быть выражен как

$$g(t) = |g(t)|e^{i\phi(t)} = T(t)e^{i\phi(t)}, \tag{15.3}$$

где $R(t) = |g(t)|$ — модуль огибающей, а $\phi(t)$ — ее фаза. Для чистого фазово- или частотно-модулированного сигнала $R(t)$ будет постоянным и в общем случае будет медленно изменяться по сравнению с $t = 1/f_c$.

В среде с замиранием $g(t)$ изменится на комплексный безразмерный множитель $\alpha(t)e^{-i\theta(t)}$ (его происхождение будет показано позже). *Модифицированный* узкополосный сигнал можно записать в виде $\alpha(t)e^{-i\theta(t)}g(t)$. Рассмотрим амплитуду $\alpha(t)R(t)$ этой огибающей, которую можно выразить через три положительных члена [3].

$$\alpha(t)R(t) = m(t) \times r_0(t) \times R(t) \tag{15.4}$$

Здесь $m(t)$ называют *компонентом крупномасштабного замирания* огибающей, а $r_0(t)$ — *компонентом мелкомасштабного замирания*. Иногда $m(t)$ именуют *локальным средним*, или *логарифмически нормальным замиранием*, поскольку его измеряемые значения можно статистически описать с помощью логарифмически нормальной функции распределения вероятностей; или, что равносильно, при измерении в децибелах $m(t)$ имеет гауссову функцию распределения вероятностей. Кроме того, $r_0(t)$ иногда называют *замиранием вследствие многолучевого распространения*, или *релеевским замиранием*. На рис. 15.3 показана связь между $\alpha(t)$ и $m(t)$ для мобильной радиосвязи. В этом рисунке учтено, что была передана *немодулированная* несущая волна, а это в контексте уравнения (15.4) означает, что в любое время $R(t) = 1$. Типичный график зависимости мощности полученного сигнала от смещения антенны (обычно в единицах длины волны) показан на рис. 15.3, *а*. Мощность полученного сигнала является, конечно, функцией множителя $\alpha(t)$. Можно без труда определить мелкомасштабные замирания, наложенные на крупномасштабные. Обычное изменение положения антенны, соответствующее переходу между соседними нулями изменения интенсивности сигнала вследствие мелкомасштабного замирания, равно приблизительно половине длины волны. На рис. 15.3, *б* крупномасштабное замирание или локальное среднее $m(t)$ было удалено, чтобы показать мелкомасштабное замирание $r_0(t)$, относящееся к некоторой постоянной средней мощности. Напомним, что $m(t)$ можно, как правило, оценить с помощью усреднения принятой огибающей по 10–30 длинам волн. Логарифмически нормально распределенное замирание является относительно медленно изменяющейся функцией местоположения. Следует отметить, что в приложениях, включающих движение, таких как использование радио в движущейся машине, зависимость от местоположения равносильна зависимости от времени. Ниже приведены некоторые подробности, касающиеся статистики и механизмов крупномасштабного и мелкомасштабного замираний.

15.2.1. Крупномасштабное замирание

Для систем мобильной радиосвязи Окумура (Okumura) [4] выполнил некоторые первоначальные измерения потерь в тракте для большого числа высот антенн и расстояний покрытия. Хата (Hata) [5] придал данным Окумуры вид параметрических формул. Вообще, модели распространения как для комнатных, так и для наружных каналов показывают, что средние потери в тракте $\overline{L_p}(d)$, как функция расстояния между передатчи-

ком и приемником d, пропорциональны n-й степени d, выраженного в единицах эталонного расстояния d_0. Математически это можно выразить следующим образом.

$$\overline{L_p}(d) \propto \left(\frac{d}{d_0}\right)^n \tag{15.5}$$

$\overline{L_p}(d)$ часто определяется в децибелах.

$$\overline{L_p}(d)\ (\text{дБ}) = L_s(d_0)\ (\text{дБ}) + 10n \lg\left(\frac{d}{d_0}\right) \tag{15.6}$$

Эталонное расстояние d_0 соответствует точке, размещенной в дальнем поле передающей антенны. Обычно значение d_0 берется равным 1 км для крупных ячеек, 100 м — для микроячеек и 1 м — для комнатных каналов. Кроме того, оценивается (с помощью уравнения (15.1)) или измеряется $L_s(d_0)$. $\overline{L_p}(d)$ — это средние (по всему множеству различных местоположений) потери в тракте для данного значения d.

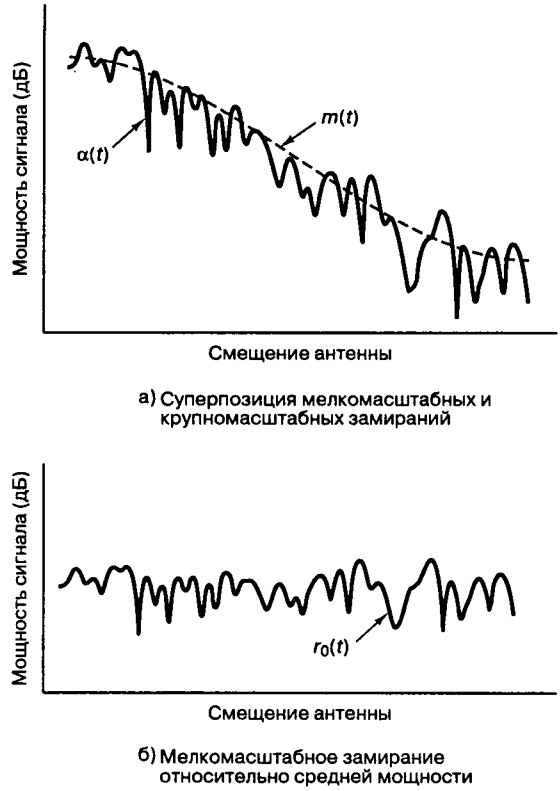


Рис. 15.3. Крупномасштабное и мелкомасштабное замирания

Если нарисовать график зависимости $\overline{L_p}(d)$ от d в логарифмическом масштабе обеих осей (для расстояний, больших d_0), то получится прямая линия с наклоном, равным $10n$. Пока-

затель степени потерь в тракте n зависит от частоты, высоты антенны и среды распространения. В свободном пространстве, где распространение сигнала происходит согласно закону обратных квадратов (как описывается в разделе 5.3.1), n равно 2, что видно из уравнения (15.1). Если имеется эффект волновода (например, при распространении по улицам города), n может быть меньше 2. При наличии препятствий n больше. На рис. 15.4 показана зависимость потерь в тракте от расстояния, полученная при измерениях, проведенных в нескольких местах Германии [6]. Здесь потери в тракте измерялись относительно эталонного расстояния $d_0 = 100$ м. Показана также линейная аппроксимация для разных значений показателя степени.

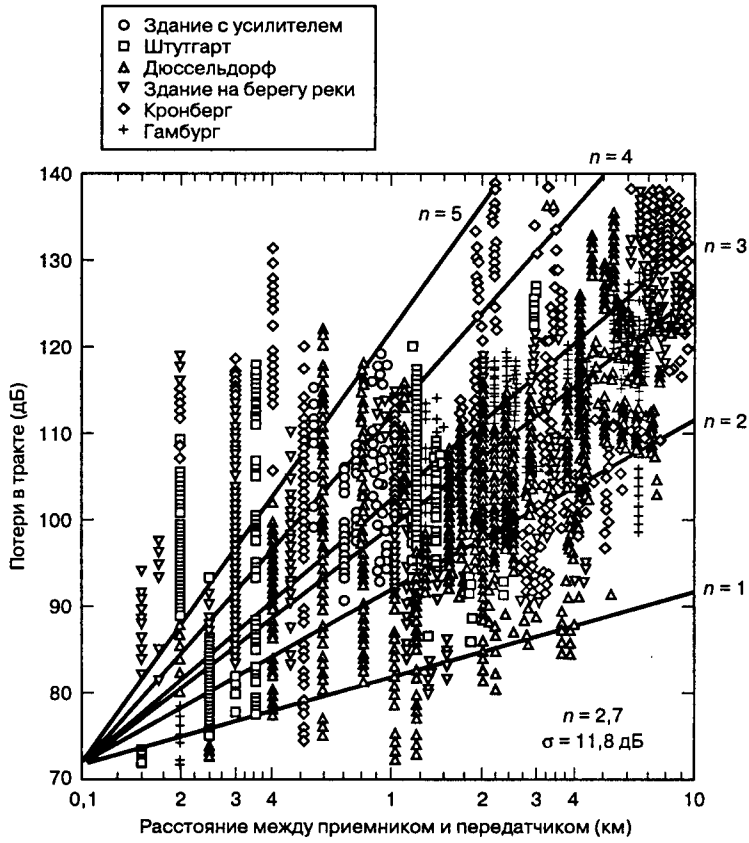


Рис. 15.4. Потери в тракте в зависимости от расстояния, измеренные в нескольких городах Германии. (Источник: Seidel S. Y. et. al. "Path Loss, Scattering and Multipath Delay Statistics in Four European Cities of Digital Cellular and Microcellular Radiotelephone". IEEE Transactions on Vehicular Technology, vol. 40, n. 4, pp. 721–730, November, 1991.)

Выражение (15.6) показывает средние потери в тракте и, следовательно, непригодно для описания конкретной конфигурации или пути распространения сигнала. Необходимо ввести отклонения от среднего значения, поскольку в различных местах среда может существенно влиять на работу системы, даже при одинаковом расположении передатчика и приемника. На рис. 15.4 показано, что разброс величины потерь

в тракте может быть весьма большим. Измерения показали, что для любых значений d потери в тракте L_p являются случайной переменной, имеющей логарифмически нормальное распределение в окрестности среднего значения $\overline{L_p}(d)$ [7]. Таким образом, потери в тракте L_p можно выразить через $\overline{L_p}(d)$, введя в уравнение (15.6) случайную переменную X_σ.

$$L_p(d)\ (\text{дБ}) = L_s(d_0)\ (\text{дБ}) + 10\,n\,\lg(d/d_0) + X_\sigma\ (\text{дБ}) \tag{15.7}$$

Здесь X_σ обозначает случайную гауссову переменную с нулевым средним (в децибелах) со среднеквадратическим отклонением σ (также в децибелах). X_σ зависит от местоположения и расстояния. Поскольку X_σ и $L_p(d)$ — это случайные переменные, то, если для вычисления потерь в тракте или энергетического резерва линии связи использовать уравнение (15.7), предварительно нужно выбрать какое-то определенное значение X_σ. Часто выбор этого значения основывается на измерениях (сделанных для большого числа взаимных размещений приемника и передатчика). Обычные значения X_σ — это 6–10 дБ или даже выше. Таким образом, для статистического описания потерь в тракте вследствие крупномасштабного замирания при произвольном расположении с определенным расстоянием между передатчиком и приемником будут необходимы такие параметры: 1) эталонное расстояние, 2) показатель степени потерь в тракте и 3) среднеквадратическое отклонение X_σ. (Имеется несколько хороших работ, касающихся измерения и оценки потерь в тракте при распространении для различных приложений и конфигураций [1, 5–9].)

15.2.2. Мелкомасштабное замирание

В этом разделе будет рассмотрен компонент мелкомасштабного замирания r_0. Анализ проводится в предположении, что антенна движется по ограниченной траектории так, что влияние крупномасштабного замирания $m(t)$ постоянно (и предполагается равным единице). Предположим, антенна перемещается и существует множество путей рассеивающих элементов, с каждым из которых связана переменная задержка распространения $\tau_n(t)$ и переменный множитель $\alpha_n(t)$. Пренебрегая шумом, можно записать полученный полосовой сигнал следующим образом.

$$r(t) = \sum_n \alpha_n(t)s[t - \tau_n(t)] \tag{15.8}$$

Подставляя уравнение (15.2) в (15.8), запишем полученный полосовой сигнал следующим образом.

$$\begin{aligned} r(t) &= \text{Re}\left(\left\{\sum_n \alpha_n(t)g[t-\tau_n(t)]\right\}e^{2\pi i f_c[t-\tau_n(t)]}\right) = \\ &= \text{Re}\left(\left\{\sum_n \alpha_n(t)e^{-2\pi i f_c\tau_n(t)}g[t-\tau_n(t)]\right\}e^{2\pi i f_c t}\right) \end{aligned} \tag{15.9}$$

Из уравнения (15.9) следует, что соответствующий полученный узкополосный сигнал будет иметь следующий вид.

$$z(t) = \sum_n \alpha_n(t) e^{-2\pi i f_c \tau_n(t)} g[t - \tau_n(t)] \tag{15.10}$$

Рассмотрим передачу *немодулированной* несущей на частоте f_c. Иными словами, в любой момент времени $g(t) = 1$. Тогда для немодулированной несущей частоты и дискретных компонентов многолучевого распространения, выраженных в форме (15.10), принятый узкополосный сигнал упростится до следующего вида

$$z(t) = \sum_n \alpha_n(t) e^{-2\pi i f_c \tau_n(t)} = \sum_n \alpha_n(t) e^{-i\theta_n(t)}, \tag{15.11}$$

где $\theta_n(t) = 2\pi f_c \tau_n(t)$. Узкополосный сигнал $z(t)$ состоит из суммы переменных во времени векторов, имеющих амплитуду $\alpha_n(t)$ и фазу $\theta_n(t)$. Следует отметить, что $\theta_n(t)$ будет изменяться на 2π радиан, когда τ_n изменится на $1/f_c$ (обычно, это очень маленькая задержка). При работе сотового радиопередатчика на частоте $f_c = 900$ МГц задержка $1/f_c$ равна 1,1 наносекунд. В свободном пространстве это соответствует изменению пути распространения сигнала на 33 см. Таким образом, в уравнении (15.11) $\theta_n(t)$ может существенно измениться при относительно небольших изменениях задержки распространения. В этом случае, если два компонента многолучевого распространения сигнала отличаются по длине пути на 16,5 см, то один прибывающий сигнал будет отличаться по фазе от другого на 180 градусов. Иногда векторы сигналов суммируются конструктивно, а иногда — деструктивно, что приводит в результате к изменениям амплитуды или замиранию $z(t)$. Уравнение (15.11) можно записать более компактно в виде суммарной полученной огибающей, просуммированной по всем рассеивающим элементам.

$$z(t) = \alpha(t) e^{-i\theta(t)} \tag{15.12}$$

Здесь $\alpha(t)$ — результирующая амплитуда, а $\theta(t)$ — результирующая фаза. В правой части уравнения (15.12) представлен тот комплексный множитель, который ранее описывался в разделе 15.2. Уравнение (15.12) является важным результатом, поскольку из него видно, что хотя *полосовой* сигнал $s(t)$, как показано в уравнении (15.2), подвержен замиранию, что приводит к приему сигнала $r(t)$, это замирание можно описать, анализируя $r(t)$ на уровне *узкополосного* сигнала.

На рис. 15.5 показан основной механизм, приводящий к замиранию в каналах с многолучевым распространением, как описывается уравнениями (15.11) и (15.12). На рисунке отраженный сигнал запаздывает по фазе (из-за увеличения расстояния распространения) относительно ожидаемого сигнала. Отраженный сигнал также имеет меньшую амплитуду (функция коэффициента отражения препятствия). Отраженные сигналы можно описать с помощью ортогональных компонентов $x_n(t)$ и $y_n(t)$, где $x_n(t) + iy_n(t) = \alpha_n(t) e^{-i\theta_n(t)}$. Если количество таких стохастических компонентов велико и ни один из них не преобладает, то *в фиксированный момент времени* переменные $x_r(t)$ и $y_r(t)$, являющиеся результатом суммирования всех $x_n(t)$ и $y_n(t)$, соответственно, будут иметь гауссову функцию распределения вероятностей. Эти ортогональные компоненты дают то же мелкомасштабное замирание r_0, которое было определено в уравнении (15.4). При немодулированной несущей волне, как показано в уравнении (15.12), $r_0(t)$ является модулем $z(t)$

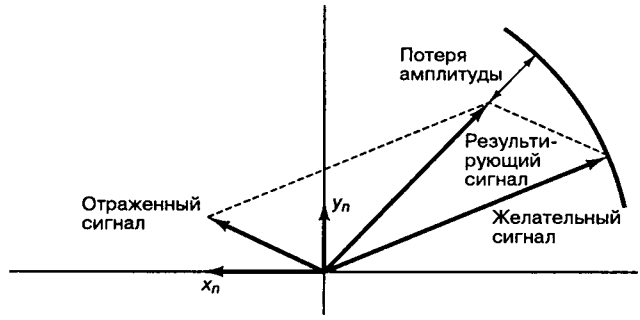


Рис. 15.5. Влияние многолучевого отражения сигнала на ожидаемый сигнал

$$r_0(t) = \sqrt{x_r^2(t) + y_r^2(t)} \tag{15.13}$$

Если полученный сигнал составлен из множественных отраженных лучей и значительного (незамирающего) компонента, распространяемого в пределах прямой видимости, амплитуда полученной огибающей имеет райсовскую функцию распределения плотности вероятности, показанную ниже, а замирание называют *райсовским* [2].

$$p(r_0) = \begin{cases} \dfrac{r_0}{\sigma^2} \exp\left[-\dfrac{(r_0^2 + A^2)}{2\sigma^2} \right] I_0\left(\dfrac{r_0 A}{\sigma^2} \right) & \text{для } r_0 \geq 0,\ A \geq 0 \\ 0 & \text{для других } r_0,\ A \end{cases} \tag{15.14}$$

Хотя $r_0(t)$ динамически изменяется во время движения, в любой *фиксированный момент времени* — это случайная переменная, которая является положительным действительным числом. Поэтому, описывая функцию плотности вероятности, можно опустить ее зависимость от времени. Параметр σ^2 — это средняя мощность многолучевого сигнала до обнаружения, A — максимальное значение *незамирающего* компонента сигнала (называемом *зеркальным компонентом*), а $I_0(\cdot)$ — модифицированная функция Бесселя первого рода нулевого порядка [11]. Распределение Райса часто записывают через параметр K, который определяется как отношение мощности зеркального компонента к мощности многолучевого сигнала. Математически это записывается как $K = A^2/(2\sigma^2)$. При приближении к нулю амплитуды зеркального компонента функция плотности вероятностей Райса стремится к функции плотности вероятности Релея, имеющей следующий вид.

$$p(r_0) = \begin{cases} \dfrac{r_0}{\sigma^2} \exp\left[-\dfrac{r_0^2}{2\sigma^2} \right] & \text{для } r_0 \geq 0 \\ 0 & \text{для других } r_0 \end{cases} \tag{15.15}$$

Релеевский замирающий компонент иногда называется *случайным*, *рассеянным* или *диффузным*, а плотность вероятности Релея является результатом отсутствия зеркального компонента сигнала; таким образом, для одиночной линии связи (без разнесения) она представляет собой функцию распределения вероятностей, связанную с наибольшим замиранием, приходящимся на среднюю мощность полученного сигнала. В

остальной части этой главы будет предполагаться (если не оговорено иное), что снижение отношения сигнал/шум (signal-to-noise ratio — SNR) вследствие замирания описывается моделью Релея. Будем также считать, что распространение сигнала происходит в полосе УВЧ, включающей сотовую и персональную службы связи, которым выделены частоты 1 ГГц и 2 ГГц.

Как показано на рис. 15.1 (блоки 4–6), мелкомасштабное замирание проявляется двумя способами: (1) путем расширения цифровых импульсов сигнала и (2) посредством переменного во времени поведения канала, вызванного движением (например, принимающая антенна находится на движущейся платформе). На рис. 15.6 последствия этого показаны как реакция многолучевого канала на короткий импульс в зависимости от задержки при различных местоположениях антенны (или различном времени, предполагая, что перемещение происходит с постоянной скоростью). На рис. 15.6 важно различать задержку τ и время передачи или наблюдения t. Задержка — это следствие расширения во времени, являющегося результатом неоптимальной импульсной характеристики канала с замираниями. Время передачи связано с передвижением антенны или пространственными изменениями, учитывающими изменения пути распространения, которые определяют нестационарное поведение канала. Нужно заметить, что при постоянной скорости, как предполагается на рис. 15.6, для иллюстрации переменного во времени поведения можно использовать либо местоположение антенны, либо время передачи. На рис. 15.6, *а–в* показана последовательность полученных профилей мощности импульса при проходе антенной равных расстояний. Ситуации, изображенные на рисунках, отличаются изменением положения антенны на $0{,}4\lambda$ [12], где λ — длина волны несущей частоты. Для каждого из показанных случаев модели отклика канала существенно отличаются по времени замирания наибольшего компонента сигнала, по количеству копий сигнала, их амплитуде и общей полученной мощности (площадь под каждым полученным профилем мощности). На рис. 15.7 обобщаются названные механизмы мелкомасштабного замирания, и в двух областях (время или задержка и частота или доплеровское смещение) рассматриваются механизмы и категории ухудшения качества передачи, связанные с каждым механизмом. Отметим, что всякий механизм, описанный во временной области, также хорошо можно описать и в частотной области. Таким образом, как представлено на рис. 15.7, механизм расширения по времени во временной области будет характеризоваться задержкой многолучевого распространения, а в частотной области — полосой когерентности канала. Подобным образом нестационарный механизм во временной области будет характеризоваться временем когерентности канала, а в частотной области — скоростью замирания в канале или доплеровским расширением. Эти механизмы и связанные с ними категории ухудшения характеристик рассматриваются в следующих разделах.

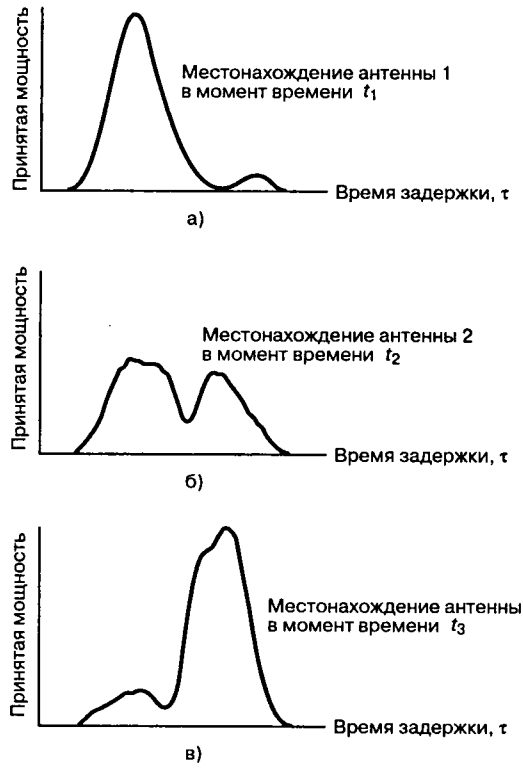


Рис. 15.6. Реакция многолучевого канала на короткий импульс в зависимости от задержки для различных местоположений антенны

Рис. 15.7. Мелкомасштабное замирание: механизмы, категории и следствия

15.3. Расширение сигнала во времени

15.3.1. Расширение сигнала во времени, рассматриваемое в области задержки

Простой способ моделирования явлений замирания был предложен Белло (Bello) [13] в 1963 году; он ввел понятие стационарного в широком смысле некоррелированного рассеяния (wide-sense stationary uncorrelated scattering — WSSUS). В такой модели сигналы, поступающие на антенну приемника с различными задержками, рассматриваются как некоррелирующие. Можно показать [2, 13], что такие каналы являются эффективно стационарными в широком смысле, как во временной, так и в частотной области. Применив такую модель к каналу с замиранием, Белло смог определить функции, которые применимы для любого момента времени и любой частоты. На рис. 15.8 для мобильного канала указаны четыре такие функции, составляющие названную модель [2, 10, 13–15]. Рассмотрим функции, начиная с рис. 15.8, *а* и двигаясь против часовой стрелки в направлении рис. 15.8, *г*.

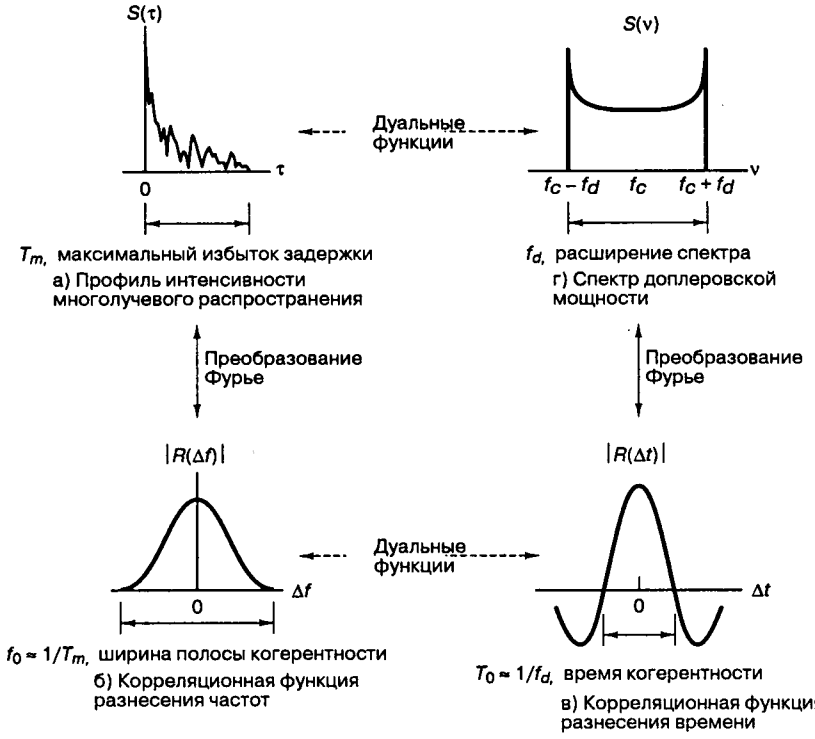


Рис. 15.8. Соотношения между корреляционными функциями канала и функциями плотности мощности

На рис. 15.8, *а* отображен *профиль интенсивности многолучевого распространения* (зависимость $S(\tau)$ от задержки τ). Зная $S(\tau)$, можно определить, как для переданного импульса полученная мощность зависит от временной задержки τ. Термин "временная задержка" (time delay) используется для обозначения избыточной задержки распростране-

ния сигнала. Он представляет задержку данного сигнала относительно времени поступления на приемник первого сигнала. Для типичного беспроводного канала полученный сигнал обычно состоит из нескольких дискретных многолучевых компонентов, приводящих к появлению изолированных пиков $S(\tau)$, называемых иногда *пальцами*, или *отраженными сигналами*. Для некоторых каналов, таких как тропосферный канал с рассеянием, принятые сигналы выглядят как континуум многолучевых компонентов [10, 15]. В таких случаях $S(\tau)$ — это относительно гладкая (непрерывная) функция τ. Для измерения профиля интенсивности многолучевого распространения необходимо воспользоваться широкополосными сигналами (импульсы или сигналы с расширенным спектром) [15]. Для единичного переданного импульса время T_m между приемом первого и последнего компонентов представляет собой *максимальную избыточную задержку распространения*, после которой мощность многолучевого сигнала падает ниже определенного порогового уровня относительно самого мощного компонента. Пороговый уровень можно выбрать на 10 или 20 дБ ниже уровня самого мощного луча. Отметим, что в идеальной системе (нулевая избыточная задержка) функция $S(\tau)$ состояла бы из идеального импульса с весовым коэффициентом, равным общей средней мощности полученного сигнала.

15.3.1.1. Категории ухудшения качества передачи вследствие расширения сигнала во времени, рассматриваемого в области задержки

В канале с замираниями взаимосвязь между максимальной избыточной задержкой распространения T_m и временем передачи символа T_s можно рассматривать с позиции двух различных категорий ухудшения качества передачи: *частотно-селективного замирания* (frequency-selective fading) и *частотно-неселективного* (frequency nonselective fading), или амплитудного замирания (flat fading) (см. рис. 15.1, блоки 8 и 9, и рис. 15.7). Говорят, что канал обнаруживает частотно-селективное замирание, если $T_m > T_s$. Это условие реализуется, когда полученный многолучевой компонент символа выходит за пределы длительности передачи символа. Такая многолучевая дисперсия порождает тот же тип искажений ISI, что и электронный фильтр. Фактически другим названием этой категории ухудшения передачи вследствие замирания является *вводимая каналом ISI*. При частотно-селективном замирании возможно уменьшение искажений, поскольку многие многолучевые компоненты разрешаются приемником. (Несколько подобных методов борьбы с замиранием описаны в следующих разделах.)

Говорят, что канал является *частотно-неселективным* или проявляется *амплитудное замирание*, если $T_m < T_s$. В этом случае все полученные многолучевые компоненты символа поступают в течение времени передачи символа; поэтому компоненты не разрешаются. В данном случае отсутствуют искажения за счет вводимой каналом ISI, так как расширение сигнала во времени не приводит к существенному наложению соседних полученных символов. Однако ухудшение характеристик все же имеет место, поскольку неразрешенные компоненты вектора сигнала могут деструктивно суммироваться, что приводит к значительному уменьшению SNR. К тому же сигнал, классифицированный как проявляющий амплитудное замирание, может иногда испытывать частотно-селективное замирание. Это будет объяснено позже, при рассмотрении ухудшения характеристик в частотной области, в которой такие явления описываются проще. При уменьшении SNR за счет амплитудного замирания можно использовать специальные методы подавления замирания, улучшающие принимаемое значение SNR (или уменьшающие требуемое SNR). Для цифровых систем наиболее эффективным способом является введение каких-либо форм разнесения сигналов и использование кодов коррекции ошибок.

15.3.2. Расширение сигнала во времени, рассматриваемое в частотной области

Полностью аналогичное описание дисперсии сигнала можно привести и в частотной области. На рис. 15.8, *б* можно видеть функцию $|R(\Delta f)|$, обозначенную как корреляционная функция *разнесения частоты*; это Фурье-образ $S(\tau)$. Функция $R(\Delta f)$ представляет корреляцию между реакциями канала на два сигнала как функцию разности частот этих сигналов. Ее можно рассматривать так, как частотную передаточную функцию канала. Следовательно, расширение сигнала во времени можно рассматривать как следствие процесса фильтрации. Зная $R(\Delta f)$, можно определить, какова корреляция между полученными сигналами, разнесенными по частоте на $\Delta f = f_1 - f_2$. Функцию $R(\Delta f)$ можно измерить, передавая пару синусоид, разнесенных по частоте на Δf, изучая взаимную корреляцию спектров двух полученных сигналов и повторяя этот процесс многократно посредством увеличения Δf. Таким образом, измерение $R(\Delta f)$ можно проводить с помощью синусоид, смещающихся по частоте вдоль интересующей полосы (широкополосный сигнал). *Полоса когерентности* (coherence bandwidth) f_0 является статистической мерой диапазона частот, по которому канал пропускает все спектральные компоненты с приблизительно равным коэффициентом усиления и линейным изменением фазы. Таким образом, полоса когерентности представляет диапазон частот, в пределах которого частотные компоненты сигнала имеют большую вероятность амплитудной корреляции. Иными словами, на все спектральные компоненты этого диапазона канал влияет одинаково, например, проявляя или не проявляя замирание. Следует отметить, что f_0 и T_m взаимосвязаны (с точностью до постоянного множителя). Можно сказать, что приблизительно

$$f_0 \approx 1/T_m \tag{15.16}$$

Максимальная избыточная задержка T_m не обязательно является наилучшим показателем того, как будет функционировать произвольная система при распространении сигнала в канале, поскольку различные каналы с одинаковым значением T_m могут иметь весьма различный профиль интенсивности сигнала в период задержки. Более подходящим параметром является разброс задержек, который чаще всего описывается через среднеквадратическое значение и называется среднеквадратическим разбросом задержек.

$$\sigma_\tau = \sqrt{\overline{\tau^2} - (\overline{\tau})^2} \tag{15.17}$$

Здесь $\overline{\tau}$ — это средняя избыточная задержка, $(\overline{\tau})^2$ — квадрат среднего, $\overline{\tau^2}$ — второй момент, а σ_τ — квадратный корень второго центрального момента $S(\tau)$ [1].

Не существует универсального соотношения между полосой когерентности и разбросом задержек. Однако, используя метод Фурье-преобразований и измерения дисперсии реальных сигналов в различных каналах, можно получить полезную аппроксимацию. В настоящее время разработано несколько приблизительных соотношений. Если полоса когерентности определена как интервал частот, в пределах которого комплексная частотная передаточная функция канала имеет корреляцию не менее 0,9, то полосу когерентности можно приблизительно записать в следующем виде [16].

$$f_0 \approx \frac{1}{50\sigma_\tau} \tag{15.18}$$

Для мобильной радиосвязи в качестве подходящей модели описания распространения в городской среде обычно берут совокупность рассеивающих элементов, имеющих радиальное равномерное распределение, равные коэффициенты отражения, но независимые случайные фазовые углы отражения [17, 18]. Эту модель называют моделью канала с *плотным размещением рассеивающих элементов*. При ее использовании полоса когерентности частот определяется подобным образом [17]: интервал частот, в пределах которого комплексная частотная передаточная функция канала имеет корреляцию не менее 0,5.

$$f_0 = \frac{0{,}276}{\sigma_\tau} \tag{15.19}$$

При изучении ионосферных эффектов часто используют следующее определение [19].

$$f_0 = \frac{1}{2\pi\sigma_\tau} \tag{15.20}$$

Более распространенным приближением для f_0, соответствующим определению, где корреляция должна быть не меньше 0,5, является следующее [1].

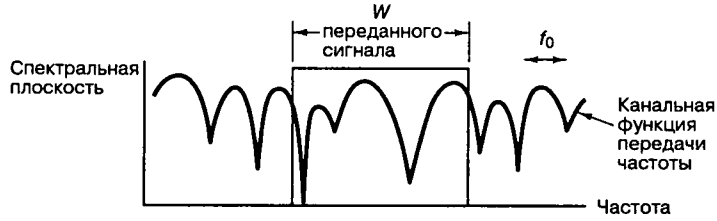
$$f_0 \approx \frac{1}{5\sigma_\tau} \tag{15.21}$$

Разброс задержек и полоса когерентности связаны с характеристиками многолучевого распространения в канале и отличаются для разных путей распространения (городская черта, пригород, холмистая местность, помещения и т.д.). Важно отметить, что параметры в уравнении (15.21) не зависят от скорости передачи сигналов. Скорость передачи влияет только на ширину полосы пропускания, W.

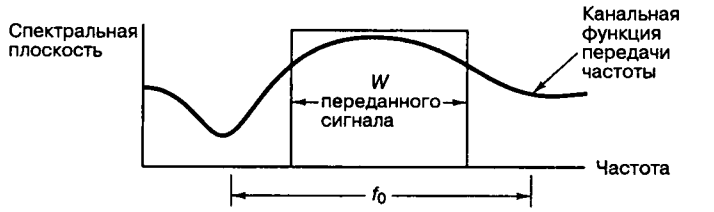
15.3.2.1. Категории ухудшения качества передачи вследствие расширения сигнала во времени, рассматриваемого в частотной области

Канал называется частотно-селективным (frequency-selective), если $f_0 < 1/T_s \approx W$, где скорость передачи символов $1/T_s$ номинально берется равной скорости передачи сигналов или ширине полосы частот сигнала W. На практике W может отличаться от $1/T_s$ из-за системной фильтрации или выбора типа модуляции данных (например, QPSK, MSK, расширение спектра и т.д.) [20]. Частотно-селективное замирание проявляется тогда, когда канал неодинаково влияет на разные спектральные компоненты сигнала. Некоторые спектральные компоненты сигнала, не входящие в полосу когерентности, будут подвергаться различному (и независимому) воздействию, в отличие от тех компонентов, которые приходятся на полосу когерентности. На рис. 15.9 приведено три примера. В каждом из них показана зависимость спектральной плотности от частоты переданного сигнала, имеющего полосу W Гц. На графике (рис. 15.9, *а*) на сигнал наложена частотная передаточная функция частотно-селективного канала ($f_0 < W$). На рис. 15.9, *а* показано, что различные спектральные компоненты переданного сигнала будут подвергаться различному воздействию.

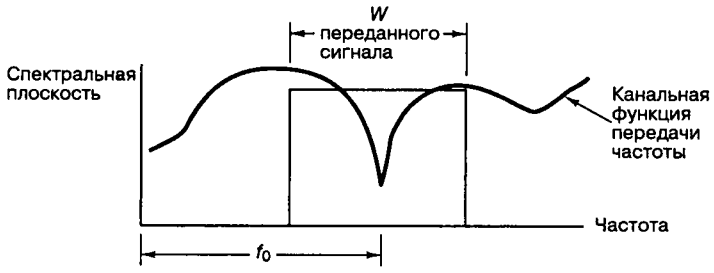
Частотно-неселективное, или амплитудное, ухудшение характеристик происходит тогда, когда $f_0 > W$. Следовательно, все спектральные компоненты сигнала будут подвергаться одинаковому воздействию со стороны канала (например, замирать или не замирать). Это показано на рис. 15.9, *б*, где изображена спектральная плотность того же переданного сигнала, имеющего полосу W Гц.



а) Типичный случай частотно-селективного замирания ($f_0 < W$)



б) Типичный случай амплитудного замирания ($f_0 > W$)



в) Нуль канальной функции передачи частоты попадает на центр полосы сигнала ($f_0 > W$)

Рис. 15.9. Связь между частотной передаточной функцией канала и переданным сигналом с полосой W

Однако на этот сигнал теперь наложена частотная передаточная функция канала с амплитудным замиранием ($f_0 > W$). Из рис. 15.9, *б* видно, что воздействие на все спектральные компоненты будет приблизительно равным. Амплитудное замирание не привносит искажений, связанных с внесенной каналом ISI, однако все же стоит ожидать ухудшения характеристик сигнала, выражающегося в уменьшении SNR. Чтобы избежать искажения вследствие внесенной каналом ISI, необходимо, чтобы канал проявлял амплитудное замирание. Это происходит при следующем условии.

$$f_0 > W \approx \frac{1}{T_s} \tag{15.22}$$

Следовательно, полоса когерентности f_0 устанавливает верхний предел скорости передачи, которую можно использовать, не включая в приемник эквалайзер.

На рис. 15.9, *б* показано обычное графическое представление амплитудного замирания, когда $f_0 > W$ (или $T_m < T_s$). Однако если мобильный радиоприемник будет менять свое местонахождение, некоторое время получаемый сигнал будет подвергаться частотно-селективному искажению, несмотря на то что $f_0 > W$. Соответствующая иллюстрация

приведена на рис. 15.9, *в*, где нуль частотной передаточной функции канала находится около середины полосы спектральной плотности переданного сигнала. Когда это происходит, узкополосный импульс может искажаться собственными смещенными низкочастотными компонентами. Одним из последствий этого является отсутствие надежного максимума импульса, составляющего основу синхронизации или предназначенного для выборки фазы несущей, переносимой импульсом [17]. Таким образом, хотя канал (на основе среднеквадратических соотношений) отнесен к каналам с амплитудным замиранием, он может периодически проявлять и частотно-селективное замирание. Стоит отметить, что канал мобильной радиосвязи, классифицированный как канал с амплитудным замиранием, не может все время проявлять амплитудное замирание. Когда f_0 становится намного больше W (или T_m становится намного меньше T_s), все меньший интервал времени реализуется состояние, показанное на рис. 15.9, *в*. Очевидно, что замирание на рис. 15.9, *а* не зависит от места в полосе частот сигнала, так что частотно-селективное замирание происходит не эпизодически, а все время.

15.3.3. Примеры амплитудного и частотно-селективного замирания

На рис. 15.10 показано несколько примеров амплитудного и частотно-селективного замирания для систем со спектром, расширенным методом прямой последовательности (direct-sequence spread-spectrum — DS/SS) [19, 20]. На этом рисунке изображены три графика зависимости выхода коррелятора псевдослучайного (pseudonoise — PN) кода от задержки как функции времени (времени передачи или наблюдения). Каждый график зависимости амплитуды от задержки подобен зависимости $S(\tau)$ от τ, показанной на рис. 15.8, *а*. Ключевое различие состоит в том, что амплитуды, показанные на рис. 15.10, представляют выход коррелятора; следовательно, форма сигнала является функцией импульсной характеристики не только канала, но и коррелятора. Задержка выражена в единицах длительности элементарных сигналов, где элементарный сигнал (chip) определяется как минимальный (по длительности) операционный блок системы расширенного спектра. На каждом графике время наблюдения отложено на оси, перпендикулярной плоскости зависимости амплитуды от задержки. Рис. 15.10 составлен по данным канала связи спутник-земля, проявляющего сцинтилляцию вследствие атмосферных помех. В то же время рис. 15.10 является полезной иллюстрацией трех различных состояний канала, которые могут быть применены для мобильной радиосвязи. Как показано на рисунке, на мобильный радиоприемник, движущийся вдоль оси времени наблюдения, влияют изменения профиля многолучевого распространения вдоль маршрута распространения. Ось времени наблюдения проградуирована в единицах элементарных сигналов. На рис. 15.10, *а* дисперсия сигнала (один пик отраженного сигнала) составляет порядка длительности элементарного сигнала T_{ch}. В типичной системе DS/SS, ширина полосы сигнала расширенного спектра приблизительно равна $1/T_{ch}$; таким образом, нормированная полоса когерентности f_0T_{ch} на рис. 15.10, *а* приблизительно равна единице, из чего следует, что ширина полосы когерентности равна порядка ширины полосы расширенного спектра. Это характерно для канала, который можно назвать частотно-неселективным, или слабо частотно-селективным. На рис. 15.10, *б*, где $f_0T_{ch} = 0{,}25$, дисперсия сигнала выражена более резко. Существует явно выраженная интерференция между элементарными сигналами, возникающая вследствие того, что ширина полосы когерентности составляет приблизительно 25 процентов от ширины полосы расширенного спектра. На рис. 15.10, *в*, где $f_0T_{ch} = 0{,}1$, дисперсия сигнала выражена еще более явно; интерференция между элементарными сигналами возросла вследствие того, что ширина полосы когерентности составляет приблизительно 10 процентов от полосы расши-

ренного спектра. Полосы когерентности (относительно скорости передачи сигнала расширенного спектра), показанные на рис. 15.10, *б*, *в*, описывают каналы, которые можно назвать, соответственно, умеренно и сильно селективными по частотам. Позже будет показано, что системы DS/SS, работающие с частотно-селективными каналами на уровне элементарных сигналов, не обязательно испытывают частотно-селективные искажения на уровне символов.

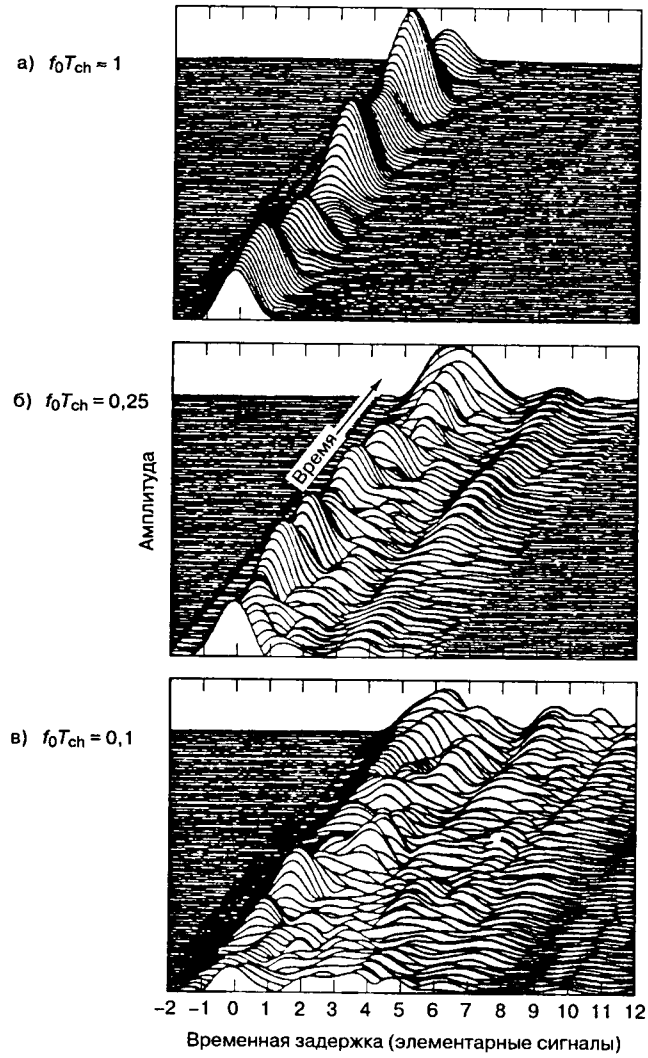
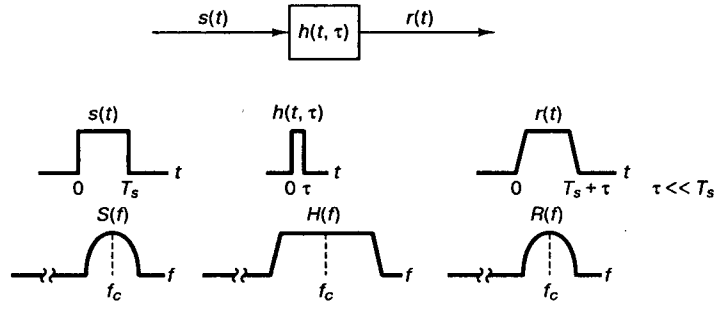
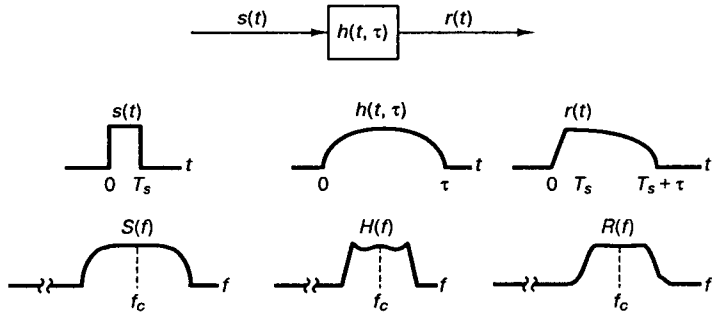


Рис. 15.10. Примеры временной развертки выхода согласованного фильтра DS/SS для трех случаев, где T_{ch} — длительность элементарного сигнала. (Источник: Bogusch R. L. "Digital Communications in Fading Channels: Modulation and Coding". Mission Research Corp., Santa Barbara, California, Report no. MRC-R-1034, March, 11, 1987.)

Проявление дисперсии сигнала в каналах с замираниями является аналогом расширения сигнала, характерного для электронного фильтра. На рис. 15.11, *а* изображен широкополосный фильтр (короткая импульсная характеристика) и его влияние на сигнал во временной и частотной областях. Этот фильтр похож на канал с амплитудным замиранием, выход которого относительно неискажен. На рис. 15.11, *б* показан узкополосный фильтр (широкая импульсная характеристика). Выходной сигнал претерпевает большее искажение как во временной, так и в частотной области. Данный процесс подобен происходящему в частотно-селективном канале.



а) Характеристики канала с амплитудным замиранием



б) Характеристики канала с частотно-селективным замиранием

Рис. 15.11. Характеристики частотно-селективного и амплитудного замирания. (Источник: Rappaport T. S. "Wireless Communications". Prentice-Hall, Upper Saddle River, New Jersey, 1996.)

15.4. Нестационарное поведение канала вследствие движения

15.4.1. Нестационарное поведение канала, рассматриваемое во временной области

Дисперсия сигнала и ширина полосы когерентности описывают в локальной области свойства канала, связанные с расширением во времени. В то же время они не дают информации о переменном во времени поведении канала, являющемся следствием относительного движения передатчика и приемника или передвижения объектов

внутри канала. Применяемые в мобильной радиосвязи каналы нестационарны, поскольку движение передатчика и приемника приводит в результате к изменениям пути распространения. Для переданного непрерывного сигнала это вызывает изменения амплитуды и фазы сигнала в приемнике. Если все рассеивающие элементы, составляющие канал, являются стационарными, то при прекращении движения амплитуда и фаза полученного сигнала будут оставаться постоянными, т.е. канал якобы будет стационарным во времени. Как только движение возобновится, поведение канала снова станет переменным во времени. Поскольку характеристики канала зависят от положения передатчика и приемника, переменное во времени поведение в этом случае эквивалентно переменному пространственному поведению.

На рис. 15.8, *в* показана функция $R(\Delta t)$, обозначающая *пространственно-временную* корреляционную функцию; это автокорреляционная функция отклика канала на поданную синусоиду. Эта функция определяет степень корреляции между откликом канала на синусоиду, отправленную в момент времени t_1, и откликом на аналогичную синусоиду, отправленную в момент t_2, где $\Delta t = t_2 - t_1$. *Время когерентности* (coherence time) T_0 — это мера ожидаемого времени, за которое характеристика канала существенно инвариантна. Ранее измерение дисперсии сигнала и полосы когерентности проводилось с помощью широкополосных сигналов. Теперь для измерения нестационарной природы канала используется узкополосный сигнал [15]. Для измерения $R(\Delta t)$ можно передать одну и ту же синусоиду ($\Delta f = 0$) в моменты времени t_1 и t_2, после чего будет определена функция взаимной корреляции полученных сигналов. Функция $R(\Delta t)$ и параметр T_0 несут в себе информацию о скорости замирания в канале. Отметим, что для идеального *стационарного* канала (например, передатчик и приемник абсолютно неподвижны) отклик канала будет иметь сильную корреляцию для всех значений Δt; таким образом, $R(\Delta t)$ как функция Δt будет постоянной. Например, если расположение стационарного пользователя характеризуется нулем многолучевого распространения, то этот нуль остается неизменным, пока не появится какое-либо движение (либо со стороны передатчика или приемника, либо со стороны объектов на пути распространения). При использовании описанной ранее модели канала с плотным размещением рассеивающих элементов при постоянной скорости перемещения V и немодулированным непрерывным сигналом с длиной волны λ, нормированная $R(\Delta t)$ будет иметь следующий вид.

$$R(\Delta t) = J_0(kV\Delta t) \tag{15.23}$$

Здесь $J_0(\cdot)$ — функция Бесселя первого рода нулевого порядка [11], $V\Delta t$ — пройденное расстояние, а $k = 2\pi/\lambda$ — фазовая постоянная свободного пространства (переводящая расстояние в радианы). Время когерентности можно измерить с помощью времени либо пройденного расстояния (полагая скорость фиксированной). Аморозо (Amoroso) описал такое измерение, используя непрерывный сигнал и модель канала с плотным размещением рассеивающих элементов [17]. Он определил статистическую корреляцию между комбинацией принятой амплитуды и фазы, измеренных при определенном расположении антенны x_0, и соответствующей комбинацией амплитуды и фазы, измеренных при несколько смещенном расположении $x_0 + \zeta$, причем смещение измерялось в единицах длины волны λ. Когда смещение между двумя положениями антенны ζ составляет $0{,}4\lambda$, совокупные амплитуды и фазы полученного непрерывного сигнала являются статистически некоррелирующими. Иными словами, наблюдение

сигнала в точке x_0 не дает никакой информации о сигнале в точке $x_0 + \zeta$. Отметим также, что при данной скорости это смещение без труда преобразуется во время (время когерентности).

15.4.1.1. Независимость основных проявлений замирания

Для движущейся антенны замирание принятого несущего сигнала обычно рассматривается как случайный процесс, даже если замирание может быть полностью предопределено, исходя из расположения рассеивающих элементов и геометрии распространения между передатчиком и принимающей антенной. Это объясняется тем, что один и тот же сигнал, полученный двумя антеннами, разнесенными, по крайней мере, на 0,4λ, статистически не коррелирует [17, 18]. Поскольку такие малые расстояния (порядка 13 см для несущей 900 МГц) соответствуют статистической декорреляции принятых сигналов, основные проявления замирания, дисперсия сигнала и скорость замирания, могут рассматриваться независимо друг от друга. Здесь нам может помочь любой из случаев, изображенных на рис. 15.10. В каждый момент времени (соответствующий некоторому пространственному размещению) видим профиль интенсивности многолучевого распространения $S(\tau)$ как функцию задержки τ. Профили многолучевого распространения изначально определяются местностью (строения, растительность и т.д.). Рассмотрим рис. 15.10, *б*, где стрелочкой, помеченной *время* (можно было также пометить как *смещение антенны*), указано направление движения через области с различными профилями многолучевого распространения. При движении мобильного радиопередатчика к новому пространственному положению, которое характеризуется иным профилем, будут происходить изменения в состоянии замирания канала, как обуславливает профиль в новом местоположении. Однако вследствие того, что один профиль декоррелирует с другим уже на расстоянии порядка 13 см (для несущей 900 МГц), скорость таких изменений зависит только от скорости движения, но не от общей геометрии местности.

15.4.1.2. Понятие дуальности

Математическому понятию дуальности (duality) можно дать следующее определение: два процесса (функции, элемента или системы) *дуальны* друг другу, если математические соотношения между ними остаются неизменными с точностью до замены параметров. В этой главе интересно отметить дуальность при изучении соотношений во временной области по сравнению с соотношениями в частотной области.

Из рис. 15.8 можно определить функции, которые ведут себя одинаково в разных областях. Для понимания модели канала с замираниями рассмотрим *дуальные функции* (duals). Например, явление дисперсии сигнала можно описать в частотной области с помощью функции $R(\Delta f)$, как это показано на рис. 15.8, *б*. Эта функция несет в себе информацию о диапазоне частот, в котором два спектральных компонента полученного сигнала имеют большую вероятность амплитудной и частотной корреляции. Скорость замирания во временной области описывается функцией $R(\Delta t)$, как это показано на рис. 15.8, *в*. Эта функция несет в себе информацию об интервале времени, в течение которого два полученных сигнала имеют большую вероятность амплитудной и фазовой корреляции. На рисунке эти две корреляционные функции, $R(\Delta f)$ и $R(\Delta t)$, помечены как дуальные. Это отмечено также на рис. 15.1, где дуальными названы блоки 10 и 13, и на рис. 15.7, где дуальны механизм расширения во времени в частотной области и механизм нестационарности во временной области.

15.4.1.3. Категории ухудшения качества передачи вследствие нестационарного поведения канала, рассматриваемого во временной области

Нестационарную природу, или механизм скорости замирания в канале, можно рассматривать с позиции категорий ухудшения качества передачи, указанных на рис. 15.7, — *быстрого* и *медленного замирания*. Термин "быстрое замирание" (fast fading) используется для описания каналов, в которых $T_0 < T_s$, где T_0 — время когерентности канала, а T_s — длительность символа. Быстрое замирание описывает условие, когда временной интервал, в течение которого поведение канала имеет корреляционный характер, мал по сравнению со временем, необходимым для передачи символа. Таким образом, можно ожидать, что характер замирания в канале будет изменяться несколько раз за время передачи символа, что приведет к искажению вида узкополосного импульса. Данное искажение аналогично описанному ранее, которое вызывается внесенной каналом ISI, поскольку принятые компоненты сигнала не сильно коррелируют во времени. Поэтому быстрое замирание может искажать узкополосный импульс, что, как правило, приводит к частому появлению неустранимых ошибок. Такие искаженные импульсы вызывают проблемы синхронизации (сбои в работе приемников, использующих фазовую автоподстройку частоты). Кроме того, существуют трудности, связанные с адекватной разработкой согласованного фильтра.

Обычно говорят, что канал вносит медленное замирание (slow fading), если $T_0 > T_s$. Здесь временной интервал, в течение которого поведение канала имеет корреляционный характер, велик по сравнению со временем, необходимым для передачи символа. Следовательно, можно ожидать, что состояние канала будет оставаться практически неизменным в течение времени передачи символа. Распространяющиеся символы, вероятнее всего, не пострадают в результате искажений импульса, описанных ранее. Основное ухудшение качества передачи в канале с медленным замиранием, как и в случае с амплитудным замиранием, связано с уменьшением SNR.

15.4.2. Нестационарное поведение канала, рассматриваемое в области доплеровского сдвига

Аналогично характеристика нестационарной природы канала может быть представлена в области доплеровского сдвига (частот). На рис. 15.8, *г* показана *доплеровская спектральная плотность мощности* (или доплеровский спектр) $S(v)$, изображенная в виде функции от доплеровского сдвига частот. Для модели с плотным размещением рассеивающих элементов, вертикальной принимающей антенной с постоянным азимутальным усилением, однородным угловым распределением входящего сигнала по всем углам в интервале $(0, 2\pi)$ и немодулированным непрерывным сигналом спектр сигнала в точках приема будет иметь следующий вид.

$$S(v) = \frac{1}{\pi f_d \sqrt{1 - \left(\frac{v - f_c}{f_d}\right)^2}} \tag{15.24}$$

Равенство сохраняется для сдвига частот v, находящегося в интервале $\pm f_d$, в окрестности несущей частоты f_c; за пределами этого интервала оно обращается в нуль. Профиль радиочастотного доплеровского спектра, который описывается уравнением (15.24), имеет классическую форму чаши, что видно из рис. 15.8, *г*. Следует заметить, что профиль спектра

является результатом принятия модели канала с плотным размещением рассеивающих элементов. Уравнение (15.24) было введено для согласования экспериментальных данных, собранных для каналов мобильной радиосвязи [22]; однако для разных приложений профили спектра различны. Например, модель с плотным размещением рассеивающих элементов несправедлива для каналов радиосвязи внутри помещений; модель канала для областей внутри помещения предполагает, что $S(\nu)$ является равномерным спектром [23].

На рис. 15.8, *г* заостренность и крутизна границ спектра доплеровских частот является следствием резкого верхнего предела доплеровского сдвига, вызванного перемещением передвижной антенны среди стационарных рассеивающих элементов в модели плотного размещения. Наибольшая величина (бесконечность) $S(\nu)$ соответствует случаю, когда рассеивающий элемент находится прямо перед движущейся платформой антенны или прямо позади нее. В этом случае величина сдвига частот описывается формулой

$$f_d = \frac{V}{\lambda}, \tag{15.25}$$

где V — относительная скорость, а λ — длина волны сигнала. Если передатчик и приемник движутся навстречу друг другу, то f_d положительна, а если они удаляются друг от друга, то f_d отрицательна. Что касается рассеивающих элементов, находящихся в направлении поперечного излучения движущейся платформы, то для них величина частотного сдвига равна нулю. Отметим, что хотя доплеровские компоненты, поступившие точно под углами 0° и 180°, имеют бесконечно большую спектральную плотность мощности, это не представляет проблемы, поскольку угол имеет непрерывное распределение, а вероятность поступления компонентов точно под этими углами равна нулю [1, 18].

$S(\nu)$ является Фурье-образом $R(\Delta t)$. Известно, что Фурье-образ автокорреляционной функции временного ряда равен квадрату амплитуды Фурье-образа исходного временного ряда. Следовательно, измерения могут проводиться просто путем передачи синусоиды (узкополосный сигнал) и с использованием Фурье-анализа для получения спектра мощности полученной амплитуды [15]. Этот доплеровский спектр мощности канала дает информацию о спектральном расширении переданной синусоиды (импульса в частотной области) в области доплеровского сдвига. Как показано на рис. 15.8, функцию $S(\nu)$ можно рассматривать как дуальную профилю интенсивности многолучевого распространения $S(\tau)$, поскольку последняя несет информацию о расширении во времени переданного импульса в области задержки. Это также отмечено на рис. 15.1 в виде дуальности между блоками 7 и 16, а на рис. 15.7 — между механизмом расширения во времени в области задержки и механизмом нестационарного поведения канала в области доплеровского смещения.

Знание $S(\nu)$ делает возможным приблизительное вычисление величины расширения спектра как функции скорости изменения состояний канала. Ширина доплеровского спектра мощности (обозначенная f_d) в литературе называется по-разному: *доплеровское расширение* (Doppler spread), *скорость замирания* (fading rate), *ширина полосы замирания* (fading bandwidth) или *спектральное расширение* (spectral broadering). Уравнение (15.25) описывает доплеровский сдвиг частоты. В обычной для многолучевого распространения окружающей среде полученный сигнал движется по нескольким отраженным путям, каждый из которых имеет отличные от других расстояние и угол поступления. Доплеровский сдвиг для каждого из путей поступления сигнала, как правило, различен. Воздействие на полученный сигнал, как правило, проявляется в виде доплеровского расширения переданной частоты сигнала, а не как сдвиг. Нужно помнить, что доплеровское расширение f_d и время когерентности T_0 обратно пропор-

циональны (с точностью до постоянного множителя), что позволяет записать следующее приблизительное соотношение между этими двумя параметрами.

$$T_0 \approx \frac{1}{f_d} \tag{15.26}$$

Поэтому доплеровское расширение f_d (или $1/T_0$) рассматривается как обычная *скорость замирания* в канале. Ранее T_0 определялся как ожидаемый интервал времени, в течение которого отклик канала на синусоиду существенно инвариантен. Если T_0 определять более точно, как интервал времени, в течение которого отклики канала на синусоиды имеют между собой корреляцию не менее 0,5, соотношение между T_0 и f_d будет приблизительно следующим.

$$T_0 \approx \frac{9}{16\pi f_d} \tag{15.27}$$

Известным эмпирическим правилом является определение T_0 через геометрическое среднее уравнений (15.26) и (15.27). Это приводит к следующему.

$$T_0 \approx \sqrt{\frac{9}{16\pi f_d^2}} = \frac{0{,}423}{f_d} \tag{15.28}$$

Для мобильной радиосвязи на частоте 900 МГц, на рис. 15.12 показано типичное влияние релеевского замирания на огибающую амплитуды сигнала в зависимости от времени [1]. На рисунке показано, что расстояние, пройденное мобильным аппаратом за интервал времени, соответствующий двум соседним нулям (мелкомасштабное замирание), равно по порядку половине длины волны ($\lambda/2$). Таким образом, из рис. 15.12 и уравнения (15.25) ясно, что время, требуемое для прохождения расстояния $\lambda/2$ (равное приблизительно времени когерентности) при движении с постоянной скоростью V, будет следующим.

Рис. 15.12. Типичный профиль огибающей при релеевском замирании на частоте 900 МГц. (Rappaport T. S. Wireless Communications. Chapter 4, Prentice-Hall, Upper Saddle River, New Jersey, 1996.*)*

$$T_0 \approx \frac{\lambda/2}{V} = \frac{0{,}5}{f_d} \tag{15.29}$$

Таким образом, когда расстояние между периодами замирания приблизительно равно $\lambda/2$, как показано на рис. 15.12, результирующее выражение для T_0 в уравнении (15.29) близко к геометрическому среднему, показанному в уравнении (15.28). Из уравнения (15.29), используя параметры, показанные на рис. 15.12 (скорость — 120 км/ч, несущая частота — 900 МГц), можно получить, что время когерентности канала — приблизительно 5 мс, а доплеровское расширение (скорость замирания в канале) — приблизительно 100 Гц. Следовательно, если в этом примере представлен канал, по которому передается оцифрованная речь с типичной скоростью 10^4 символов/с, скорость замирания значительно меньше скорости передачи символов. При таких условиях канал будет проявлять эффекты медленного замирания. Нужно сказать, что если бы абсцисса на рис. 15.12 была проградуирована в единицах длины волны, а не в единицах времени, то отображенные характеристики замирания выглядели бы так же для любой радиочастоты и любой скорости движения антенны.

15.4.2.1. Аналогия спектрального расширения в каналах с замираниями

Рассмотрим причину, по которой сигнал испытывает спектральное расширение при распространении или приеме подвижной платформой, и то, почему спектральное расширение (называемое также скоростью замирания в канале) является функцией скорости движения. Для объяснения этого явления можно воспользоваться следующей аналогией. На рис. 15.13 показана манипуляция цифровым сигналом (такая, как амплитудная или частотная манипуляция), где тон $\cos 2\pi f_c t$, определенный в интервале $-\infty < t < \infty$, характеризуется в частотной области импульсами ($\pm f_c$). Такое представление в частотной области является идеальным (т.е. нулевая ширина полосы частот), поскольку тон — это одна частота с бесконечной длительностью. В практических приложениях при передаче цифрового сигнала происходит включение и выключение (манипуляция) сигналов с требуемой скоростью. Манипуляция может рассматриваться как умножение тона бесконечной длительности на рис. 15.13, *а* на идеально прямоугольную функцию манипуляции (коммутации) на рис. 15.13, *б*. Описание такой коммутационной функции в частотной области имеет вид $\mathrm{sinc}\, fT$ (см. приложение А, табл. А.1).

На рис. 15.13, *в* показан полученный в результате умножения тон $\cos 2\pi f_c t$, теперь ограниченный по длительности. Результирующий спектр получается путем свертки спектральных импульсов (рис. 15.13, *а*) с функцией $\mathrm{sinc}\, fT$ (рис. 15.13, *б*); этот результирующий расширенный спектр показан на 15.13, *в*. Далее видно, что если передача сигналов происходит с более высокой скоростью, которой соответствует прямоугольник меньшей длины (рис. 15.13, *г*), то для результирующего спектра сигнала (рис. 15.13, *д*) характерно большее расширение спектра. Изменение состояния канала с замиранием является в какой-то мере аналогом амплитудной модуляции цифровых сигналов. Канал ведет себя как коммутатор, "включающий и выключающий" сигнал. Чем выше скорость изменения состояния канала, тем большее расширение спектра испытывает сигнал, распространяющийся по такому каналу. Это неточная аналогия, поскольку включение и выключение сигналов может привести к разрыву фазы, в то время как для типичных рассеивающих элементов при многолучевом распространении характерна непрерывность фазы.

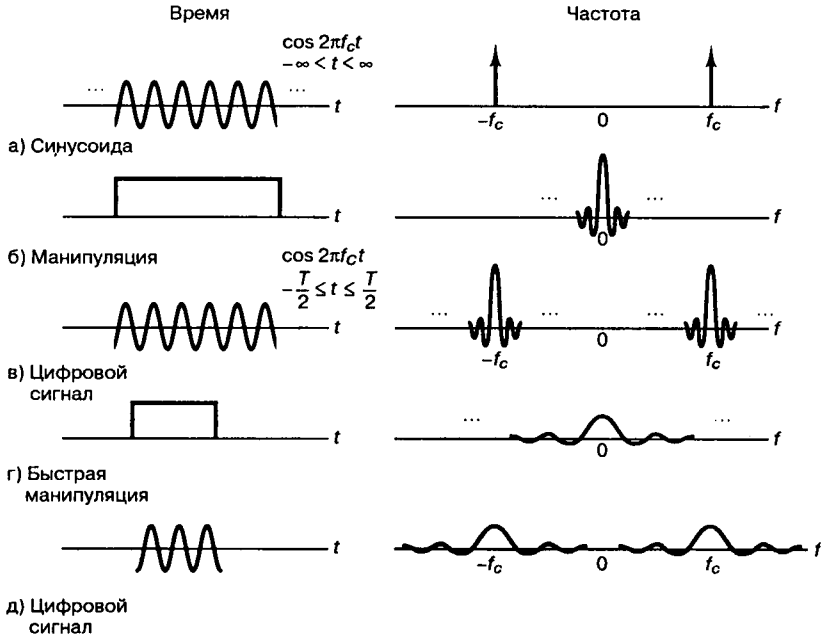


Рис. 15.13. Аналогия между расширением спектра при замирании и расширением спектра при манипуляции цифровым сигналом

15.4.2.2. Категории ухудшения характеристик вследствие нестационарной природы канала, рассматриваемые в области доплеровского сдвига

Говорят, что в канале имеет место быстрое замирание, если скорость передачи символов $1/T_s$ (приблизительно равная скорости передачи сигналов или ширине полосы частот W) меньше скорости замирания $1/T_0$ (приблизительно равной f_d), т.е. быстрое замирание характеризуется следующими соотношениями.

$$W < f_d \tag{15.30}$$

или

$$T_s > T_0 \tag{15.31}$$

Наоборот, в канале имеет место медленное замирание, если скорость передачи сигналов больше скорости замирания. Таким образом, чтобы избежать искажения сигнала, вызванного быстрым замиранием, нужно создать канал, который будет подвержен медленному замиранию, что обеспечивается за счет большей скорости передачи сигнала по сравнению со скоростью замирания.

$$W > f_d \tag{15.32}$$

или

$$T_s < T_0 \tag{15.33}$$

В уравнении (15.22) показано, что вследствие дисперсии сигнала ширина полосы когерентности f_0 устанавливает *верхний предел* скорости передачи сигналов, при которой отсутствует частотно-селективное искажение. Аналогично в уравнении (15.32) показано, что в результате доплеровского расширения скорость замирания в канале f_d устанавливает нижний предел скорости передачи сигнала, при которой отсутствует искажение, связанное с

быстрым замиранием. Для систем связи высоких частот, если телетайпное сообщение или сообщение в азбуке Морзе было передано с низкой скоростью передачи данных, в каналах часто наблюдаются характерные особенности быстрого замирания. В то же время большинство современных наземных каналов мобильной радиосвязи чаще всего можно охарактеризовать как каналы с медленным замиранием.

Уравнений (15.32) и (15.33) недостаточно для описания желаемого поведения канала. Лучшим способом задания требований для избежания быстрого замирания было бы условие $W >> f_d$ (или $T_s << T_0$). Если это условие не удовлетворено, то случайная частотная модуляция (frequency modulation — FM), вызванная переменными доплеровскими сдвигами, будет существенно ухудшать характеристики системы. Эффект Доплера приводит к частому появлению неустранимых ошибок, которые нельзя компенсировать простым увеличением E_b/N_0 [24]. Это частое появление неустранимых ошибок наиболее резко выражено во всевозможных схемах передачи, использующих модуляцию несущей фазы. Отдельный отраженный доплеровский путь (без рассеивающих элементов) регистрирует мгновенный сдвиг, традиционно вычисляемый как $f_d = V/\lambda$. Однако комбинация отраженных и многолучевых компонентов порождает довольно сложную временную зависимость мгновенной частоты, которая может вызвать колебания частоты, сильно превышающие $\pm V/\lambda$ при восстановлении информации детектором мгновенной частоты (который является нелинейным устройством) [25]. На рис. 15.14 показано, как это происходит. В результате движения переносного устройства в момент времени t_1 отраженный вектор поворачивается на угол θ, в то время как суммарный вектор поворачивается на угол ϕ, который приблизительно в четыре раза больше θ. Скорость изменения фазы в момент времени, близкий к этому конкретному периоду замирания, приблизительно равна скорости изменения отраженной доплеровской фазы, умноженной на 4. Следовательно, сдвиг мгновенной частоты $d\phi/dt$ был бы в 4 раза больше отраженного доплеровского сдвига. Образование резких максимумов мгновенных сдвигов частот в моменты времени, близкие к сильному замиранию, подобно появлению "щелчков" или "пиков", характерных для сигнала FM. На рис. 15.15 продемонстрирована серьезность этой проблемы. На рисунке показан график зависимости частоты появления однобитовых ошибок от E_b/N_0 для передачи сигнала $\pi/4$ с модуляцией DQPSK на частоте $f_0 = 850$ МГц для различных моделируемых скоростей переносного устройства [26]. Должно быть ясно, что при высоких скоростях кривая характеристики спускается до уровня частоты появления ошибок, который может быть недопустимо высок. В идеале, когерентный демодулятор, который захватывает и отслеживает информационный сигнал, должен был бы гасить влияние такого шума частотной модуляции, таким образом исключая влияние доплеровского сдвига. Однако при больших значениях f_d восстановление несущей реализовать сложно, поскольку нужно построить очень широкополосные (по отношению к скорости передачи данных) схемы фазовой автоподстройки частоты (phase-lock loop — PLL, ФАПЧ). Для приложений речевой связи с частотой появления ошибок в интервале от 10^{-3} до 10^{-4} учитывается большое значение доплеровского сдвига, которое считается равным по порядку величине $0{,}01 \times W$. Следовательно, во избежание искажений, вызванных быстрым замиранием, и частого появления неустранимых ошибок, вызванных эффектом Доплера, скорость передачи сигнала должна превышать скорость замирания в 100–200 раз [27]. Точное значение зависит от типа модуляции сигнала, строения приемника и требуемой частоты появления ошибок [1, 25–29]. Девериан (Davarian) [29] показал, что система, отслеживающая частоту, может посредством дифференциальной манипуляции с минимальным сдвигом (differential minimum-shift keying — DMSK)

снизить (но не устранить) частоту появления неустранимых ошибок в мобильных системах связи.

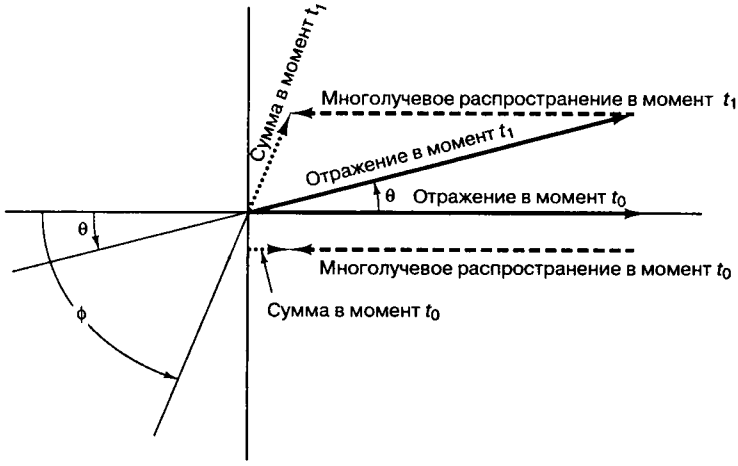


Рис. 15.14. Комбинация отраженного и многолучевого компонентов может давать большее колебание частоты, чем $\pm v/\lambda$. (Источник: Amoroso F. Instantaneous Frequency Effects in a Doppler Scattering Environment. IEEE International Conference on Communications, June 7–10, 1987, pp. 1458–1466.*)*

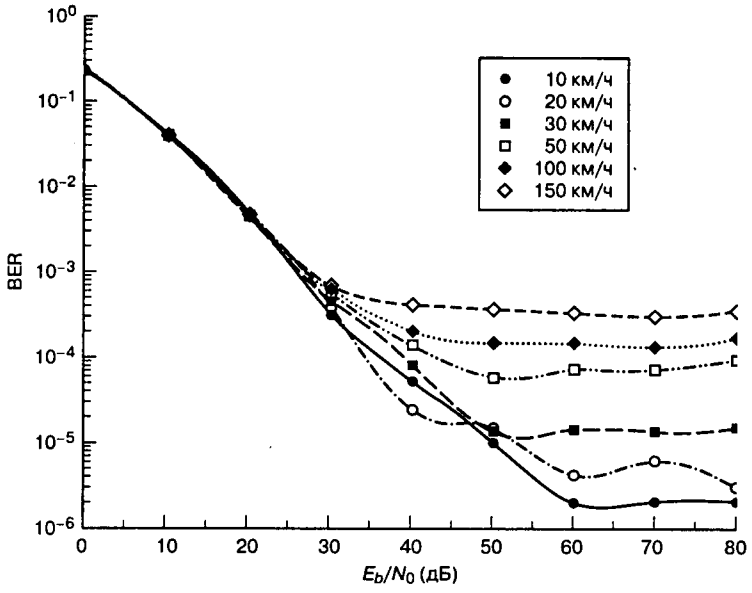


Рис. 15.15. Зависимость вероятности частоты появления ошибки от E_b/N_0 для схемы π/4 DQPSK при разных скоростях движения: f_c = 850 МГц, $R_s = 24 \times 10^3$ символов/с. (Источник: Fung V., Rappaport T. S. and Thoma B. Bit-Error Simulation for π/4 DQPSK Mobile Radio Communication Using Two-Ray and Measurement-Based Impulse Response Models. IEEE Journal on Selected Areas in Communication, Vol. 11, n. 3, April, 1993, pp. 393–405.*)*

15.4.3. Релеевский канал с медленным и амплитудным замиранием

При дискретном многолучевом канале с комплексной огибающей $g(t)$, описываемой уравнением (15.3), демодулированный сигнал (шумом пренебрегаем) описывается уравнением (15.10), которое повторно приводится ниже.

$$z(t) = \sum \alpha_n(t) e^{-2\pi i f_c \tau_n(t)} R[t - \tau_n(t)] e^{i\phi(t-\tau_n)} \qquad (15.34, \text{а})$$

Здесь $R(t) = |g(t)|$ — модуль огибающей, а $\phi(t)$ — ее фаза. Предположим, что канал проявляет амплитудное замирание, так что многолучевые компоненты не разрешаются. Тогда слагаемые $\{\alpha_n(t)\}$ в уравнении (15.34, а) за один период передачи сигнала T нужно выразить как результирующую амплитуду $\alpha(T)$ всех n векторов, полученных за этот промежуток времени. Аналогично фазовые члены в уравнении (15.34, а) за один период передачи сигнала нужно выразить как результирующую фазу $\theta(T)$ всех n замирающих векторов плюс информационную фазу, полученную за этот промежуток. Пусть канал проявляет медленное замирание, так что с помощью применения контура фазовой автоподстройки частоты (phase-lock loop — PLL, ФАПЧ) или другого подходящего метода фазу (с незначительной погрешностью) можно вычислить из полученного сигнала. Следовательно, в канале с медленным и амплитудным замиранием для каждого периода передачи сигнала можно записать включающую шум $n_0(T)$ тестовую статистику вне демодулятора.

$$z(T) = \alpha(T) R(T) e^{-i[\theta(T) - \phi(T)]} + n_0(T) \qquad (15.34, \text{б})$$

Далее для простоты вместо $\alpha(T)$ будем писать α. При двоичной передаче по каналу AWGN с фиксированным коэффициентом замирания $\alpha = 1$ вероятность появления битовой ошибки для основной когерентной и некогерентной схем PSK и ортогональной схемы FSK представлена в главе 4, табл. 4.1. Все графики зависимости вероятности появления ошибочного бита от E_b/N_0 для таких схем передачи сигнала показывают классическую экспоненциальную зависимость ("водопадоподобный" вид, ассоциируемый с каналом AWGN). Однако, при условии многолучевого распространения, если отсутствует отраженный компонент сигнала, α является случайной переменной с релеевским распределением; или, что равнозначно, α^2 описывается плотностью вероятности χ^2. На рис. 15.16 отображен график вероятности ошибки для такого релеевского замирания. Если $(E_b/N_0)\,\mathbf{E}(\alpha^2) >> 1$, где $\mathbf{E}(\cdot)$ выражает математическое ожидание, то формулы для вероятности битовой ошибки при использовании основных схем двоичной передачи сигналов, показанных на рис. 15.16, даны в табл. 15.1. Каждая схема передачи сигнала, которая в канале AWGN давала график в виде водопада, представленный на рис. 4.25, теперь, в результате релеевского замирания, описывается приблизительно линейной функцией.

Таблица 15.1. Релеевская граничная вероятность битовой ошибки (где $(E_b/N_0)\,\mathbf{E}(\alpha^2) >> 1$)

Модуляция	P_B
PSK (когерентная)	$\dfrac{1}{4(E_b/N_0)\mathbf{E}(\alpha^2)}$

Окончание табл. 15.1

Модуляция	P_B
DPSK (дифференциально-когерентная)	$\frac{1}{2(E_b/N_0)\mathbf{E}(\alpha^2)}$
Ортогональная FSK (когерентная)	$\frac{1}{2(E_b/N_0)\mathbf{E}(\alpha^2)}$
Ортогональная FSK (некогерентная)	$\frac{1}{(E_b/N_0)\mathbf{E}(\alpha^2)}$

Proakis J. D. *Digital Communications.* McGraw-Hill, New York, 1983.

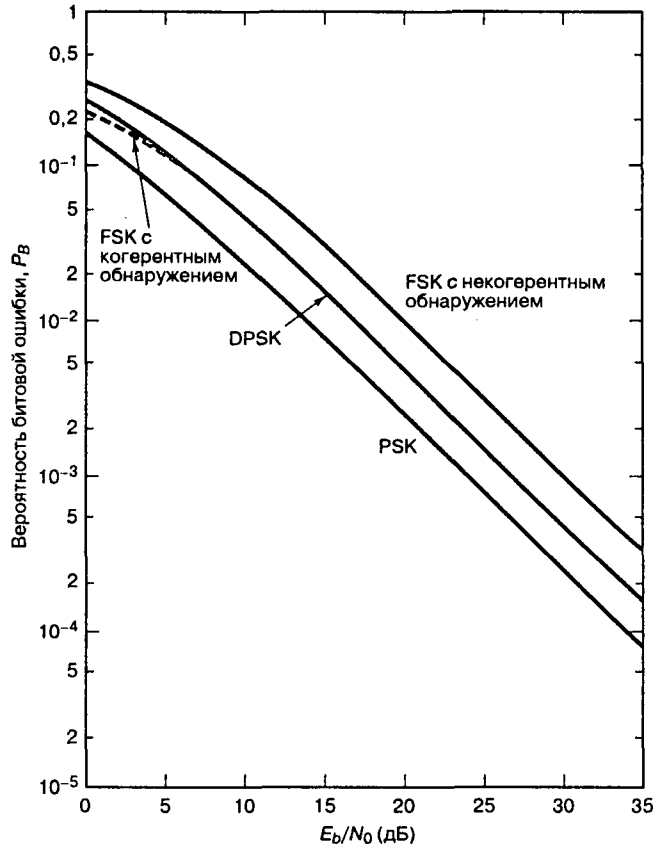


Рис. 15.16. Вероятность ошибки при двоичной передаче по каналу с медленным релеевским замиранием. (Источник: Proakis J. G. Digital Communications, McGraw-Hill Book Company, New York, 1983.*)*

15.5. Борьба с ухудшением характеристик, вызванным эффектами замирания

В подписи к рис. 15.17 "хорошая, плохая, ужасная" отражены три основные категории характеристик, выраженных через вероятность битовой ошибки в зависимости от E_b/N_0. Крайняя левая кривая, имеющая экспоненциальную форму, соответствует ожидаемому поведению данной зависимости при использовании любых номинальных схем модуляции при AWGN. Видно, что при разумном уровне E_b/N_0 можно ожидать хорошей достоверности передачи. Средняя кривая, названная *релеевским пределом*, демонстрирует ухудшение достоверности передачи, вытекающее из уменьшения E_b/N_0, что характерно для амплитудного или медленного замирания при отсутствии компонента, распространяющегося вдоль линии прямой видимости. Кривая является функцией, обратно пропорциональной E_b/N_0, так что для значений E_b/N_0, представляющих практический интерес, характеристики будут "плохими". При релеевском замирании, чтобы указать на то, что проводится усреднение по "лучшим " и "худшим" случаям замирания, часто вводятся параметры с чертой. Следовательно, часто можно увидеть графики вероятности битовой ошибки с усредненными параметрами, обозначенными $\overline{P_B}$ и $\overline{E_B}/N_0$. Такое обозначение акцентирует внимание на том, что каналы с замираниями имеют память; таким образом, принятые выборки сигнала коррелируют друг с другом во времени. Следовательно, при создании таких графиков вероятности ошибки для каналов с замиранием, необходимо изучить процесс в течение промежутка времени, намного превышающего время когерентности канала. Кривая, достигающая непоправимого уровня ошибок, часто называется *дном ошибок* (error floor) и представляет "ужасную" характеристику, при этом вероятность битовой ошибки может выходить на постоянный уровень, близкий к 0,5. Это соответствует эффекту сильного ухудшения характеристик, который может проявиться при частотно-селективном или быстром замирании.

Если в результате замирания канал вносит искажения в сигнал, для системы может быть характерен неисправимый уровень ошибок, превышающий допустимую частоту появления ошибок. В этом случае никакое увеличение E_b/N_0 не поможет достичь желаемого уровня достоверности передачи, и единственно доступным подходом, допускающим улучшение, является использование каких-либо иных методов устранения или уменьшения искажений. Метод борьбы зависит от того, вызвано ли искажение частотно-селективным или быстрым замиранием. Когда искажение сигнала будет смягчено, зависимость P_B от E_b/N_0 может перейти из категории "ужасно" в категорию, близкую к "плохо" — кривая релеевского предела. Далее можно использовать дополнительные методы борьбы с эффектами, вызванными замиранием, приложив усилия к приближению характеристики системы к характеристикам канала AWGN, применив некоторые виды разнесения, чтобы снабдить приемник набором некоррелирующих копий сигнала, и воспользовавшись мощным кодом коррекции ошибок.

На рис. 15.18 перечислено несколько методов борьбы как с искажением сигнала, так и с уменьшением SNR. Если рис. 15.1 и 15.7 играют роль проводника по описанию явлений замирания и их следствий, то рис. 15.18 аналогичным образом может служить для описания методов борьбы с этими явлениями и их следствиями. Предлагаемые подходы используются, когда проектирование системы рассматривается в два основных этапа: первый — выбор метода борьбы, уменьшающего или устраняющего любые ухудшения характеристик, вызванные искажениями; второй — выбор типа разнесения, который позволил бы наилучшим образом приблизиться к характеристикам канала AWGN.

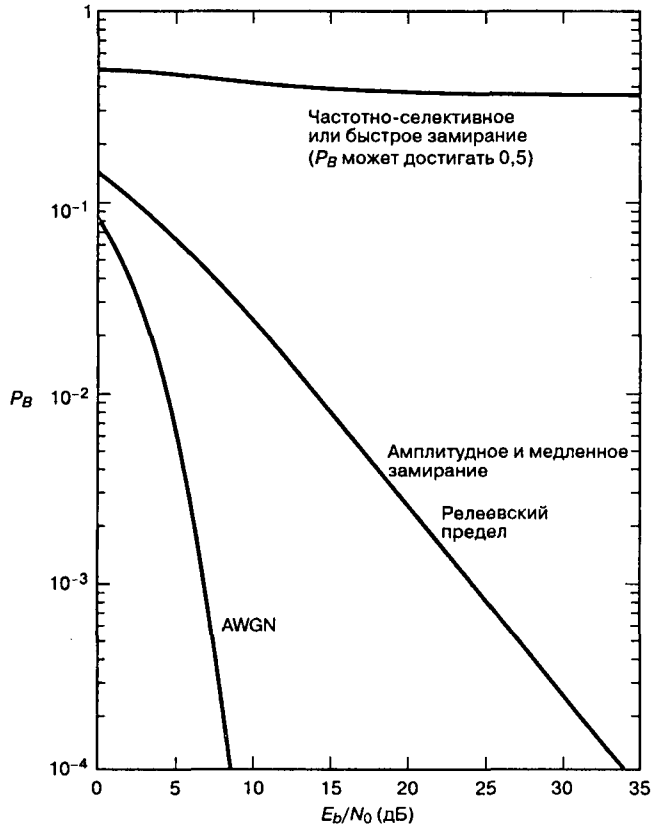


Рис. 15.17. Достоверность передачи сигналов: хорошая, плохая, ужасная

Меры против искажения	Меры против потери SNR
Частотно-селективное искажение • Адаптивное выравнивание (например, использование эквалайзеров с обратной связью по принятию решения или эквалайзеров Витерби) • Расширение спектра (методом прямой последовательности или перестройки частоты) • Ортогональное FDM (OFDM) • Контрольный сигнал	**Быстрое и медленное замирание** • Некоторое разнообразие для получения дополнительных некоррелирующих оценок сигнала • Коды коррекции ошибок
Искажение вследствие быстрого замирания • Устойчивая модуляция • Избыточность для увеличения скорости передачи сигналов • Кодирование и чередование	**Типы разнесения** • Время (например, чередование) • Частота (например, расширение полосы, спектра методом прямой последовательности или перестройки частоты) • Пространство (например, разнесенные принимающие антенны) • Поляризация

Рис. 15.18. Основные типы борьбы с искажением и снижением SNR

15.5.1. Борьба с частотно-селективными искажениями

Для борьбы с вызванной каналом ISI, которая возникает вследствие частотно-селективного замирания, может использоваться выравнивание. Иными словами, выравнивание изменяет характеристики системы, описываемые кривой, которая на рис. 15.17 названа "ужасно", на характеристики, близкие к кривой "плохо". Процесс выравнивания для уменьшения воздействия ISI заключается в использовании методов, собирающих рассеянную энергию символа в ее исходный временной интервал. По сути, эквалайзер (устройство выравнивания) является обратным фильтром канала. Если канал является частотно-селективным, эквалайзер усиливает частотные компоненты с малыми амплитудами и ослабляет с большими. Целью комбинации канала и выравнивающего фильтра является получение плоской частотной характеристики и линейного изменения фазы [30]. Поскольку в мобильных системах характеристика канала меняется со временем, выравнивающий фильтр должен изменяться или приспосабливаться к нестационарным характеристикам канала. Следовательно, такие фильтры являются адаптивными устройствами, которые предназначены не только для борьбы с искажениями; они также обеспечивают разнесение. Поскольку ослабление искажений выполняется путем сбора рассеянной энергии символа в исходный временной интервал символа (так, чтобы это не мешало обнаружению других символов), эквалайзер попутно предоставляет приемнику энергию символа, которая в противном случае была бы утрачена.

Эквалайзер с обратной связью по решению (decision feedback equalizer — DFE) имеет участок прямой связи, являющийся линейным трансверсальным фильтром [30], размер регистра и весовые коэффициенты отводов которого выбраны так, чтобы когерентно собирать практически всю энергию текущего символа. Эквалайзер DFE имеет также участок обратной связи, который удаляет энергию, оставшуюся от ранее обнаруженных символов [10, 30–32]. Принцип работы DFE основан на следующем: когда обнаруживается информационный символ, ISI, вносимая им в последующие символы, может быть оценена и вычтена до обнаружения последующих символов (см. раздел 3.4.3.2.).

Эквалайзер, работающий по принципу оценки последовательности с максимальным правдоподобием (maximum-likelihood sequence estimation — MLSE), проверяет все возможные последовательности данных (вместо того чтобы обнаруживать каждый полученный символ отдельно) и выбирает ту, которая является наиболее вероятной из всех кандидатов. Эквалайзер MLSE впервые был предложен Форни (Forney) [33] и реализован с использованием алгоритма декодирования Витерби [34]. Принцип MLSE оптимален в том смысле, что он минимизирует вероятность ошибки последовательности. Поскольку при реализации эквалайзера MLSE обычно используется алгоритм декодирования Витерби, это устройство часто называют *эквалайзером Витерби* (Viterbi equalizer). Позже в этой главе будет продемонстрировано адаптивное выравнивание, применяемое в системе GSM (Global System for Mobile — глобальная система мобильной связи), где используется эквалайзер Витерби.

Расширение спектра методом прямой последовательности (direct-sequence spread-spectrum — DS/SS) может использоваться для уменьшения искажений, вызванных частотно-селективной ISI, поскольку отличительной особенностью систем расширенного спектра является их способность отфильтровывать помехи, а ISI — это один из видов помех. Рассмотрим систему DS/SS, в которой используется двоичная фазовая манипуляция (binary phase-shift keying — BPSK) и канал связи, содержащий один прямой и один отраженный путь. Пусть распространение от передатчика к приемнику приводит к многолучевому распространению сигнала, запаздывающего на τ по сравнению с прямым сигналом. Пренебрегая шумом, многолучевой сигнал можно выразить следующим образом.

$$r(t) = Ax(t)g(t)\cos(2\pi f_c t) + \alpha Ax(t-\tau)g(t-\tau)\cos(2\pi f_c t + \theta) \qquad (15.35)$$

Здесь $x(t)$ — информационный сигнал, $g(t)$ — шумоподобный (pseudonoise — PN) код расширения, τ — разность во времени распространения между двумя путями, а α — поглощение многолучевого сигнала по сравнению с сигналом, распространяющимся по прямому пути. Кроме того, предполагается, что случайная фаза θ равномерно распределена в интервале $(0, 2\pi)$. Приемник умножает поступающий сигнал $r(t)$ на код $g(t)$. Если приемник синхронизирован с сигналом, распространяющимся по прямому пути, умножение на кодовый сигнал дает следующее.

$$r(t)g(t) = Ax(t)g^2(t)\cos(2\pi f_c t) + \alpha Ax(t-\tau)g(t)g(t-\tau)\cos(2\pi f_c t - \theta), \qquad (15.36)$$

где $g^2(t) = 1$. Если τ больше длительности элементарного псевдошумового сигнала, тогда

$$\left|\int g(t)g(t-\tau)dt\right| << \left|\int g^2(t)dt\right| \qquad (15.37)$$

по некоторому удобному интервалу интегрирования (корреляция). Таким образом, система расширенного спектра эффективно устраняет многолучевую интерференцию за счет корреляционного (по коду) приемника. Хотя наличие введенной каналом ISI обычно не заметно для систем DS/SS, такие системы подвержены потерям энергии, содержащейся в многолучевых компонентах, отклоняемых приемником. Необходимость сбора утраченной энергии, принадлежащей подобным многолучевым элементарным сигналам, стала причиной разработки RAKE-приемника (RAKE receiver) [35–37]. В этом приемнике каждому многолучевому компоненту выделяется отдельный коррелятор. Приемник когерентно суммирует энергию каждого луча, избирательно задерживая их (более ранние компоненты задерживаются дольше) таким образом, чтобы они объединялись когерентно.

Ранее описывался канал, который можно классифицировать как канал с амплитудным замиранием, но который время от времени (когда нуль частотной передаточной функции канала попадает на центр полосы сигнала) проявляет частотно-селективное искажение. Использование DS/SS является удобным методом борьбы с таким искажением, поскольку широкополосный сигнал SS может охватить большое число периодов характеристики частотно-селективного ослабления. Таким образом, большая часть энергии импульса пройдет через среду рассеивающих элементов, что отличается от воздействия нулей канала на узкополосный сигнал [17] (см. рис. 15.9, *в*). Способность спектра сигнала охватывать большое число периодов передаточной функции частотно-селективного канала является ключевой, позволяющей сигналу DS/SS преодолевать искажающее влияние многолучевой среды. Необходимое условие: ширина полосы частот расширенного спектра W_{ss} (или скорость передачи элементарных сигналов R_{ch}) должна быть больше ширины полосы когерентности f_0. Чем больше отношение W_{ss} к f_0, тем более эффективным будет подавление искажений. Временное представление такого подавления выражено в уравнениях (15.36) и (15.37). Таким образом, чтобы разрешить многолучевые компоненты (либо отбросить их, либо использовать в RAKE-приемнике), необходимо, чтобы дисперсия сигнала расширенного спектра была больше скорости передачи элементарных сигналов.

Расширение спектра методом скачкообразной перестройки частоты (frequency hopping spread spectrum — FH/SS) может использоваться для борьбы с искажениями, вызванными частотно-селективным замиранием, причем скорость изменения частоты

должна быть не меньше скорости передачи символов. Ослабление искажений происходит в данном случае благодаря механизмам, отличным от использованных в DS/SS. Приемники с перестройкой частоты избегают эффектов искажения вследствие многолучевого распространения, быстро меняя в передатчике полосу несущей частоты; таким образом, помехи не возникают, поскольку изменение положения полосы частот приемника происходит до поступления многолучевого сигнала.

Ортогональное уплотнение с частотным разделением (orthogonal frequency-division multiplexing — OFDM) может использоваться при передаче сигнала в каналах с частотно-селективным замиранием для увеличения периода передачи символа, что позволит избежать применения эквалайзера. Принцип работы заключается в разделении (разуплотнении) последовательности с высокой скоростью передачи на N групп символов так, чтобы каждая группа содержала последовательность с более низкой скоростью передачи символов (в N раз меньшую), чем у исходной последовательности. Полоса сигнала состоит из N ортогональных несущих сигналов, каждый из которых модулируется отличной от других группой символов. Целью является снижение скорости передачи символов (скорости передачи сигналов) $W \approx 1/T_s$ на каждой несущей так, чтобы она была меньше ширины полосы когерентности канала f_0. Метод OFDM, изначально именуемый *Kineplex*, — это метод, реализованный в мобильных системах радиосвязи США [38] и использованный в Европе под названием кодированное OFDM (Coded OFDM — COFDM) в телевидении высокой четкости (high-definition television — HDTV) [39].

Контрольный сигнал (pilot signal) — это сигнал, способствующий когерентному обнаружению сигналов. Контрольные сигналы можно реализовать в частотной области как внутриполосные тоны [40] или во временной области как цифровые последовательности, которые могут также предоставлять информацию о состоянии канала и таким образом улучшать достоверность передачи при замирании [41].

15.5.2. Борьба с искажениями, вызванными быстрым замиранием

Искажения, вызванные быстрым замиранием, приводят к необходимости использования помехоустойчивой схемы модуляции (некогерентной или дифференциально-когерентной), которая не требует сопровождения фазы и снижает время интеграции детектора [19]. Кроме того, можно увеличить скорость передачи символов $W \approx 1/T_s$, чтобы она превышала скорость замирания $f_d \approx 1/T_0$, путем введения избыточности сигнала. Кодирование с коррекцией ошибок может также вносить улучшения; взамен повышения энергии сигнала код снижает E_b/N_0, требуемое для получения заданной достоверности передачи. При данном E_b/N_0 при наличии кодирования дно ошибок вне демодулятора не будет опускаться, при этом вне декодера может быть достигнута меньшая частота появления ошибок [19]. Таким образом, при кодировании можно получить приемлемую достоверность передачи и, по сути, допустить более высокий уровень ошибок в сигналах, поступающих от демодулятора, который в противном случае был бы неприемлем. Чтобы воспользоваться преимуществами кодирования, ошибки вне демодулятора должны не коррелировать (что обычно бывает в среде с быстрым замиранием) либо в систему должно внедряться устройство чередования.

Если одновременно происходит ухудшение характеристик в результате быстрого замирания и частотной избирательности, улучшение может обеспечить один интересный метод фильтрации. Частотно-селективное ухудшение характеристик можно снизить, используя набор сигналов с OFDM. В то же время обычные сигналы OFDM искажаются в результате быстрого замирания (доплеровское расширение нарушает ортогональность поднесущих

OFDM). В этом случае для формирования сигнала во временной области и кодирования с частичным откликом (см. раздел 2.9) с целью уменьшения боковых спектральных лепестков набора сигналов (что помогает сохранить их ортогональность) используется метод полифазной фильтрации [24]. Процесс вносит известную ISI и помехи соседнего канала (adjacent channel interference — ACI), которые затем устраняются последующей обработкой на эквалайзере и применением гасящего фильтра [43].

15.5.3. Борьба с уменьшением SNR

После реализации некоторых методов борьбы с ослаблением сигнала вследствие частотно-селективного и быстрого замирания, следующим шагом является использование методов разнесения для перемещения рабочей точки системы с кривой достоверности передачи, помеченной "плохо" на рис. 15.17, на кривую, приближающуюся к характеристике AWGN. Термин "разнесение" (diversity) применяется для обозначения различных методов, пригодных для некоррелированного воспроизведения приемником интересующего сигнала. Некоррелированность является здесь важной особенностью, поскольку дополнительные копии сигнала ничем не помогли бы приемнику, если бы все эти копии были одинаково плохи. Ниже перечислены некоторые способы реализации методов разнесения.

- *Разнесение во времени* (time diversity) может обеспечиваться путем передачи сигнала в L различных временных интервалах с разнесением не менее чем на T_0. Пример разнесения во времени — чередование, использованное совместно с кодированием с коррекцией ошибок.
- *Разнесение по частоте* (frequency diversity) может обеспечиваться путем передачи сигнала на L различных несущих с частотным разнесением не менее f_0. Пример разнесения по частоте — расширение полосы частот. Полоса частот сигнала W расширяется так, чтобы превышать f_0, предоставляя приемнику несколько независимо замирающих копий сигнала. При этом достигается частотное разнесение порядка $L = W/f_0$. Когда W становится больше f_0, то, если не используется выравнивание, существует возможность частотно-селективного искажения. Таким образом, расширенная полоса частот может улучшить характеристики системы (посредством разнесения) только в том случае, если ослаблено частотно-селективное искажение, связанное с этим разнесением.
- *Системы расширенного спектра* (spread-spectrum systems) — это системы, в которых для исключения интерферирующих сигналов используются методы расширения полосы частот. Если спектр расширяется методом прямой последовательности (direct-sequence spread-spectrum — DS/SS), то, как было показано ранее, многолучевые компоненты отбрасываются, если задержка их поступления превышает длительность одного элементарного сигнала. Однако чтобы приблизиться к характеристикам AWGN, необходимо компенсировать потерю энергии, которая содержится в этих отброшенных компонентах. RAKE-приемник (описанный позже) дает возможность когерентно объединять энергию нескольких многолучевых компонентов, поступивших по различным путям (с достаточно различающимися задержками). Таким образом, можно сказать, что при использовании RAKE-приемника в системе DS/SS получается разнесение по пути распространения. RAKE-приемник нужен при приеме, когерентном по фазе; но при дифференциально-когерентном обнаружении битов можно реали-

зовать простую задержку (равную комплексно сопряженной длительности одного бита) [44].

- *Расширение спектра методом скачкообразной перестройки частоты (frequency-hopping spread-spectrum — FH/SS)* также иногда используется в качестве механизма разнесения. В системе GSM применяется медленная перестройка частоты (217 скачков/с) для компенсации в трех случаях, когда объект движется очень медленно (или совсем не движется) и испытывает сильное замирание вследствие спектральных нулей.
- *Пространственное разнесение* (spatial diversity) обычно осуществляется посредством множественных принимающих антенн, разнесенных на расстояние, не меньшее 10 длин волн при размещении на базовой станции (и меньше, при размещении на мобильном объекте). Для выбора наилучшего выхода антенн или для когерентного объединения всех выходов следует реализовать специальные методы обработки сигналов. В настоящее время также реализованы системы с множественными передатчиками, размещенными в разных местах, например система GPS (Global Positioning System — глобальная система навигации и определения положения).
- *Поляризационное разнесение* (polarization diversity) [45] — это еще один из способов получения дополнительных некоррелированных наборов сигнала.
- Любую схему разнесения можно рассматривать как тривиальную форму кода с повторениями (repetition code) в пространстве и во времени. В то же время существуют методы улучшения отношения SNR в каналах с замиранием, которые эффективнее и мощнее кодов с повторениями. Уникальный метод борьбы с ухудшением — это кодирование с коррекцией ошибок, поскольку он не обеспечивает большую энергию сигнала, а снижает требуемое E_b/N_0, необходимое для достижения желаемой вероятности ошибки. Применение кодирования с коррекцией ошибок совместно с чередованием [19, 46–51] — это, пожалуй, наиболее распространенная схема улучшения рабочих характеристик системы в среде с замиранием. Следует отметить, что механизм рассеивания ошибок во время замирания посредством разнесения во времени зависит от движения переносного устройства. Чем больше скорость мобильного устройства, тем эффективнее эта схема; при низких скоростях эффективность мала. (Зависимость скорости передвижного устройства от характеристик устройства чередования продемонстрирована в разделе 15.5.6.)

15.5.4. Методы разнесения

Задачей реализации методов разнесения является использование дополнительных независимых (или, по крайней мере, некоррелирующих) путей прохождения сигнала для улучшения получаемого SNR. Разнесение может улучшить рабочие характеристики системы при сравнительно небольших затратах; в отличие от выравнивания, разнесение не требует служебных расходов на настройку. В этом разделе будет показано улучшение достоверности передачи, которое можно получить с помощью методов разнесения. Вероятность битовой ошибки $\overline{P}_B$, усредненная по всем "подъемам и спадам" канала с медленным замиранием, можно вычислить следующим образом.

$$\overline{P}_B = \int_0^\infty P_B(x)p(x)dx \tag{15.38}$$

Здесь $P_B(x)$ — вероятность битовой ошибки для данной схемы модуляции при заданном значении SNR $= x$, где $x = \alpha^2 E_b/N_0$, а $p(x)$ — плотность вероятности x при замирании. При постоянных E_b и N_0, α используется для обозначения изменений амплитуды вследствие замирания (см. раздел 15.2.2).

При релеевском замирании α имеет релеевское распределение, так что α^2 и x имеют χ^2-распределение. Таким образом, согласно уравнению (15.15),

$$p(x) = \frac{1}{\Gamma}\exp\left(-\frac{x}{\Gamma}\right) \qquad x \geq 0, \tag{15.39}$$

где $\Gamma = \overline{\alpha^2} E_b/N_0$ — это SNR, усредненное по всем подъемам и спадам замирания. Если каждая разнесенная ветвь (сигнала) имеет мгновенное значение SNR $= \gamma_i$ и предполагается, что каждая ветвь имеет одинаковое среднее значение SNR, равное Γ, то получаем следующее.

$$p(\gamma_i) = \frac{1}{\Gamma}\exp\left(-\frac{\gamma_i}{\Gamma}\right) \qquad \gamma_i \geq 0 \tag{15.40}$$

Вероятность того, что отдельная ветвь имеет SNR, меньшее порогового значения γ, равна следующему.

$$\begin{aligned} P(\gamma_i \leq \gamma) = \int_0^{\gamma} p(\gamma_i)\, d\gamma_i = \int_0^{\gamma} \frac{1}{\Gamma}\exp\left(-\frac{\gamma_i}{\Gamma}\right) d\gamma_i = \\ = 1 - \exp\left(-\frac{\gamma}{\Gamma}\right) \end{aligned} \tag{15.41}$$

Вероятность того, что все M независимых разнесенных ветвей сигнала получены одновременно с SNR, меньшим некоторого порогового значения γ, равна следующему.

$$P(\gamma_1, \ldots, \gamma_M \leq \gamma) = \left[1 - \exp\left(-\frac{\gamma}{\Gamma}\right)\right]^M \tag{15.42}$$

Вероятность того, что любая ветвь сигнала имеет значения SNR $> \gamma$, равна следующему.

$$P(\gamma_i > \gamma) = 1 - \left[1 - \exp\left(-\frac{\gamma}{\Gamma}\right)\right]^M \tag{15.43}$$

Выражение (15.43) — это вероятность превышения порогового значения при разнесении с автовыбором.

Пример 15.1. Преимущество разнесения

Пусть используется разнесение на 4 ветви, и каждая ветвь получает независимый сигнал с релеевским замиранием. Среднее SNR равно $\Gamma = 20$ дБ. Определите вероятность одновременного приема всех 4 ветвей с SNR, меньшим 10 дБ (а также вероятность того, что этот порог будет превышен). Сравните результаты с использованием разнесения и без него.

Решение

Используя уравнение (15.42) при $\gamma = 10$ дБ и $\gamma/\Gamma = 10$ дБ $- 20$дБ $= -10$ дБ $= 0{,}1$, найдем вероятность того, что SNR упадет ниже 10 дБ.

$$P(\gamma_1, \gamma_2, \gamma_3, \gamma_4 \leq 10 \text{ дБ}) = [1 - \exp(-0{,}1)]^4 = 8{,}2 \times 10^{-5}$$

При использовании разнесения получаем следующее.

$$P(\gamma_i > 10 \text{ дБ}) = 1 - 8{,}2 \times 10^{-5} = 0{,}9999$$

Без разнесения

$$P(\gamma_1 \leq 10 \text{ дБ}) = [1 - \exp(-0{,}1)]^1 = 0{,}095$$

$$P(\gamma_1 > 10 \text{ дБ}) = 1 - 0{,}095 = 0{,}905$$

15.5.4.1. Комбинированные методы разнесения

Наиболее распространенные методы объединения разнесенных сигналов — это *разнесение с автовыбором* (selection diversity), *разнесение с обратной связью* (feedback diversity), *разнесение с максимальным отношением* (maximal ratio diversity) и *разнесение с равным усилением* (equal gain diversity). В системах, использующих пространственное разнесение, выбор включает дискретизацию M сигналов антенн и передачу на демодулятор наибольшего из них. При разнесении с автовыбором объединение сигналов реализуется относительно просто, однако оно не является оптимальным, поскольку в нем не используются одновременно все полученные сигналы.

При разнесении с обратной связью или при сканирующем разнесении (scanning diversity) не используется самый мощный из M сигналов; вместо этого M сигналов сканируются в определенной последовательности до тех пор, пока не будет найден сигнал, превышающий данное пороговое значение. Именно этот сигнал используется до тех пор, пока его уровень не опустится ниже установленного порогового значения, после чего процесс сканирования начинается снова. Достоверность этого метода несколько ниже, чем других методов, однако разнесение с обратной связью довольно просто реализовать.

При объединении разнесенных сигналов по принципу максимального отношения сигналы со всех M ветвей взвешиваются согласно их личным отношениям SNR, а затем суммируются. Перед суммированием требуется достичь синфазности суммируемых сигналов. Алгоритмы определения требуемого опережения или задержки сигнала аналогичны используемым в эквалайзерах и RAKE-приемниках. Суммирование с максимальным отношением дает среднее SNR $\overline{\gamma_M}$, равное сумме отдельных средних SNR, как показано ниже.

$$\overline{\gamma_M} = \sum_{i=1}^{M} \overline{\gamma_i} = \sum_{i=1}^{M} \Gamma = M\Gamma \tag{15.44}$$

Здесь предполагалось, что каждая ветвь имеет среднее SNR, равное $\overline{\gamma_i} = \Gamma$. Таким образом, объединение сигналов с максимальным отношением может дать приемлемое среднее SNR, даже если ни одно из средних значений $\overline{\gamma_i}$ не является приемлемым. В этом методе M ветвей суммируются синфазно, т.е. они умножаются на соответствующий весовой коэффициент так, чтобы на приемник подавался сигнал с наибольшим возможным SNR.

Объединение с равным усилением аналогично объединению с максимальным отношением, за исключением того, что все весовые коэффициенты равны единице. По-прежнему остается возможность достичь приемлемого значения SNR на выходе при большом числе неприемлемых значений на входе. Достоверность передачи при этом незначительно уступает достоверности при объединении с максимальным отношением (см. [52] для более детального ознакомления с объединением разнесенных сигналов).

15.5.5. Типы модуляции для каналов с замираниями

Очевидно, что схема передачи сигнала, основанная на преобразованиях амплитуды, такая как амплитудная манипуляция (amplitude shift keying — ASK) или квадратурная амплитудная модуляция (quadrature amplitude modulation — QAM), по сути, подвержена ухудшению качества передачи в среде с замиранием. Таким образом, для каналов с замираниями предпочтительно выбирать схемы передачи сигнала с частотным или фазовым типом модуляции.

При рассмотрении ортогональных схем модуляции FSK для каналов с замираниями удобно использовать схему MFSK (с $M = 8$ или больше), поскольку ее достоверность выше, чем у схемы с передачей двоичного сигнала. В каналах с медленным релеевским замиранием двоичная DPSK и 8-FSK отличаются не более чем на 0,1 дБ друг от друга [19]. На первый взгляд, может показаться, что при повышении порядка ортогонального алфавита расширяется полоса пропускания, которая в какой-то момент превысит полосу когерентности, что приведет к частотно-селективному замиранию. Однако для схемы MFSK требуется доступная полоса передачи, намного превышающая ширину полосы распространяющегося сигнала. Например, рассмотрим схему 8-FSK и скорость передачи 10 000 символов/с. Ширина полосы пропускания равна $MR_s = 80\,000$ Гц. Это ширина полосы частот, которая должна быть доступна для использования системой. Однако каждый раз при передаче символа отправляется не весь алфавит, а только один однополосный тон (занимающий в спектре 10 000 Гц). При рассмотрении модуляции PSK для каналов с замираниями алфавиты модуляции более высокого порядка показывают плохую производительность, поэтому схем MPSK с $M = 8$ или выше следует избегать [19]. Ниже в качестве некоторого обоснования такой точки зрения приводится пример 15.2, в котором рассмотрена система мобильной связи.

Пример 15.2. Изменения в системе мобильной связи

Доплеровское расширение $f_d = V/\lambda$ показывает, что скорость замирания непосредственно зависит от скорости движения. В табл. 15.2 показано доплеровское расширение в зависимости от скорости движения передвижного устройства для несущих частот 900 МГц и 1800 МГц. Вычислите изменение фазы, приходящееся на один символ, для передачи сигнала с модуляцией QPSK при скорости $24{,}3 \times 10^3$ символов/с. Предполагается, что несущая частота равна 1800 МГц, а скорость передвижного устройства равна 50 миль/ч (80 км/ч). Повторите вычисления для скорости передвижного устройства, равной 100 миль/ч.

Решение

$$\begin{aligned}\Delta\theta/\text{символ} &= \frac{f_d \text{ Гц}}{R_s \text{ символ/с}} \times 360^\circ = \\ &= \frac{132 \text{ Гц}}{24{,}3 \times 10^3 \text{ символ/с}} \times 360^\circ = \\ &= 2^\circ/\text{символ}\end{aligned}$$

При скорости 100 миль/ч: $\Delta\theta$/символ = 4°/символ

Таким образом, должно быть очевидно, почему MPSK со значением $M > 4$ обычно не используется для передачи информации в среде с многолучевым распространением.

Таблица 15.2. Доплеровское расширение в зависимости от скорости мобильного устройства

Скорость		Доплеровское расширение (Гц)	Доплеровское расширение (Гц)
миль/ч	км/ч	900 МГц ($\lambda = 33$ см)	1800 МГц ($\lambda = 16{,}6$ см)
3	5	4	8
20	32	27	54
50	60	66	132
80	108	106	212
120	192	160	320

15.5.6. Роль чередования

В разделе 8.2 были описаны различные свойства чередования. Для передачи в среде с многолучевым распространением основным преимуществом чередования является осуществление временного разнесения (при использовании совместно с кодированием с коррекцией ошибок). Чем больше интервал времени, в течение которого канальные символы разделены, тем больше шансов, что смежные биты (после восстановления исходного порядка) будут подвержены некоррелированным проявлениям замирания, таким образом, больше шансов достичь эффективного разнесения. На рис. 15.19 показаны преимущества введения интервала времени чередования $T_{\text{Ш}}$, большего времени когерентности канала T_0. Система имеет следующие параметры: модуляция DBPSK, декодирование согласно мягкой схеме принятия решений, сверточный код со степенью кодирования 1/2, $K = 7$, канал испытывает медленное релеевское замирание. Должно быть очевидно, что устройство чередования, имеющее наибольшее отношение $T_{\text{Ш}}/T_0$, будет работать лучше всего (высокая частота появления ошибок при демодуляции ведет к низкой частоте появления ошибок декодирования). Это позволяет заключить, что $T_{\text{Ш}}/T_0$ должно быть каким-нибудь большим числом, скажем 1000 или 10 000. В то же время в системах связи реального времени это невозможно, поскольку характерная временная задержка, связанная с чередованием, была бы чрезмерной. Как описывалось в разделе 8.2.1 для блочного чередования, перед передачей первой строки и первого столбца в память должен быть загружен практически весь массив. Подобным образом в приемнике перед операцией восстановления массива почти весь он должен быть сохранен. Это приведет к задержке, равной длительности одного блока данных, как в передатчике, так и приемнике. В примере 15.2 показано, что для сотовой системы телефонной связи с несущей частотой 900 МГц отношение $T_{\text{Ш}}/T_0$, равное 10, приблизительно составляет предел, при котором еще не наблюдается чрезмерной задержки.

Интересно отметить, что чередование не дает никаких преимуществ в отношении многолучевого распространения при отсутствии относительного движения передатчика и приемника (или движения объектов на путях распространения сигналов). Преимущества (касающиеся достоверности передачи в системе) обнаруживаются при увеличении скорости движения. (Не нужно использовать это в качестве оправдания превышения скорости на шоссе.)

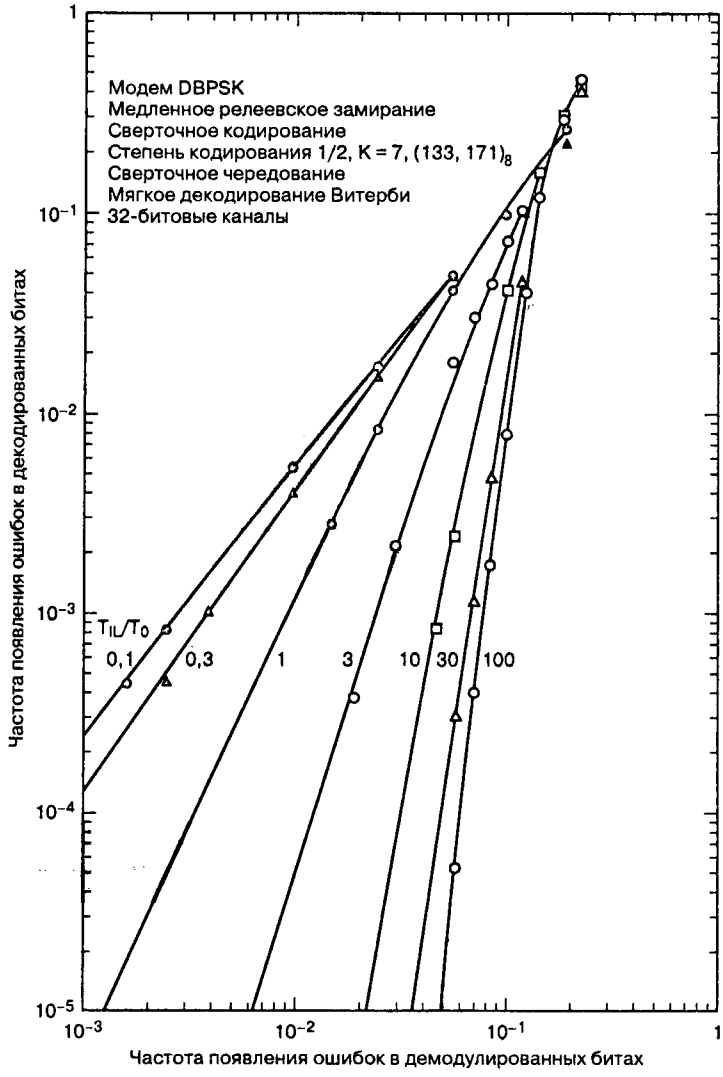
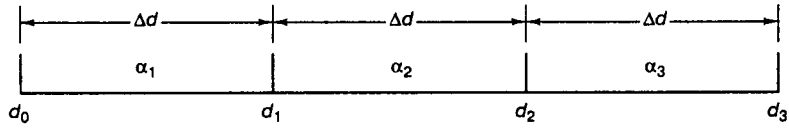


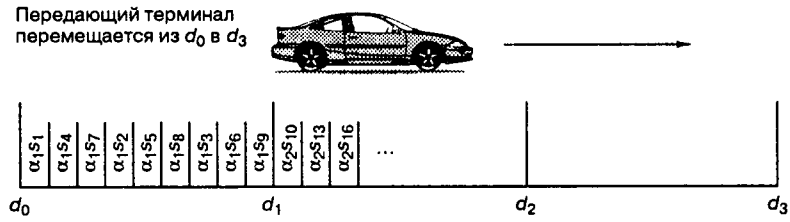
Рис. 15.19. Вероятность ошибки для различных отношений времени чередования к времени когерентности

На рис. 15.20, *а* показаны области, характеризуемые разными функциями замирания $\{\alpha_i\}$. В области между точками d_0 и d_1 функция замирания равна α_1, между точками d_1 и d_2 — α_2 и т.д. Пусть точки d_i расположены через равное расстояние Δd. На рис. 15.20, *б* показан автомобиль, движущийся с небольшой скоростью; когда он перемещается на расстояние Δd, его передатчик успевает излучить девять символов. Допустим, что рабочий интервал устройства чередования — это три символа, так что символы s_1—s_9 появляются в произвольном порядке, показанном на рис. 15.20, *б*. Отметим, что все девять символов испытывают одинаковое замирание α_1, так что после восстановления исходного сигнала мы не обнаружим никакого преимущества чередования. Рассмотрим теперь рис. 15.20, *в*, на котором автомобиль движется в 3 раза бы-

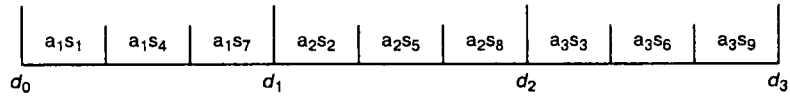
стрее, чем на рис. 15.20, *б*; таким образом, когда он переместится на расстояние Δd, его передатчик излучит только три символа. Как и ранее, символы подвержены замиранию, характерному для этой области. В результате этого получаем последовательность из девяти символов, показанную на рис. 15.20, *в*. После восстановления исходной последовательности из последовательности, показанной на рис. 15.20, *в*, получаем следующие пары "множитель замирания/символ": $\alpha_1 s_1$, $\alpha_2 s_2$, $\alpha_3 s_3$, $\alpha_1 s_4$, $\alpha_2 s_5$, $\alpha_3 s_6$, $\alpha_1 s_7$, $\alpha_2 s_8$, $\alpha_3 s_9$. Можно видеть, что смежные символы искажаются вследствие влияния различных множителей замирания. Таким образом, чередование с временным периодом, слишком малым, чтобы давать хотя бы какие-нибудь преимущества при низких скоростях, оправдывает себя при более высоких скоростях.



а) Зависимость поглощения a_i от расстояния



б) Передача симолов s_i после чередования (нескоростная машина)



в) Передача символов s_i после чередования (скоростная машина)

Рис. 15.20. Преимущества чередования при увеличении скорости радиостанции

На рис. 15.21 также показано, что хотя с увеличением скорости мобильного устройства качество связи и ухудшается (увеличивается скорость замирания), польза от чередования при этом возрастает. На рис. 15.21 представлены результаты эксплуатационных испытаний, проведенных на системе CDMA, удовлетворяющей стандарту Interim Specification 95 (IS-95), в канале, состоящем из движущегося устройства и базовой станции [53]. На рисунке показана зависимость отношения E_b/N_0, требуемого для поддержания частоты ошибок в кадрах (20 мс данных), равной 1%, от скорости передвижного устройства. Наилучшие характеристики (наименьшее требуемое E_b/N_0) достигаются при низких скоростях от 0 до 20 км/ч. Это область низких скоростей, в которой методы регулирования мощности в системе могут наиболее эффективно компенсировать эффекты медленного замирания; при низких скоростях чередование не приносит какой-либо пользы, и на графике показано сильное ухудшение характеристик как функция скорости. При скорости порядка 20–60 км/ч крутизна этого ухудшения уменьшается. Это область, в которой регулирование мощности в системе уже не позволяет полностью справиться с возрастанием скорости замирания, и в то же

время использование чередования еще не приносит достаточной пользы. На скорости 60 км/ч достоверность передачи для такой системы достигает наихудшего значения. Когда устройство движется более 60 км/ч, контроль мощности уже не позволяет каклибо бороться с замиранием, однако чередование обеспечивает неизменное улучшение характеристик при увеличении скорости. Задача устройства чередования, заключающаяся в преобразовании эффектов глубокого замирания (коррелирующие во времени события) в случайные события, упрощается с ростом скорости. Итак, достоверность передачи по каналу с замираниями обычно ухудшается с ростом скорости, поскольку возрастает доплеровское расширение или скорость замирания. В то же время использование чередования, которое становится более эффективно при высоких скоростях, приводит к ослаблению эффектов ухудшения. Эта тенденция повышения достоверности передачи не может продолжаться бесконечно. В конечном счете производительность системы достигает уровня неустранимых ошибок, показанного на рис. 15.15. Следовательно, если бы измерения, показанные на рис. 15.21, проводились при скоростях, превышающих 200 км/ч, то на графике была бы точка, в которой кривая развернулась бы круто вверх, что соответствовало бы ухудшению рабочих характеристик, вызванному возрастанием доплеровского эффекта.

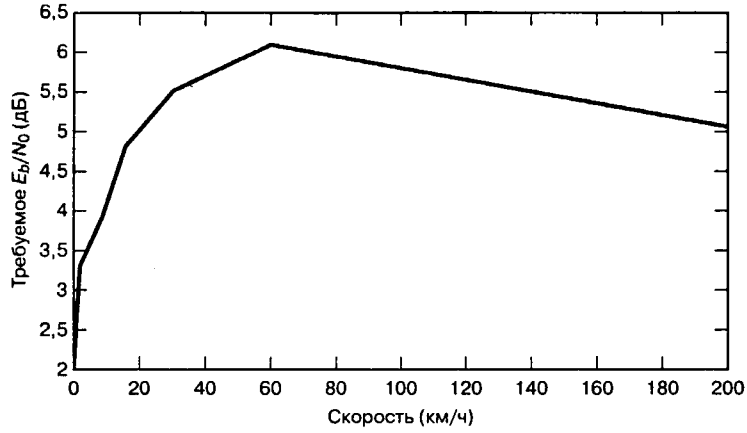


Рис. 15.21. Типичная зависимость требуемого E_b/N_0 от скорости движения. Используется релеевский канал с двумя независимыми путями распространения, частота передачи 850 МГц, частота появления ошибочных кадров 1%

15.6. Краткий обзор ключевых параметров, характеризующих каналы с замираниями

Подытожим условия, которым должна удовлетворять система, чтобы канал не вносил частотно-селективного искажения и искажения, вызванного быстрым замиранием. Объединив выражения (15.22), (15.32) и (15.33), получаем следующее.

$$f_0 > W > f_d \tag{15.45}$$

или

$$T_m < T_s < T_0 \tag{15.46}$$

Иными словами, желательно, чтобы полоса когерентности канала превышала скорость передачи сигналов, которая, в свою очередь, должна превышать скорость замирания в канале. Напомним, что если не бороться с искажениями, то f_0 устанавливает верхний предел, а f_d — нижний предел скорости передачи сигнала.

15.6.1. Искажения вследствие быстрого замирания: случай 1

Если условия (15.45) и (15.46) не выполняются, искажения будут происходить до тех пор, пока не будут приняты подходящие меры. Рассмотрим быстрое замирание, при котором скорость передачи сигналов меньше скорости замирания в канале.

$$f_0 > W < f_d \tag{15.47}$$

Борьба с искажениями заключается в использовании одного или нескольких перечисленных ниже методов (см. рис. 15.18).

- Выбирается метод модуляции/демодуляции, наиболее устойчивый в условиях быстрого замирания. Это значит, например, что необходимо избегать схем, которые требуют контуров ФАПЧ для восстановления несущей, поскольку быстрое замирание может не позволить контурам ФАПЧ достичь синхронизации.
- Вводится достаточная избыточность, чтобы скорость передачи символов превышала скорость замирания в канале, но в то же время не превышала ширины полосы когерентности. Тогда канал можно классифицировать как проявляющий амплитудное замирание. Однако, как было показано в разделе 15.3.3, даже каналы с амплитудным замиранием будут испытывать частотно-селективное замирание всегда, когда передаточная функция проявляет спектральный нуль вблизи центра полосы сигнала. Поскольку это происходит только иногда, бороться с искажением можно путем выбора адекватного кода коррекции ошибок и использования чередования.
- Описанные выше два способа борьбы с искажением могут привести к тому, что демодулятор будет работать возле релеевского предела [19] (см. рис. 15.17). В то же время график зависимости вероятности ошибки от E_b/N_0 может спрямляться (как это показано на рис. 15.15) вследствие частотно-модулированного шума, который является результатом случайного доплеровского расширения. Использование внутриполосного контрольного тона и контура стабилизации частоты может снизить уровень, при котором характеристика спрямляется.
- Чтобы избежать эффекта дна ошибки вследствие случайного доплеровского расширения, скорость передачи сигналов должна увеличиться до величины, превышающей скорость замирания приблизительно в 100–200 раз [27]. Это один из мотивов разработки мобильных систем связи, работающих в режиме множественного доступа с временным разделением (time-division multiple access — TDMA).
- Применяется кодирование с коррекцией ошибок и чередование для дополнительного улучшения рабочих характеристик системы.

15.6.2. Искажения вследствие частотно-селективного замирания: случай 2

Рассмотрим частотно-селективное замирание, при котором ширина полосы когерентности меньше скорости передачи символов, в то время как скорость передачи символов больше доплеровского расширения.

$$f_0 < W > f_d \tag{15.48}$$

Поскольку скорость передачи символов превышает скорость замирания в канале, искажения вследствие быстрого замирания отсутствуют. В то же время необходимо ослабить частотно-селективные эффекты. Борьба с искажениями заключается в использовании одного или нескольких перечисленных ниже методов (см. рис. 15.18).

- Адаптивное выравнивание, расширение спектра (методом прямой последовательности или скачкообразной перестройки частоты), OFDM, контрольный сигнал. В европейской системе GSM в каждый временной интервал передачи выводится некоторая контрольная последовательность, помогающая приемнику определить импульсную характеристику канала. Для ослабления частотно-селективных искажений применяется эквалайзер Витерби (рассматривается ниже).
- Когда воздействие искажений ослаблено, для приближения к характеристикам канала AWGN можно использовать методы частотного разнесения (а также кодирование с коррекцией ошибок и чередование). Для передачи спектра, расширенного методом прямой последовательности (direct-sequence spread-spectrum — DS/SS), разнесение может реализоваться посредством использования RAKE-приемника (рассматривается ниже), выполняющего когерентное объединение многолучевых компонентов, которые в противном случае были бы утеряны.

15.6.3. Искажения вследствие быстрого и частотно-селективного замирания: случай 3

Пусть ширина полосы когерентности канала меньше скорости передачи сигналов, которая, в свою очередь, меньше скорости замирания. Это условие математически выражается следующим образом.

$$f_0 < W < f_d \tag{15.49}$$

или

$$f_0 < f_d \tag{15.50}$$

Очевидно, что канал проявляет как быстрое, так и частотно-селективное замирание. Напомним из уравнений (15.45) и (15.46), что f_0 устанавливает верхний предел, а f_d — нижний предел скорости передачи сигналов. Таким образом, условие (15.50) представляет собой сложную проектную задачу, поскольку, если не обеспечено подавление искажений, *максимально* допустимая скорость передачи сигнала будет, собственно говоря, *меньше минимально* допустимой скорости передачи сигналов. Борьба с искажением в этом случае выполняется подобно тому, как это рекомендовалось в случае 1.

- Выбирается метод модуляции/демодуляции, наиболее устойчивый в условиях быстрого замирания.
- Для увеличения скорости передачи символов используется избыточность передачи.
- Вводятся какие-либо типы подавления искажений, вызванных частотно-селективным замиранием, подобно описанным в случае 2.

- Когда воздействие искажений было подавлено, вводится какой-либо тип разнесения (а также кодирование с коррекцией ошибок и чередование) с целью приближения к характеристикам канала AWGN.

Пример 15.3. Эквалайзеры и устройства чередования в мобильной связи

Рассмотрим сотовый телефон, который размещен на объекте, движущемся со скоростью 60 миль в час (96 км/ч). Несущая частота равна 900 МГц. С помощью тестового профиля эквалайзера GSM, показанного на рис. 15.22, определите следующее: а) среднеквадратический разброс задержек σ_τ; б) максимально допустимую ширину полосы сигнала $W \approx 1/T_s$, при которой не требуется эквалайзер; в) считая, что разброс задержек в канале равен найденному в п. а, какая из следующих систем требует использования эквалайзера: цифровой сотовый стандарт США (United States Digital Cellular Standard — USDC), известный как IS-54 (новая версия — IS-136), глобальная система мобильной связи (Global System for Mobile — GSM), системы CDMA, разработанные согласно IS-95; ширина полос и скорость передачи символов для этих систем равны следующему: USDC — $W = 30$ кГц, $1/T_s = 24{,}3 \times 10^3$ символа/с; GSM — $W = 200$ кГц, $1/T_s = 271 \times 10^3$ символа/с; IS-95 — $W = 1{,}25$ МГц, $1/T_s = 9{,}6 \times 10^3$ символа/с; г) общую (передатчик плюс приемник) задержку, вносимую устройством чередования, когда отношение рабочего интервала устройства к времени когерентности $T_{\text{Ц}}/T_0$ равно 10 (если общая приемлемая задержка (передатчик плюс приемник) для речи равна 100 мс, можно ли использовать устройство с описанными выше характеристиками для передачи речи?); д) повторите пп. а–г для несущей частоты 1900 МГц.

Рис. 15.22. Тестовый профиль эквалайзера GSM

Решение

а) На рис. 15.22 тестовый профиль системы GSM показывает идеализированный компонент многолучевого распространения, расположенный через каждые шесть интервалов задержки $\{\tau_k\}$ в промежутке от 0 до 16 мкс. Каждый компонент можно обозначить через $S(\tau_k)$, его среднюю относительную мощность, которая на этом профиле одинакова для всех компонентов (0 дБ). Профиль представляет *мнимую* многолучевую среду, используемую при тестировании перед выравниванием [15]. При таком расположении компонентов, как показано на рисунке, средний разброс задержек будет иметь следующий вид.

$$\bar{\tau} = \frac{\sum_k S(\tau_k)\tau_k}{\sum_k S(\tau_k)} = \frac{0 + 3{,}2 + 6{,}4 + 9{,}6 + 12{,}8 + 16{,}0}{6} = 8 \text{ мкс}$$

Второй момент разброса задержек $\overline{\tau^2}$ и среднеквадратический разброс задержек σ_τ имеют следующий вид:

$$\overline{\tau^2} = \frac{\sum_k S(\tau_k)\tau_k^2}{\sum_k S(\tau_k)} = \frac{0+3{,}2^2+6{,}4^2+9{,}6^2+12{,}8^2+16{,}0^2}{6} = 93{,}87 \text{ мкс}^2$$

и, с помощью уравнения (15.17),

$$\sigma_\tau = \sqrt{\overline{\tau^2} - (\overline{\tau})^2} = \sqrt{93{,}87 - 8^2} = 5{,}5 \text{ мкс}$$

б) С помощью уравнения (15.21) полоса когерентности канала будет определена следующим образом.

$$f_0 \approx \frac{1}{5\sigma_\tau} = \frac{1}{5 \times 5{,}5 \text{ мкс}} = 36{,}4 \text{ кГц}$$

Таким образом, максимально допустимая полоса пропускания сигнала, при которой не нужно использовать эквалайзер, будет $W = 36{,}4$ кГц.

в) Для полос пропускания различных систем, данных в этом примере, очевидно, что использование эквалайзера в USCD не обязательно, тогда как в GSM он действительно нужен. Относительно систем, которые разрабатывались согласно IS-95, можно сказать следующее: поскольку скорость передачи сигналов или полоса пропускания W, равная 1,25 МГц, значительно превышает полосу когерентности 36,4 кГц, система проявляет частотно-селективное замирание. В то же время в таких системах с расширением спектра методом прямой последовательности (direct-sequence spread spectrum — DS/SS), W умышленно расширяется с целью превышения f_0 и, следовательно, подавления эффектов частотно-селективного замирания. Необходимость в эквалайзере возникает только тогда, когда проблему представляет межсимвольная интерференция (intersymbol interference — ISI), но ISI не является проблемой, если скорость передачи символов меньше полосы когерентности (или длительность символа больше многолучевого разброса). Следовательно, в случае IS-95 эквалайзер не нужен, поскольку скорость передачи $9{,}6 \times 10^3$ символов/с значительно ниже полосы когерентности. Для разнесения путей применяется описываемый в разделе 15.7.2 RAKE -приемник; на уровне элементарных сигналов его реализация сходна с реализацией эквалайзера.

г) Чтобы определить задержку, вносимую устройством чередования, рассчитаем доплеровское расширение и время когерентности с помощью уравнений (15.25) и (15.29).

$$f_d = \frac{V}{\lambda} = \frac{\frac{96 \text{км/ч}}{3600 \text{с/ч}}}{\frac{3 \times 10^8 \text{ м/с}}{9 \times 10^8 \text{ Гц}}} = 80 \text{Гц}, \text{ следовательно, } T_0 \approx \frac{0{,}5}{f_d} = 6{,}3 \text{мс}$$

Исходя из того, что $T_{IL}/T_0 = 10$, рабочий интервал устройства чередования равен $T_{IL} = 63$ мс. Из этого следует, что общая задержка передатчика и приемника равна 126 мс. Для передачи речи это значение несколько превышает приемлемое. В мобильных системах часто применяются устройства с более короткими рабочими интервалами, которые дают односторонние задержки порядка 20–40 мс.

д) Повторяем расчеты для несущей частоты 1900 МГц. На вычисление полосы когерентности смена несущей не оказывает никакого влияния, а вот доплеровское расширение, время когерентности и задержку чередования нужно рассчитывать заново. Итак,

$$f_d = \frac{V}{\lambda} = 169\,\text{Гц}, \text{ следовательно, } T_0 \approx \frac{0{,}5}{f_d} = 3\,\text{мс}$$

Таким образом, рабочий интервал устройства чередования равен $T_{IL} = 30$ мс; это даст общую задержку передатчика и приемника, равную 60 мс, что является приемлемым значением для речевого сигнала.

15.7. Приложения: борьба с эффектами частотно-селективного замирания

15.7.1. Применение эквалайзера Витерби в системе GSM

На рис. 15.23 показан кадр (длительность 4,615 мс) схемы множественного доступа с временным разделением (time-division multiple access — TDMA) в системе GSM, состоящий из 8 слотов, каждый из которых присвоен активному мобильному клиенту. Обычный пакет передачи, занимающий один интервал, состоит из 57 бит сообщения, расположенных по обе стороны от 26-битовой последовательности, иногда называемой *зондирующей* (sounding) или *настроечной* (training). Длительность одного слота составляет 0,577 мс (или скорость передачи равна 1733 слота/с). Задача внутренней контрольной последовательности — помочь приемнику в адаптивном определении импульсной характеристики канала (за время передачи одного слота, т.е. 0,577 мс). Чтобы данный метод был эффективным, характеристики замирания в канале должны оставаться неизменными в течение времени, приблизительно равного длительности одного слота. Иначе говоря, за время передачи одного слота, пока приемник анализирует искажение контрольного блока, не должно проявиться быстрое замирание; в противном случае компенсация замирания в канале окажется неэффективной. В качестве примера можно взять приемник GSM, находящийся на скоростном поезде, который движется с постоянной скоростью 200 км/ч (около 55,56 м/с). Частота несущей 900 МГц (длина волны $\lambda = 0{,}33$ м). Из уравнения (15.29) время, соответствующее проходу половины длины волны, равно следующему.

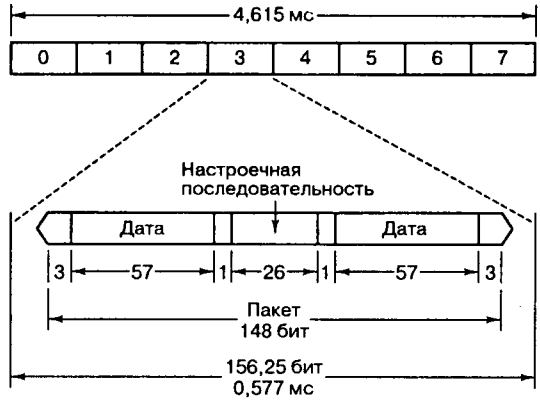


Рис. 15.23. Кадр TDMA GSM и временной слот, содержащий нормальный пакет

$$T_0 \approx \frac{\lambda/2}{V} \approx 3 \text{ мс} \tag{15.51}$$

Как показывает уравнение (15.51), это приблизительно отвечает времени когерентности. Следовательно, время когерентности канала более чем в 5 раз превышает время передачи одного слота (0,577 мс). Время, необходимое для значительного изменения характеристик замирания в канале, относительно велико по сравнению со временем передачи одного слота. Отметим, что выбор, сделанный в системе GSM при подборе времени передачи слота TDMA и контрольного блока, несомненно, был осуществлен при учете необходимости устранения эффектов быстрого замирания, которые могут свести на нет эффективность работы эквалайзера. Скорость передачи символов в стандарте GSM (или скорость передачи битов, если используется двоичная модуляция) равна 271 000 символов/с, а полоса пропускания W составляет 200 кГц. Поскольку среднеквадратический разброс задержек σ_τ в городской местности равен порядка 2 мкс, то, исходя из уравнения (15.21), можно видеть, что результирующая полоса когерентности f_0 будет приблизительно равна 100 кГц. Следовательно, очевидно, что поскольку $f_0 < W$, приемник GSM должен иметь средства для борьбы с частотно-селективным искажением. Как правило, для этого используется эквалайзер Витерби.

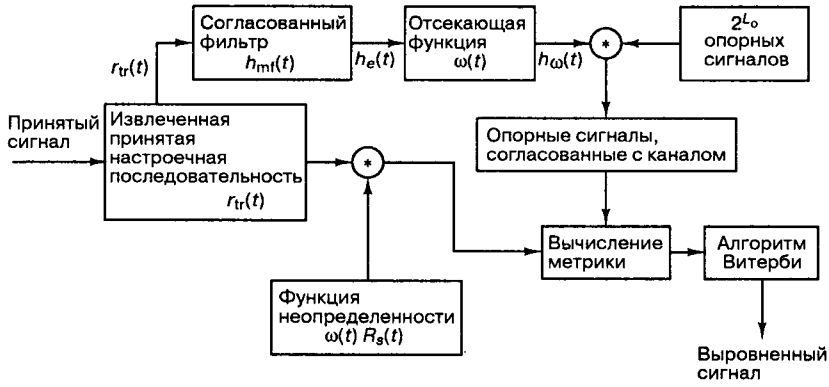


Рис. 15.24. Применение эквалайзера Витерби в системе GSM

На рис. 15.24 изображена блок-схема приемника GSM для оценки импульсной характеристики канала. Эта оценка нужна детектору для получения опорных сигналов, согласованных с состоянием канала [54], что будет объяснено ниже. Наконец, для оценки битов сообщения с максимальным правдоподобием используется алгоритм Витерби. Принятый сигнал можно описать через переданный сигнал, свернутый с импульсной характеристикой канала. Обозначим через $s_{tr}(t)$ переданную контрольную последовательность, а через $r_{tr}(t)$ — соответствующую принятую последовательность.

$$r_{tr}(t) = s_{tr}(t) * h_c(t) \tag{15.52}$$

В данном случае "*" означает операцию свертки, причем шумом мы пренебрегаем. В приемнике, поскольку $r_{tr}(t)$ является частью принятого нормального пакета, она извлекается и отсылается на фильтр с импульсной характеристикой $h_{mf}(t)$, который согласован с $s_{tr}(t)$. Этот согласованный фильтр выдает оценку $h_c(t)$, обозначаемую как $h_e(t)$, которая, согласно (15.25), записывается следующим образом.

$$h_e(t) = r_{tr}(t) * h_{mf}(t) = \\ = s_{tr}(t) * h_c(t) * h_{mf}(t) = \\ = R_s(t) * h_c(t) \tag{15.53}$$

Здесь $R_s(t) = s_{tr}(t) * h_{mf}(t)$ — автокорреляционная функция $s_{tr}(t)$. Если $s_{tr}(t)$ предназначена для получения очень короткой (импульсного типа) автокорреляционной функции $R_s(t)$, тогда $h_e(t) \approx h_c(t)$. Далее, при использовании отсекающей функции $w(t)$, $h_e(t)$ усекается до функции $h_w(t)$, которую уже можно обрабатывать численно. Временная длительность $w(t)$, обозначаемая как L_o, должна быть достаточно большой для компенсации эффектов типичной ISI, введенной каналом. Член L_o образуется в результате двух вкладов, а именно: L_{CISI}, соответствующий управляемой ISI, вызванной гауссовой фильтрацией полосового сигнала (который затем модулирует несущую согласно схеме MSK), и L_C, соответствующий вводимой каналом ISI, которая вызвана многолучевым распространением. Таким образом, L_o можно записать следующим образом.

$$L_o = L_{CISI} + L_C$$

В системе GSM требуется обеспечить подавление искажений, вызванных дисперсией сигнала, имеющего разброс задержек порядка 15–20 мкс. Поскольку в GSM длительность бита составляет 3,69 мкс, L_o можно выразить в единицах битовых интервалов. Следовательно, эквалайзер Витерби, применяемый в системе GSM, обладает памятью от 4 до 6 битовых интервалов. На каждом интервале L_o бит задача эквалайзера Витерби состоит в нахождении наиболее правдоподобной последовательности, длиной L_o бит, среди 2^{L_o} возможных, которые могли быть переданы. Определение наиболее правдоподобной L_o-битовой последовательности, которая могла быть передана, требует создания 2^{L_o} значащих опорных сигналов путем модификации (или искажения) 2^{L_o} идеальных сигналов (генерируемых приемником) таким образом, как канал искажает передаваемый слот. Следовательно, 2^{L_o} опорных сигналов сворачиваются с усеченной оценкой импульсной характеристики канала $h_w(t)$ с целью генерации искаженных или своего рода подогнанных под канал опорных сигналов. Затем подкорректированные сигналы сравниваются с принятыми информационными сигналами для расчета метрик. Отметим, что перед сравниванием принятые данные сворачиваются с известной усеченной автокорреляционной функцией $w(t)R_s(t)$, преобразовывая ее подобно опорным сигналам. Такой фильтрованный сигнал сообщения сравнивается с 2^{L_o} возможными подкорректированными опорными сигналами, причем способ получения метрик подобен способу, использованному в алгоритме декодирования Витерби (Viterbi decoding algorithm — VDA). Алгоритм VDA дает максимально правдоподобную оценку переданной последовательности данных [34].

Отметим, что в большинстве методов выравнивания для компенсации неоптимальных свойств $h_c(t)$ применяются фильтры, т.е. выравнивающие фильтры пытаются модифицировать искаженные формы импульсов. В то же время эквалайзер Витерби работает иным образом. Он включает измерение $h_c(t)$, а затем предоставляет способ подгонки приемника под среду канала. Целью такой подгонки является попытка помочь детектору в оценке искаженной последовательности импульсов. При наличии эквалайзера Витерби искаженные выборки не меняют формы и не компенсируются прямо каким-либо иным методом; приемник не подавляет сигнал, он перестраивается таким образом, что становится способен к более эффективной обработке искаженных фрагментов.

15.7.2. RAKE-приемник в системах с расширением спектра методом прямой последовательности

Стандарт IS-95 определяет систему сотовой связи DS/SS, в которой для разнесения путей распространения используется RAKE-приемник (RAKE receiver) [35–37]. Данный приемник изучает различные многолучевые задержки на предмет кодовой корреляции, потом соответствующим образом восстанавливает задержанные сигналы, которые затем оптимально сочетаются с выходом других независимых коррелляторов. На рис. 15.25 показаны профили мощности сигнала, соответствующие пяти передачам элементарных сигналов кодовой последовательности 1 0 1 1 1, причем моменты наблюдения обозначены как t_{-4} — для самого раннего наблюдения и t_0 — для самого позднего. На осях абсцисс показаны три компонента, поступающих с задержками τ_1, τ_2 и τ_3. Полагается, что интервалы между моментами передачи t_i и интервалы между моментами задержек τ_i равны по длительности одному элементарному сигналу. Отсюда можно сделать вывод, что компонент, поступающий на приемник в момент t_{-4} с задержкой τ_3, совпадает по времени с двумя другими компонентами, а именно: поступающими в моменты t_{-3} и t_{-2} с задержками τ_2 и τ_1, соответственно. Поскольку в этом примере задержанные компоненты разделены, по крайней мере, временем одного элементарного сигнала, то их можно разрешить. В приемнике должен быть блок зондирования, предназначенный для оценки времени задержки τ_i. Следует отметить, что для мобильных наземных систем радиосвязи скорость замирания относительно низка (порядка миллисекунд) или, иначе говоря, когерентность канала довольно высока по сравнению с длительностью элементарного сигнала ($T_0 > T_{ch}$). Таким образом, изменения τ_i проявляются достаточно слабо, чтобы приемник успел подстроиться к ним.

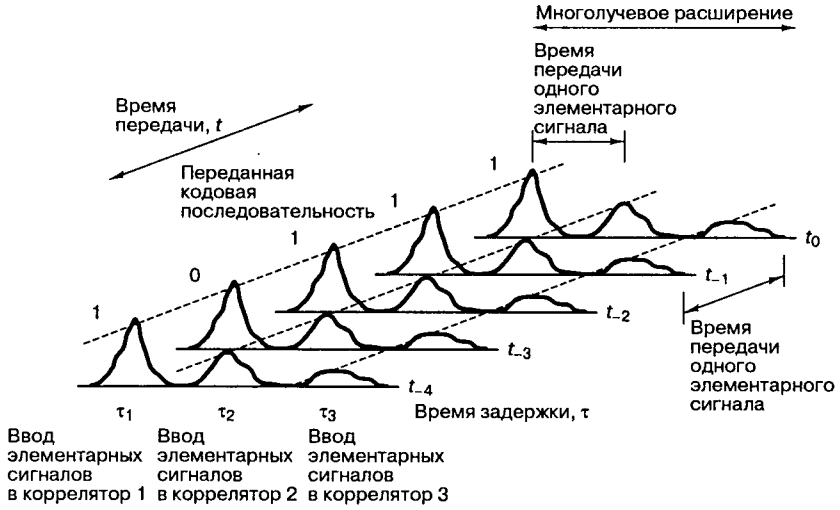


Рис. 15.25. Принимаемые элементарные сигналы в трехкомпонентном RAKE-приемнике

После оценки задержек τ_i для восстановления каждого разрешимого многолучевого компонента используется отдельный коррелятор. В данном примере подразумевается три таких коррелятора, каждый из которых будет обрабатывать запаздывающую версию

одной и той же последовательности элементарных сигналов 1 0 1 1 1. На рис. 15.25 каждый коррелятор принимает элементарные сигналы с профилем мощности, представляющим собой последовательность компонентов, расположенную вдоль диагональной линии. Для простоты все элементарные сигналы показаны как положительные сигнальные посылки. В действительности эти элементарные сигналы образуют шумоподобную последовательность, которая, конечно, содержит и положительные, и отрицательные импульсы. Каждый из корреляторов пытается скоррелировать эти поступающие элементарные сигналы с таким же соответствующим образом синхронизированным псевдослучайным кодом. В конце символьного интервала (как правило, на один символ приходится сотни или даже тысячи элементарных сигналов) выходы корреляторов когерентно объединяются, после чего принимается решение относительно значения принятого символа. На рис. 15.26 показано фазовое вращение компонентов (F_i), выполняемое RAKE-приемником для облегчения когерентного объединения сигналов. На уровне элементарных сигналов RAKE-приемник подобен эквалайзеру, но его действительная функция заключается в разнесении путей распространения.

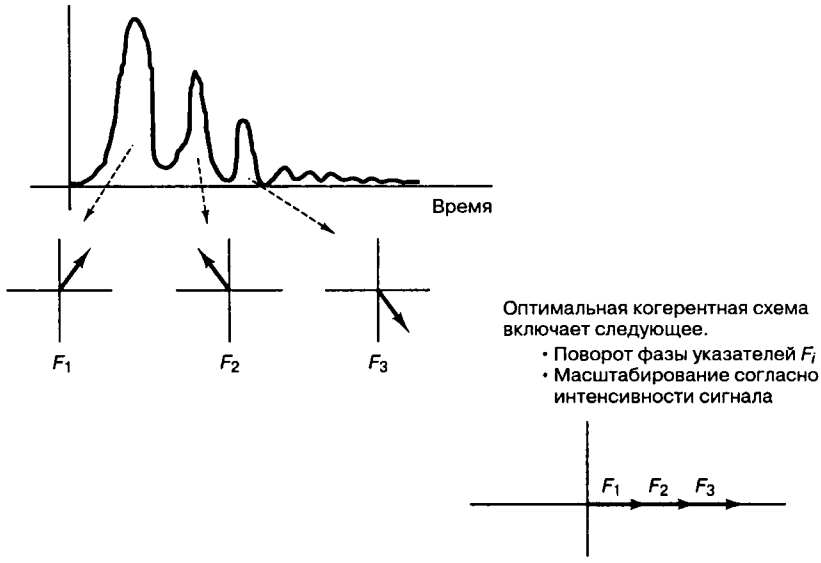


Рис. 15.26. Когерентное объединение многолучевых вкладов в RAKE-приемнике

Способность систем DS/SS к подавлению помех основывается на том, что кодовая последовательность, поступающая на приемник со сдвигом по времени лишь на один элементарный сигнал, будет иметь очень низкую корреляцию с конкретным псевдослучайным кодом, с которым коррелировала исходная последовательность. Следовательно, любые кодовые элементарные сигналы, запаздывающие на один или более элементарных интервалов, будут подавляться коррелятором. Задержанные элементарные сигналы всего лишь вносят вклад в возрастание уровня интерференции (корреляционных боковых лепестков). Подавление, которое осуществляет RAKE-приемник, можно назвать *разнесением путей распространения*, так как он осуществляет объединение энергии элементарных сигналов, которые поступают по многим путям распространения. Без RAKE-приемника эта энергия была бы потеряна для приемника DS/SS. Если на рис. 15.25 обратить внимание на картину над точкой τ_3, можно сделать вывод, что существует интерференция между

элементарными сигналами вследствие одновременного поступления разных компонентов. Эффективность, получаемая в результате расширения спектра, позволяет системе выдерживать такую интерференцию на уровне элементарных сигналов. Считается, что другой коррекции в стандарте IS-95 не нужно.

15.8. Резюме

В этой главе охарактеризованы основные эффекты, вносящие вклад в замирание в определенных каналах связи. Здесь представлен рис. 15.1, который является путеводной нитью при рассмотрении явлений замирания. Описаны два типа замирания, крупно- и мелкомасштабное. Изучены два проявления мелкомасштабного замирания (дисперсия сигнала и скорость замирания). Рассмотрение проводилось с двух точек зрения — частотной и временной. В главе определены две категории ухудшения качества для дисперсии: частотно-селективное и амплитудное замирание. Кроме того, две категории определены для скорости замирания: быстрое и медленное замирание. Категории ухудшения вследствие мелкомасштабного замирания представлены на рис. 15.7. На рис. 15.8 показаны математические модели, в которых используются корреляционные функции и функции плотности мощности. Эти модели позволяют получить удобное симметричное описание, благодаря которому можно наглядно представить преобразование Фурье и соотношение дуальности, описывающие явления замирания. Здесь также представлены методы борьбы с эффектами каждой из категорий замирания; эти методы показаны на рис. 15.18. В заключение показано применение методов подавления в системах GSM и CDMA, удовлетворяющих стандарту IS-95.

Литература

1. Rappaport T. S. *Wireless Communications.* Chapter 3 and 4, Prentice Hall, Upper Saddle River, New Jersey, 1996.
2. Greenwood D. and Hanzo L. *Characterization of Mobile Radio Channels.* Mobile Radio Communications, edited by R. Steele, Chapter 2, Pentech Press, London, 1994.
3. Lee W. C. Y. *Elements of Cellular Mobile Radio Systems.* IEEE Trans. on Vehicular Technology, vol. V-35, n. 2, May, 1986, pp. 48–56.
4. Okumura Y., et. al. *Field Strength and its Variability in VHF and UHF land Mobil Radio Service.* Review of the Elec. Comm. Lab., vol. 16, n. 9 & 10, 1968, pp. 825–873.
5. Hata M. *Empirical Formulae for Propagation Loss in Land Mobile Radio Services.* IEEE Trans. on Vehicular Technology, vol. VT-29, n. 3, 1980, pp. 317–325.
6. Seidel S. Y. et. al. *Path Loss, Scattering and Multipath Delay Statistics in Four European Cities for Digital Cellular and Microcellular Radiotelephone.* IEEE Transactions on Vehicular Technogy, vol. 40, n. 4, November, 1991, pp. 721–730.
7. Cox D. C., Murray R. and Norris, A. *800 MHz Attenuation Measured in and around Suburban Houses.* AT&T Bell Laboratory Technical Journal, vol. 673, n. 6, July–August, 1984, pp. 921–954.
8. Schilling D. L., et. al. *Broadband CDMA for Personal Communications Systems.* IEEE Communications Magazine, vol. 29, n. 11, November 1991, pp. 86–93.
9. Andersen J. B., Rappaport T. S., Yoshida S. *Propagation Measurements and Models for Wireless Communications Channels.* IEEE Communications Magazine, vol. 33, n. 1, January, 1995, pp. 42–49.
10. Proakis J. G. *Digital Communications,* Chapter 7. McGraw-Hill Book Company, New York, 1983.
11. Schwartz M. *Information, Transmission, Modulation, and Noise,* Second Edition. McGraw-Hill, New York, 1970.
12. Amoroso F. *Investigation of Signal Variance, Bit Error Rates and Pulse Dispersion for DSPN Signaling in a Mobil Dense Scatterer Ray Tracing Model.* Int'l Journal of Satellite Communications, vol. 12, 1994, pp. 579–588.

13. Bello P. A. *Characterization of Randomly Time-Variant Linear Channels.* IEEE Trans. on Commun. Syst., December, 1963, pp. 360–393
14. Green P. E. Jr. *Radar Astronomy Measurement Techniques.* MIT Licoln Laboratory, Lexington, Mass., Tech Report No. 282, December, 1962.
15. Pahlavan K. and Levesque A. H. *Wireless Information Nerworks.* Chapters 3 and 4. John Wiley and Sons, New York, 1995.
16. Lee W. Y. C. *Mobil Cellular Communications.* McGraw-Hill Book Co., New York, 1989.
17. Amoroso F. *Use of DS/SS Signaling to Mitigate Rayleigh Fading in a Dens Scatterer Environment.* IEEE Personal Communications, vol. 3, n. 2, April, 1996, pp. 52–61.
18. Clarke R. H. *A Statistical Theory of Mobile radio Reception.* Bell System Technical J., vol. 47, n. 6, July–August, 1968, pp. 957–1000.
19. Bogusch, R. L. *Digital Communications in Fading Channels: Modulation and Coding.* Mission Research Corp., Santa Barbara, California, Report no. MRC-R-1043, March, 11, 1987.
20. Amoroso F. *The Bandwidth of Digital Data Signals.* IEEE Communications Magazine, vol. 18, n. 6, November, 1980, pp. 13–24.
21. Bogusch R. L. et. al. *Frequency Selective Propagation Effects on Spread-Spectrum Receiver Tracking.* Prceedings of the IEEE, vol. 69, n. 7, July, 1981, pp. 787–796.
22. Jakes W. C. (Ed.) *Microwave Mobile Communications.* John Wiley & Sons, New York, 1974.
23. *Joint Tchnical Committee of Committee T1 R1P1.4 and TIA TR46.33/TR45.4.4 on Wireless Access.* "Draft Final Report on RF Channel Characterization," Paper No. JTC(AIR)/94.01.17-238R4, January, 17, 1994.
24. Bello, P. A. and Nelin, B. D., "The Influence of Fading Spectrum on the Binary Error Probabilities of Incoherent and Differentially Coherent Matched Filter Receivers," *IRE Transactions on Commun. Syst.,* vol. CS-10, June, 1962, pp. 160–68.
25. Amoroso F. *Instantaneous Frequently Effects in a Doppler Scattering Environment.* IEEE Enternational Conference on Communications, June, 7–10, 1987, pp. 1458–1466.
26. Fung V., Fappaport T. S. and Thoma B. *Bit-Error Simulation for $\pi/4$ DQPSK Mobile Radio Communication Using Two-Ray and Measurement-Base Impulse Response Models.* IEEE J. Sel. Areas Commun., vol. 11, n. 3, April, 1993, pp. 393–394.
27. Bateman A. J. and McGeehan J. P. *Data Transmission over UHF Fading Mobile Radio Channels.* IEEE Proceedings, vol. 131, Pt. F, n. 4, July, 1984, pp. 364–374.
28. Feher K. *Wireless Digital Communications,* Prentice Hall, Upper Saddle River, New Jersey, 1995.
29. Davarian F., Simon M. and Sumida J. *DMSK: A Practical 2400-bps Receiver for the Mobile Satellite Service.* Jet Propulsion Laboratory Publication 85–51 (MSAT-X Report No. 111), June, 15, 1985.
30. Rappaport T. S. *Wireless Communicatios.* Chapter 6. Prentice Hall, Upper Saddle River, New Jersey, 1996.
31. Bogousch R. L., Guigliano F. W. and Knepp D. L. *Frequency-Selective Scintillation Effects and Decision Feedback Equalization in High Data-Rate Satellite Links.* Proceedings of the IEEE, vol 71, n. 6, June, 1983, pp. 754–767.
32. Qureshi S. U. H. *Adaptive Equalization.* Proceedings of the IEEE, vol. 73, n. 9, September, 1985, pp. 1340–87.
33. Forney G. D. *The Viterbi Algorithm.* Proceedings of the IEEE, vol. 61, n. 3, March, 1978, pp. 268–278.
34. Viterbi A. J. and Omura J. K. *Principles of Digital Communication and Coding.* McGraw-Hill, New York, 1979.
35. Price R. and Green P. E. Jr. *A Communication Technique for Multipath Channels.* Proceeding of the IRE, March, 1958, pp. 555–570.
36. Turin G. L. *Introduction to Spread-Spectrum Amtimultipath Techniques and their Application to Urban Digital Radio.* Proceedings of the IEEE, vol. 68, n. 3, March, 1980, pp. 328–353.
37. Simon M. K., Omura J. K., Scholtz R. A. and Levitt B. K. *Spread Spectrum Communications Handbook.* McGraw-Hill Book Co., 1994.
38. Birchler M. A. and Jasper S. C. *A 64 kbps Digital Land Mobile Radio System Employing M-16QAM.* Proceedings of the 1992 IEEE Int'l. Conference on Selected Topics in Wireless Communications, Vancouver, British Columbia, June, 25–26, 1992, pp. 158–162.

39. Sari H., Karam G. and Jeanclaude I. *Transmission Techniques for Digital Terrestrial TV Broadcasting.* IEEE Communications Magazine, vol. 33, n. 2, February, 1995, pp. 100–109.
40. Cavers J. K. *The Performance of Phase Locked Transparent Tone-in-Band with Symmetric Phase Detection.* IEEE Trans. on Commun., vol. 39, n. 9, September, 1991, pp. 1389–1399.
41. Moher M. L. and Lodge J. H. *TCMP—A Modulation and Coding Strategy for Rician Fading Channel.* IEEE Journal on Selected Areas in Communications, vol. 7, n. 9, December, 1989, pp. 1347–1355.
42. Harris F. *On the Relationship Between Multirate Polyphase FIR Filters and Windowed, Overlapped FFT Processing.* Proceedings of the Tweinty Third Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California, October, 30 to November, 1, 1989, pp. 485–488.
43. Lowdermilk R. W. and Harris F. *Design and Performance of Fading Insensitive Orthogonal Frequency Division Multiplexing (OFDM) using Polyphase Filtering Techniques.* Proceedings of the Thirtieth Annual Asilomar Comference on Signals, Systems, and Computers, Pacific Grove, California, November, 3–6, 1996.
44. Kavehrad M. and Bodeep G. E. *Design and Experimental Results for a Direct Sequence Spread-Spectrum Radio Using Differential Phase-Shift Keying Modulation for Indoor Wireless Communications.* IEEE JSAC, vol. SAC-5, n. 5, June, 1987, pp. 815–23.
45. Hess G. C. *Land-Mobil Radio System Engineering.* Artech House, Boston, 1993.
46. Hagenauer J. and Lutz E. *Forward Error Correction Coding for Fading Compensation in Mobile Satellite Channels.* IEEE JSAC, vol. SAC-5, n. 2, February, 1987, pp. 215–225.
47. McLane P. I., et al. *PSK and DPSK Trellis Codes for Fast Fading, Shadowed Mobile Satellite Communication Channels.* IEEE Trans. on Comm., vol. 36, n. 11, November, 1988, pp. 1242–1246.
48. Schlegel C. and Costello D. J. Jr. *Bandwidth Efficient Coding for Fading Channels: Code Construction and Performance Analysis.* IEEE JSAC, vol. 7, n. 9, December, 1989, pp. 1356–1368.
49. Edbauer F. *Performance of Interleaved Trellis-Coded Differential 8-PSK Modulation over Fading Channals.* IEEE J. on Selected Areas in Comm., vol. 7, n.9, December, 1989, pp. 1340–1346.
50. Soliman S. and Mokrani K. *Performance of Coded Systems over Fading Dispersive Channels.* IEEE Trans. on Communications, vol. 40, n. 1, January, 1992, pp. 51–59.
51. Divsalar D. and Pollara, F. *Turbo Codes for PCS Applications.* Proc. ICC'95, Seattle, Washington, June, 18–22, 1995, pp. 54–59.
52. Simon M. and Alouini M-S. *Digital Communications over Fading Channels: A Unified Approach to Performance Analysis.* John Wiley, New York, 2000.
53. Padovani R. *Reverse Link Performance of IS-95 Based Cellular Systems.* IEEE Personal Communications, Third Quarter 1994, pp. 28–34.
54. Hanzo L. and Stefanov J. *The Pan-European Digital Cellular Mobile Radio System—Know as GSM.* Mobile Radio Communications, edited by R. Steele, Chapter 8, Pentech Press, London, 1992.

Задачи

15.1. Функция плотности вероятности для непрерывной случайной релеевской переменной определяется формулой (15.15).

а) Найдите выражение для *функции распределения*, как это описано в разделе 1.5.5.

б) Используйте функцию распределения для определения процента времени, в течение которого уровень сигнала будет на 15 дБ ниже среднеквадратического значения для сигнала, переданного по каналу мобильной радиосвязи, испытывающему релеевское замирание.

в) Повторите п. б для уровня сигнала, который на 5 дБ ниже среднеквадратического.

15.2. Сигнал в системе мобильной радиосвязи расширяется во времени. Скорость передачи символов $R_s = 20 \times 10^3$ символов/с. Измерения в канале показывают, что средняя избыточная задержка распространения равна 10 мкс, а второй момент избыточной задержки равен $1{,}8 \times 10^{-10}$ с^2.

а) Вычислите ширину полосы когерентности f_0, если она определена как интервал частот, в пределах которого комплексная передаточная функция имеет корреляцию не меньше 0,9.

б) Повторите п. а, если f_0 определена как интервал, имеющий корреляцию не меньше 0,5.

в) Определите, будет ли сигнал подвергаться частотно-селективному замиранию.

15.3. Рассмотрим канал, профиль плотности мощности которого состоит из трех импульсных функций со следующей мощностью и следующим расположением временной задержки: –20 дБ при 0 мкс, 0 дБ при 2 мкс и –10 дБ при 3 мкс.

а) Вычислите среднюю избыточную задержку.

б) Вычислите второй момент избыточной задержки.

в) Вычислите среднеквадратический разброс задержек.

г) Оцените ширину полосы когерентности (соответствующую корреляции не менее 0,9).

д) Вычислите приблизительное значение частоты передачи, если приемник расположен на самолете, движущемся со скоростью 800 км/ч, а время, требуемое для пересечения расстояния, равного половины длины волны, равно 100 мкс.

15.4. Дана система мобильной радиосвязи с несущей частотой $f_c = 900$ МГц и доплеровской частотой $f_d = 50$ Гц. Предполагается, что применяется модель плотного размещения рассеивающих элементов.

а) Изобразите график *доплеровской плотности спектральной мощности* $S(\nu)$ в интервале $f_c \pm f_d$ (используйте порядка 10 точек).

б) Объясните поведение $S(\nu)$ на границах.

в) Вычислите время когерентности T_0, предполагая, что отклик канала на синусоиду дает корреляцию не менее 0,5.

15.5. Для каждой из перечисленных ниже категорий замирания назовите приложение, обычно подпадающее под эту категорию. Дайте количественное обоснование.

а) Частотно-селективное, быстрое замирание.

б) Частотно-селективное, медленное замирание.

в) Амплитудное замирание, быстрое замирание.

г) Амплитудное замирание, медленное замирание.

15.6. а) Как связаны профиль плотности мощности сигнала, характеризующийся среднеквадратической задержкой σ_τ, и доплеровская спектральная плотность мощности, характеризующаяся шириной полосы замирания f_d?

б) Как связаны частотная корреляционная функция, которая характеризуется шириной полосы когерентности f_0, и временная корреляционная функция, которая характеризуется временем когерентности T_0?

15.7. Рассмотрим узкополосные системы мобильной связи для применения внутри помещений, которые характеризуются профилем плотности мощности, состоящим из четырех импульсных функций со следующей мощностью и следующим расположением временной задержки: 0 дБ при 0 нс, –3 дБ при 100 нс, –3 дБ при 200 нс и –6 дБ при 300 нс. Какую максимальную скорость передачи символов может поддерживать такая система без использования эквалайзера? Для нахождения ширины полосы когерентности воспользуйтесь определением, в котором фигурирует корреляция тонов 0,5.

15.8. Рассмотрим систему мобильной радиосвязи, использующую модуляцию QPSK при скорости передачи $24{,}3 \times 10^3$ символов/с и несущей частоте 1900 МГц. Какова наибольшая допустимая скорость транспортных средств, использующих такую систему, если требуется, чтобы изменения фазы в результате спектрального расширения (доплеровского расширения) не превышали 5°/символ?

15.9. Чтобы чередование обеспечивало значимое разнесение во времени, эмпирическое правило требует, чтобы рабочий интервал соответствующего устройства T_{IL} был, по крайней мере, в десять раз больше времени когерентности канала T_0. Покажите график зависимости T_{IL} от частоты (отобразите по три значения частоты: 300 МГц, 3 ГГц и 30 ГГц) для следующих пользователей мобильных телефонов.

а) Пешеход, идущий со скоростью 1 м/с.

б) Скоростной поезд, движущийся со скоростью 50 м/с

в) Если телефон используется для общения в реальном времени, то какая из шести точек на графике описывает случай, когда можно достичь значимого разнесения во времени при использовании рабочего интервала устройства чередования, ровно в десять раз превышающего T_0?

г) Какие общие выводы можно сделать?

15.10. Ширина полосы передаваемого сигнала равна 5 кГц, сигнал распространяется по каналу с полосой когерентности 50 кГц. Очевидно, что это один из примеров каналов с амплитудным замиранием. Объясните, как такой канал может время от времени подвергаться частотно-селективному замиранию.

15.11. Рассмотрим систему мобильной радиосвязи TDMA с несущей частотой 1900 МГц, которая работает на поездах при скоростях 180 км/ч. Для изучения импульсной характеристики канала с целью обеспечения выравнивания в передачу каждого пользователя в дополнение к информационным битам вносятся настроечные биты. Необходимо, чтобы настроечная последовательность состояла из 20 бит, при этом данное число не должно превышать 20% от общего количества бит, также настроечные биты должны внедряться в данные, по крайней мере, каждые $T_0/4$ с. Предполагая двоичную модуляцию, определите наименьшую скорость передачи, при которой эти требования удовлетворялись бы без быстрого замирания.

15.12. а) В конце 80-х в Японии была разработана система PHS (Portable Handyphone System — персональная система переносных телефонов). Спецификация PHS задает разнесение несущих, равное 300 кГц. Восприимчив ли этот стандарт к частотно-селективному замиранию в среде, в которой канал обладает среднеквадратическим разбросом задержек порядка 300 нс?

б) Стандарт телефонов DECT (Digital Enhanced Cordless Telephone —цифровые расширенные беспроводные телекоммуникации) был разработан для информационного обмена высокой плотности и ближней связи (внутри помещений). Спецификация DECT задает разнесение несущих, равное 1,728 МГц. Предполагается, что среднеквадратический разброс задержек равен 150 нс. Определите, нужно ли включать в схему приемника DECT эквалайзер.

15.13. Рабочий интервал устройства чередования должен, по крайней мере, в 10 раз превышать время когерентности канала, чтобы дать существенное разнесение по времени в мобильной системе радиосвязи. Рассмотрите использование такого устройства при проектировании системы мобильной связи, работающей на частоте 1 ГГц и предназначенной для пешеходов, идущих со скоростью 0,5 м/с. Насколько *большим* должен быть интервал? Подходит ли это для системы речевой связи реального времени?

15.14. Какое максимальное отношение рабочего интервала устройства чередования ко времени когерентности T_{IL}/T_0 можно использовать в следующих случаях, если суммарный интервал задержки передатчика и приемника необходимо удержать ниже 100 мс.

а) Скорость замирания в канале равна 100 Гц.

б) Скорость замирания в канале равна 1000 Гц.

15.15. Системы мобильной связи сконструированы так, чтобы поддерживать скорость передачи данных, равную 200 Кбит/с, используя при этом модуляцию QPSK и несущую частоту

1900 МГц. Они предназначены для использования в транспортных средствах, которые обычно движутся со скоростью 96 км/ч.

а) Какое изменение фазового угла $\Delta\theta$ на символ можно ожидать?

б) Чему будет равно $\Delta\theta$ на символ, если скорость передачи уменьшится до 100 Кбит/с?

в) Повторите п. б для скорости 48 км/ч.

г) Сделайте общие выводы для данного случая.

15.16. Среднеквадратический разброс задержек в канале, испытывающем замирание вследствие многолучевого распространения, равен $\sigma_\tau = 10$ мкс, а доплеровское расширение равно $f_d = 1$ Гц. Длительность широкополосного импульса принимается равной $T_s = 1$ мкс.

а) Чему равна ширина полосы когерентности канала?

б) Чему равно время когерентности канала?

в) Как можно было бы классифицировать канал относительно частотной избирательности и скорости замирания.

г) Как можно было бы изменить длительность импульса (скорость передачи данных), чтобы ослабить эффекты замирания?

15.17. В мобильных системах радиосвязи схема, основанная на фазовой модуляции, чрезвычайно подвержена фазовым искажениям. Этих искажений можно избежать, если скорость передачи сигнала превышает скорость замирания, по меньшей мере, в 100 раз [27]. Рассмотрим радиосистему, работающую на несущей частоте 1900 МГц и движущуюся со скоростью 96 км/ч. Какой должна быть наименьшая скорость передачи символа в такой системе, чтобы избежать искажений вследствие быстрого замирания?

15.18. Рассмотрим систему мобильной связи, обладающую кадровой структурой и распределением временных слотов (рис. 315.1).

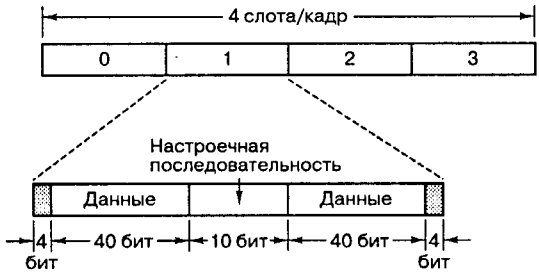


Рис. 315.1. Формат кадра TDMA

На каждый кадр приходится 4 временных слота; каждому пользователю отводится один слот на кадр. Каждый слот содержит 98 бит, как это показано на рис. 315.1. При передаче сигнала используется модуляция QPSK с несущей частотой 700 МГц. Скорость передачи равна $33{,}6 \times 10^3$ символов/с, а ширина полосы пропускания равна 47 кГц. Система должна нормально функционировать при скоростях до 100 км/ч. Измерения, проводимые в физическом канале, показали, что типичное среднеквадратическое значение разброса задержек составляет порядка 4 мкс.

а) Будут ли в канале эффекты ухудшения характеристик вследствие быстрого замирания, если предположить, что настроечная последовательность позволяет оценить импульсную характеристику канала в течение каждого сегмента времени?

б) Будет ли такая конструкция подвергаться ухудшению характеристик вследствие частотно-селективного замирания?

15.19. Общая допустимая задержка переданных данных в отдельном канале мобильной радиосвязи ограничена величиной 340 мс. Скорость передачи данных равна $19{,}2 \times 10^3$ символов/с,

данные при этом чередуются с целью разнесения во времени. Задержки, характерные для системы, показаны в табл. 315.1.

Таблица 315.1. Значение задержек в мс

Задержка, T	Значение (в мс)
Кодер	2
Модулятор	10
Канал	0,3
Демодулятор	25
Декодер	$2 \times 10^8/f_{clk}$

Задержка в миллисекундах для декодера дана в виде $2 \times 10^8/f_{clk}$, где f_{clk} — тактовая частота декодера. Вычислите минимальную тактовую частоту декодера, требуемую при следующих рабочих диапазонах устройства чередования.

а) 100 бит

б) 1000 бит

в) 2850 бит

г) Какие можно сделать выводы относительно поведения тактовой частоты декодера в результате увеличения размера рабочего интервала устройства?

15.20. Рассмотрим систему мобильной связи с ортогональной FDM (OFDM), которая предназначена для работы в транспортных средствах (со скоростью 80 км/ч в городской среде) и обладает шириной полосы когерентности 100 кГц. Несущая частота равна 3 ГГц, при этом требуется, чтобы данные передавались при скорости 1024×10^3 символов/с. Выберите подходящую схему поднесущих для следующих целей: 1) избежать использования эквалайзера и 2) минимизировать любые эффекты, вызванные быстрым замиранием. Схема должна определять, сколько необходимо поднесущих, насколько далеко они должны быть разнесены по частоте и какое должно использоваться значение отношения скорости передачи символов на поднесущую.

15.21. Системы мобильной радиосвязи используют передачу сигналов со спектром, расширенным методом прямой последовательности (direct-sequence spread-spectrum — DS/SS), для ослабления следствий того, что полученный сигнал имеет два компонента: прошедший по прямому пути и пришедший после отражения. Отраженный путь на 120 м длиннее прямого. Какой должна быть скорость передачи элементарного сигнала, чтобы такая система ослабляла эффект многолучевого распространения?

15.22. Общеизвестно, что передача сигналов со спектром, расширенным методом прямой последовательности (direct-sequence spread-spectrum — DS/SS), может использоваться как метод борьбы с вызванной каналом ISI в частотно-селективных каналах. Тем не менее, если рассмотреть рис. 15.25 в определенный момент времени, скажем τ_3, то будет присутствовать интерференция между элементарными сигналами. Нужно ли использовать дополнительные методы выравнивания, чтобы преодолеть интерференцию на уровне элементарных сигналов? Объясните.

15.23. Схемы CDMA и TDMA уникальны в том смысле, что каждая из этих схем множественного доступа имеет свои средства борьбы с замиранием. От каких типов ухудшения характеристик "естественным образом" защищает каждая схема?

15.24. Рассмотрим схему разнесения, состоящую из четырех каналов, как показано на рис. 315.2. Каждый канал отвечает за прохождение сигналов $r(t)$, независимо замирающих по Релею. В определенный момент времени полученный сигнал может быть выражен в виде четырехмерного вектора $\mathbf{r} = [r_1, r_2, r_3, r_4]$, где r_i — напряжение в канале i. Кроме того, усиление в каждом из каналов можно выразить через четырехмерный вектор $\mathbf{G} = [G_1, G_2, G_3, G_4]$, где G_i описывает усиление напряжения в канале i. Рассмотрим мо-

мент времени, в который измеренное значение **r** было равно [0,87, 1,21, 0,66, 1,90], а соответствующее усиление G — [0,5, 0,8, 1,0, 0,8]. Средняя мощность шума в каждом канале N равна 0,25.

а) Вычислите SNR сигнала, поступающего на детектор.

б) Можно показать [1], что SNR максимально, когда все G_i равны r_i^2/N. Используя этот факт, определите максимально достижимое SNR.

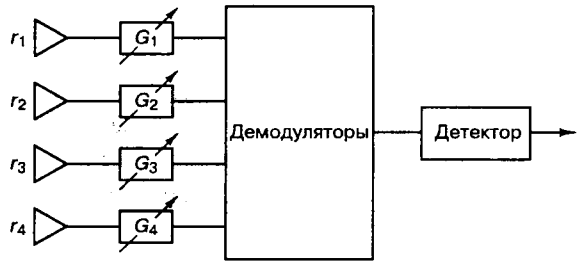


Рис. 315.2. Приемник с разнесением на четыре канала

15.25. В системе для улучшения значения SNR приемника используется разнесение каналов. Предполагается, что каждый канал получает независимо замирающий релеевский сигнал. Приемник должен удовлетворять следующему требованию: вероятность получения всеми каналами сигнала с SNR, меньшим некоторого порогового значения, равна 10^{-4}, где пороговое значение принято равным 5 дБ, а среднее SNR равно 15 дБ.

а) Вычислите количество каналов разнесенного приема M, необходимых для того, чтобы приемник удовлетворял этому условию.

б) Основываясь на результатах п. а, вычислите вероятность получения во всех каналах SNR > 5 дБ.

15.26. В приемнике с двумя каналами используется схема разнесения. Из каждого канала было получено следующее.

$$\begin{bmatrix}\text{Канал 1}\\ \text{Канал 2}\end{bmatrix} = \begin{bmatrix}1{,}85 & 1{,}91 & -1{,}311 & -1{,}58 & 1{,}21 & 1{,}93 & 1{,}11 & -1{,}67 & 2{,}13 & -2{,}25\\ 1{,}67 & 1{,}69 & -2{,}13 & -1{,}26 & 1{,}74 & 1{,}76 & 1{,}29 & -1{,}93 & 2{,}31 & -1{,}08\end{bmatrix}$$

В первой строке показаны значения напряжений в первом канале, а во второй строке — напряжения во втором канале. Каждый столбец соответствует определенному моменту времени. Считается, что средняя мощность шума в каждом канале равна 0,25 Вт, также предполагается, что упомянутые выше значения преобразованы в синфазные с последующим объединением методами максимального отношения и равного усиления. Мгновенное усиление напряжения, предоставляемое делителем для каналов 1 и 2, равно $G_1 = 1{,}2$ и $G_2 = 1{,}4$. Кроме того, разнесение с обратной связью предполагает, что пороговое значение SNR нужно установить равным 5 дБ.

Вычислите, выход какого канала будет подан на детектор, если используются следующие методы разнесения.

а) Выборочный.

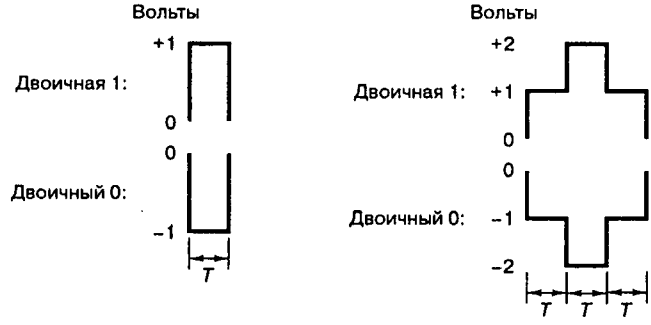
б) С обратной связью.

Вычислите величину SNR, которую имеет сигнал, поданный на детектор, если используются следующие методы разнесения.

а) Максимального отношения.

б) Равного усиления.

15.27. Отклик канала на идеальный положительный или отрицательный импульс расширяется в три раза, как это показано на рис. 315.3. Таким образом, для последовательности переданных импульсов полученный сигнал состоит из суперпозиции $L(=3)$ вкладов (сегменты от трех импульсов) — текущий импульс плюс память о двух предыдущих импульсах. Используйте *диаграмму решетчатого кодирования* для описания вызванной каналом ISI и пометьте каждую ветвь решетки значениями напряжения, являющимися результатом перехода. Изначально система была очищена до состояния 00 путем передачи двух отрицательно поляризованных импульсов. Затем рассмотрите передачу последовательности 1 1 0 1 1 с использованием идеальных импульсов, изображенных на рис. 315.3. Определите амплитуду полученного искаженного сигнала и покажите его путь по решетчатой диаграмме. *Подсказка:* эта двоичная система с конечным числом состояний имеет 2^{L-1} состояний. Воспользуйтесь миллиметровкой для вычисления суперпозиции, необходимой для представления искаженных сигналов, характеризующих канал. Построение решетчатой диаграммы описано в разделе 7.2.3. *Единственное замечание*: здесь вместо кодовых битов используются уровни напряжения.



а) Переданные идеальные биполярные импульсы

б) Принятые импульсы, искаженные ISI

Рис. 315.3

15.28. Используйте характеристики канала и настроечную последовательность, описанную в задаче 15.27, и добавьте шумовое напряжение, равное $\{+1\ -1\ +1\ -1\ +1\}$, для получения искаженного сигнала. Применяйте *диаграмму решетчатого декодирования* для иллюстрации того, как алгоритм декодирования Витерби используется в этом процессе выравнивания, и приведите вычисления, дающие первый бит сообщения. *Подсказка:* процесс подобен декодированию битов, кодированных сверточным кодом, где вместо кодовых битов используются уровни напряжения.

15.29. В мобильных системах связи для борьбы с эффектами замирания используется эквалайзер Витерби. Скорость передачи равна 160×10^3 символов/с, для модуляции используется схема BPSK. Дисперсия сигнала, являющаяся результатом вызванной каналом ISI, равна 25 мкс.

а) Вычислите приблизительный объем памяти L_0 в битовых интервалах, который необходимо включить в эквалайзер Витерби.

б) Каким должен быть объем памяти, чтобы удвоить скорость передачи символов?

Вопросы

15.1. Какие два механизма характеризуют мелкомасштабное замирание? Объясните, как временное и частотное описание этих механизмов связано через *Фурье-преобразование* и отношение дуальности (см. разделы 15.2–15.4).

15.2. Какая разница между *райсовским* и *релеевским* замиранием (см. раздел 15.2.2)?

15.3. Определите следующие параметры: среднеквадратический разброс задержек, ширина полосы когерентности, время когерентности, доплеровское расширение. Как они связаны между собой (см. разделы 15.3 и 15.4)?

15.4. Какие две *категории ухудшения характеристик* характеризуют рассеяние сигнала по времени, а какие две — нестационарную природу канала (см. разделы 15.3 и 15.4.)?

15.5. Почему два основных механизма замирания, характеризующих мелкомасштабное замирание, рассматриваются *независимо* друг от друга (см. раздел 15.4.1.1)?

15.6. Почему *искажение сигнала*, вызванное замиранием, является более серьезным эффектом искажения, чем *уменьшение* SNR (см. раздел 15.5)?

15.7. Какие методы применяются для борьбы с частотно-селективным замиранием? Какие методы используются для борьбы с быстрым замиранием (см. раздел 15.5)?

15.8. Какие существуют способы разнесения сигнала (см. раздел 15.5.3)?

15.9. Если между передатчиком и приемником *отсутствует движение*, какой рабочий интервал устройства чередования нужен для защиты от быстрого замирания (см. раздел 15.5.6)?

ПРИЛОЖЕНИЕ А

Обзор анализа Фурье

А.1. Сигналы, спектры и линейные системы

Электрические сигналы связи — это меняющиеся со временем сигналы напряжения или тока, обычно описываемые во временной области. С другой стороны, подобные сигналы также удобно описывать в частотной области, где описание сигнала называется его *спектром*. Спектральные понятия достаточно важны при анализе и проектировании систем связи; они могут описывать сигнал через его среднюю мощность или энергетическое содержание на различных частотах и показывают, какую часть (полосы) электромагнитного спектра занимает сигнал. Федеральная комиссия по средствам связи США (Federal Communications Commission — FCC) требует, чтобы теле- и радиостанции работали на выделенных им частотах при крайне малых промежутках между полосами, занятыми различными станциями. Например, амплитудно-модулированные радиоканалы разделены полосой 10 кГц, а телевизионные каналы — полосой 6 МГц. Так что наш интерес к спектрам и анализу Фурье объясняется реальными требованиями помещения сигнала в точно заданные границы.

Частотные спектральные характеристики можно приписать как к собственно сигналам, так и электрическим схемам. Если говорится, что конкретный спектр описывает *сигнал*, подразумевается, что один из способов описания сигнала — это задать его амплитуду и фазу как функции частоты. В то же время, когда мы говорим о спектральных параметрах *схемы*, имеем в виду передаточную функцию (или частотную характеристику), связывающую выход схемы с ее входом; другими словами, схема характеризуется тем, какая часть спектра входного сигнала пройдет на выход.

А.2. Применение методов Фурье к анализу линейных систем

Методы Фурье используются для анализа линейных схем или систем: (1) для предсказания реакции (отклика) системы; (2) для определения динамики системы (передаточной функции) и (3) для оценки или интерпретации результатов тестов. Предсказание реакции сис-

темы (1) схематически проиллюстрировано на рис. А.1. Пусть на вход системы подается произвольный периодический сигнал с периодом T_0 секунд. Методы Фурье-анализа, как показано на рисунке, позволяют описать подобный вход как сумму синусоидальных сигналов. Наименьшая (или *собственная*) частота этих сигналов — $1/T$ Гц; остальные частоты кратны данной ($2/T_0$, $3/T_0$, ...) и называются *гармониками*. Важной особенностью линейной системы является *принцип суперпозиции* — реакция на сумму сигналов равна сумме откликов на каждый сигнал. Фактически это свойство используется как определение линейности. Математически система линейна, если для всех a, b, $x_1(t)$ и $x_2(t)$

$y_1(t)$ — реакция системы на $x_1(t)$;

$y_2(t)$ — реакция системы на $x_2(t)$;

и

$ay_1(t) + by_2(t)$ — реакция системы на $ax_1(t) + bx_2(t)$.

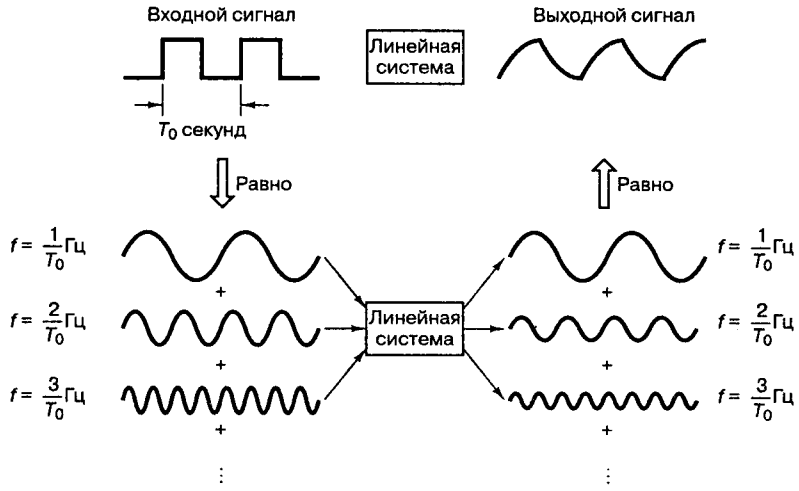


Рис. А.1. Предсказание реакции системы

Данное определение свидетельствует о том, что выходной отклик *линейной* системы с входными синусоидальными сигналами должен составляться из синусоидальных сигналов с *теми же частотами*, что и у входных сигналов; обычно подобная система задается *частотной передаточной функцией* (*частотной характеристикой*), описывающей изменение амплитуды и фазы сигнала в зависимости от частоты, как показано на рис. А.2. На рис. А.2, *а* представлена характерная зависимость амплитуды сигнала от частоты; подобным образом на рис. А.2, *б* показана зависимость фазы сигнала от частоты.

Передаточная функция является рабочей характеристикой системы, т.е. описывает отклик системы на каждую синусоиду. Следовательно, имея передаточную функцию системы, можно предсказать каждый выходной компонент. С помощью принципа суперпозиции эти выходные компоненты суммируются, что дает реакцию системы на любой входной сигнал (рис. А.1). Подобным образом, зная входной и выходной сигналы, можно определить передаточную функцию системы.

Развитие методов Фурье-анализа оказало большое влияние на анализ линейных систем; оно позволило связать переходные процессы и методы работы с гармоническими функциями, а также упростило анализ линейных систем при их активизации

произвольным входным сигналом. Как логарифм позволяет превратить операцию умножения в операцию сложения, так и методы Фурье-анализа позволяют заменить сложные сигналы гармоническими составляющими и методами работы с гармоническими функциями.

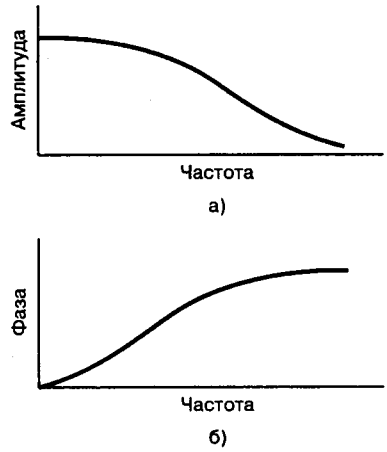


Рис. А.2. Передаточная функция системы: а) амплитудная характеристика; б) фазовая характеристика

А.2.1. Разложение в ряд Фурье

Периодические сигналы с конечной энергией, передаваемой за период, можно представить в виде *ряда Фурье*. Произвольный периодический сигнал $x(\lambda)$ выражается через бесконечное число гармоник с возрастающими частотами.

$$x(\lambda) = \tfrac{1}{2} a_0 + a_1 \cos \lambda + a_2 \cos 2\lambda + a_3 \cos 3\lambda + \ldots + \\ + b_1 \sin \lambda + b_2 \sin 2\lambda + b_3 \sin 3\lambda + \ldots \qquad (A.1)$$

Члены $\cos \lambda$ и $\sin \lambda$ называются *основными*; члены $\cos n\lambda$ и $\sin n\lambda$ при $n > 1$, где n — целое, именуются *гармоническими*. Члены a_n и b_n представляют коэффициенты гармоник, а $\frac{1}{2} a_0$ — это постоянный член или составляющая постоянного тока.

Период функции $x(\lambda)$ должен равняться 2π или кратной величине; кроме того, функция $x(\lambda)$ должна быть однозначной. Ряд Фурье можно рассматривать как "рецепт приготовления" *любого периодического* сигнала из синусоидальных составляющих. Чтобы данный ряд имел практическое значение, он должен сходиться, т.е. частичные суммы ряда должны иметь предел.

Процесс создания произвольного периодического сигнала из коэффициентов, описывающих смешивание гармоник, называется *синтезом*. Обратный процесс вычисления коэффициентов именуется *анализом*. Вычисление коэффициентов облегчается тем, что среднее от перекрестных произведений синусоиды на косинусоиду (а также среднее любой синусоиды или косинусоиды) равно нулю. Ниже приводятся формулы, иллюстрирующие основные свойства средних от гармонических функций.

$$\left.\begin{aligned}&\int_{-\pi}^{\pi}\sin m\lambda\, d\lambda = 0\\&\int_{-\pi}^{\pi}\cos m\lambda\, d\lambda = 0\\&\int_{-\pi}^{\pi}\sin m\lambda \cos n\lambda\, d\lambda = 0\end{aligned}\right\}\ \text{где } m \text{ и } n \text{ — любые целые} \qquad (A.2)$$

$$\left.\begin{aligned}&\int_{-\pi}^{\pi}\sin m\lambda \sin n\lambda\, d\lambda = 0\\&\int_{-\pi}^{\pi}\cos m\lambda \cos n\lambda\, d\lambda = 0\end{aligned}\right\}\ \text{при } m \neq n \qquad (A.3)$$

$$\left.\begin{aligned}&\int_{-\pi}^{\pi}(\sin m\lambda)^2 d\lambda = \pi\\&\int_{-\pi}^{\pi}(\cos m\lambda)^2 d\lambda = \pi\end{aligned}\right\}\ \text{при } m = n \qquad (A.4)$$

Рассмотрим, как вычисляются значения коэффициентов a_n или b_n в формуле (A.1). Например, для вычисления коэффициента a_3 обе стороны формулы (A.1) можно умножить на $\cos 3\lambda\, d\lambda$, а затем проинтегрировать.

$$\int_{-\pi}^{\pi} x(\lambda)\cos 3\lambda\, d\lambda = \underbrace{\int_{-\pi}^{\pi}\tfrac{1}{2}a_0\cos 3\lambda\, d\lambda}_{=0} + \underbrace{\int_{-\pi}^{\pi}a_1\cos\lambda\cos 3\lambda\, d\lambda}_{=0} +$$

$$+ \underbrace{\int_{-\pi}^{\pi}a_2\cos 2\lambda\cos 3\lambda\, d\lambda}_{=0} + \int_{-\pi}^{\pi}a_3(\cos 3\lambda)^2\, d\lambda + \ldots$$

$$+ \underbrace{\int_{-\pi}^{\pi}b_1\sin\lambda\cos 3\lambda\, d\lambda}_{=0} + \int_{-\pi}^{\pi}b_2\sin 2\lambda\cos 3\lambda\, d\lambda +$$

$$+ \underbrace{\int_{-\pi}^{\pi}b_3\sin 3\lambda\cos 3\lambda\, d\lambda}_{=0} + \ldots$$

$$\int_{-\pi}^{\pi} x(\lambda)\cos 3\lambda\, d\lambda = \int_{-\pi}^{\pi}a_3(\cos 3\lambda)^2\, d\lambda = a_3\pi$$

$$a_3 = \frac{1}{\pi}\int_{-\pi}^{\pi} x(\lambda)\cos 3\lambda\, d\lambda$$

Полученный вывод можно обобщить.

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} x(\lambda)\cos n\lambda\, d\lambda \tag{A.5}$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} x(\lambda)\sin n\lambda\, d\lambda \tag{A.6}$$

Коэффициент a_0 находится из (A.5) при $n = 0$. В результате получаем следующее.

$$\tfrac{1}{2}a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} x(\lambda)\, d\lambda \tag{A.7}$$

Данное выражение — это член нулевой частоты, или среднее значение периодического сигнала. Процесс синтеза уравнения (A.1) можно записать в более компактной форме.

$$x(\lambda) = \tfrac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n\cos n\lambda + b_n\sin n\lambda) \tag{A.8}$$

Существует несколько способов выражения *пары преобразований* (анализа и синтеза) Фурье. Наиболее распространенная форма — это выражение синуса и косинуса через экспоненты с комплексным показателем.

$$\cos\lambda = \frac{e^{i\lambda} + e^{-i\lambda}}{2} \tag{A.9}$$

$$\sin\lambda = \frac{e^{i\lambda} - e^{-i\lambda}}{2i} \tag{A.10}$$

Периодическая функция с периодом T_0 секунд имеет следующие частотные компоненты — f_0, $2f_0$, $3f_0$, ..., где $f_0 = 1/T_0$ называется *собственной частотой*. Иногда частотные компоненты записывают как ω_0, $2\omega_0$, $3\omega_0$, ..., где $\omega_0 = 2\pi/T_0$ именуется собственной *угловой* частотой; частота f измеряется в герцах, частота ω — в радианах в секунду. Заменим $n\lambda$ в аргументах гармонических функций в формулах (A.5)–(A.8) на $2\pi nf_0t = 2\pi nt/T_0$, где n — целое. При $n = 1$, nf_0 представляет собственную частоту, а при $n > 1$ — гармоники собственной частоты. Используя формулы (A.8)–(A.10), можно записать $x(t)$ в экспоненциальной форме.

$$x(t) = \frac{a_0}{2} + \frac{1}{2}\sum_{n=1}^{\infty}[(a_n - ib_n)e^{2\pi inf_0t} + (a_n + ib_n)e^{-2\pi inf_0t}] \tag{A.11}$$

Обозначим через c_n комплексные коэффициенты, или спектральные компоненты $x(t)$, связанные с коэффициентами a_n и b_n следующим образом.

$$c_n = \begin{cases} \frac{1}{2}(a_n - ib_n) & \text{при } n > 0 \\ \dfrac{a_0}{2} & \text{при } n = 0 \\ \frac{1}{2}(a_n + ib_n) & \text{при } n < 0 \end{cases} \tag{A.12}$$

Теперь формулу (A.11) можно упростить.

$$x(t) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n f_0 t} \tag{A.13}$$

Здесь коэффициенты экспоненциальных гармоник определяются следующим образом.

$$c_n = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x(t) e^{2\pi i n f_0 t} dt \tag{A.14}$$

Для проверки справедливости формулы (A.14) умножим обе части выражения (A.13) на $e^{2\pi i m f_0 t} dt/T_0$, проинтегрируем по промежутку $(-T_0/2, T_0/2)$ и используем следующую формулу.

$$\frac{1}{T_0} \int_{-T_0/2}^{T_0/2} e^{2\pi i (n-m) f_0 t} dt = \delta_{nm} = \begin{cases} 1 & \text{при } n = m \\ 0 & \text{при } n \neq m \end{cases} \tag{A.15}$$

Здесь δ_{nm} называется *дельта-функцией Кронекера*. После выполнения указанных действий получаем

$$\frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x(t) e^{2\pi i m f_0 t} dt = \sum_{n=-\infty}^{\infty} c_n \delta_{nm} = c_m \tag{A.16}$$

для всех целых m. В общем случае коэффициент c_n — комплексное число, выразить которое можно следующим образом.

$$c_n = |c_n| e^{i\theta_n} \tag{A.17}$$

$$c_{-n} = |c_n| e^{-i\theta_n}, \tag{A.18}$$

где

$$|c_n| = \tfrac{1}{2}\sqrt{a_n^2 + b_n^2} \tag{A.19}$$

$$\theta_n = \operatorname{arctg}\left(-\frac{b_n}{a_n}\right) \tag{A.20}$$

$$b_0 = 0 \quad \text{и} \quad c_0 = \frac{a_0}{2}$$

Значение $|c_n|$ определяет амплитуду n-й гармоники периодического сигнала, так что график зависимости $|c_n|$ от частоты, называемой *амплитудным спектром*, дает амплиту-

ду каждой из n дискретных гармоник сигнала. Подобным образом график зависимости θ_n от частоты, именуемой *фазовым спектром*, дает фазу каждой гармоники сигнала.

Коэффициенты Фурье вещественной периодической по времени функции обладают следующим свойством.

$$c_{-n} = c_n^*, \tag{A.21}$$

где c_n^* — комплексно сопряженное c_n. Таким образом, получаем следующее.

$$|c_{-n}| = |c_n| \tag{A.22}$$

Амплитудный спектр является четной функцией частоты. Подобным образом фазовый спектр θ_n — это нечетная функция частоты, поскольку из формулы (A.20) следует, что

$$\theta_{-n} = -\theta_n. \tag{A.23}$$

Итак, как отмечалось выше, ряды Фурье особенно полезны при описании произвольных периодических сигналов с конечной энергией каждого периода. Кроме того, они могут использоваться для описания непериодических сигналов, имеющих конечную энергию за конечный интервал. Впрочем, для таких сигналов более удобным является представление в виде интеграла Фурье (см. раздел A.2.3).

A.2.2. Спектр последовательности импульсов

В цифровой связи весьма важным сигналом является идеальная периодическая последовательность прямоугольных импульсов, показанная на рис. A.3. Для коэффициентов ряда Фурье последовательности импульсов $x_p(t)$ с периодом T_0 (каждый импульс имеет амплитуду A и длительность T) справедливо следующее выражение (проверить справедливость можно с помощью формул (A.14) и (A.10)).

$$c_n = \frac{AT}{T_0}\frac{\sin(\pi nT/T_0)}{\pi nT/T_0} = \frac{AT}{T_0}\operatorname{sinc}\frac{nT}{T_0} \tag{A.24}$$

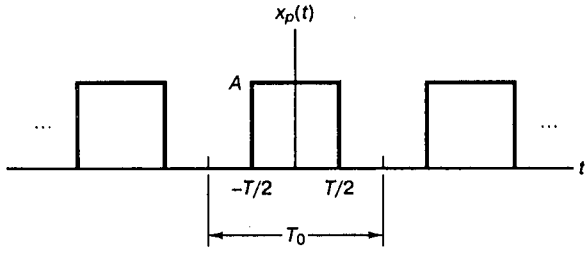


Рис. A.3. Последовательность импульсов

В данном выражении

$$\operatorname{sinc} y = \frac{\sin(\pi y)}{\pi y}$$

Функция sinc, как показано на рис. A.4, достигает максимума (единицы) при $y = 0$ и стремится к нулю при $y \to \pm\infty$, осциллируя с постепенно уменьшающейся амплитудой. Через нуль она проходит в точках $y = \pm 1, \pm 2, \ldots$. На рис. A.5, *а* как функция от-

ношения n/T_0 показан амплитудный спектр последовательности импульсов $|c_n|$, а на рис. А.5, *б* изображен фазовый спектр θ_n. Следует отметить, что положительные и отрицательные частоты двустороннего спектра — это весьма полезный способ математического выражения спектра; очевидно, что в лабораторных условиях воспроизвести можно только положительные частоты.

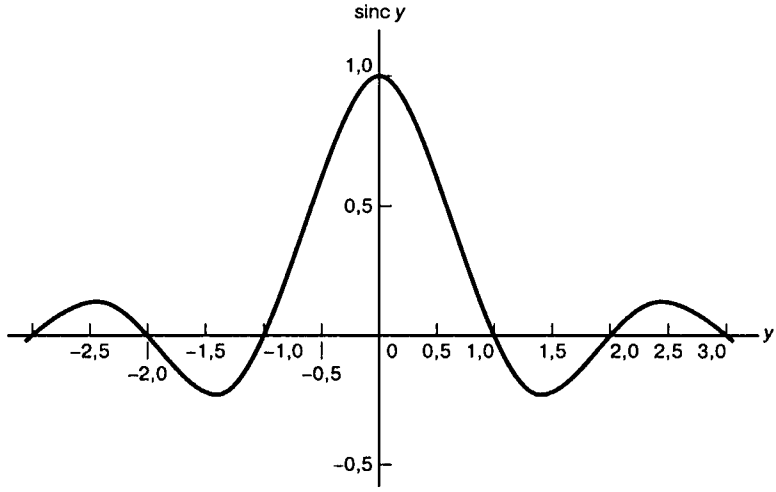


Рис. А.4. Функция sinc

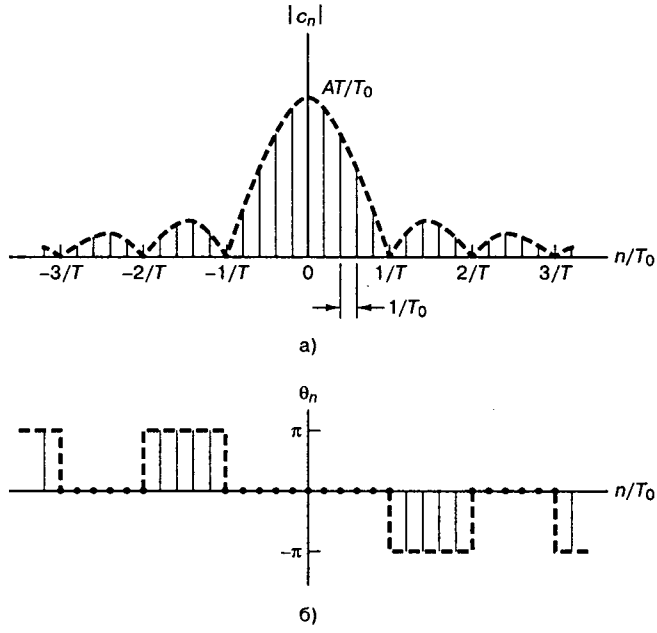


Рис. А.5. Спектр последовательности импульсов: а) амплитудный; б) фазовый

Синтез выполняется посредством подстановки коэффициентов из формулы (A.24) в формулу (A.13). Получаемый ряд представляет исходную последовательность импульсов $x_p(t)$, синтезированную из составных элементов.

$$x_p(t) = \frac{AT}{T_0} \sum_{n=-\infty}^{\infty} \operatorname{sinc} \frac{nT}{T_0} e^{2\pi i n f_0 t} \tag{A.25}$$

Идеальная периодическая последовательность импульсов включает все гармоники, кратные собственной частоте. В системах связи часто предполагается, что значительная часть мощности или энергии узкополосного сигнала приходится на частоты от нуля до первого нуля амплитудного спектра (рис. A.5, *а*). Таким образом, в качестве меры *ширины полосы* последовательности импульсов часто используется величина $1/T$ (где T — длительность импульса). Отметим, что ширина полосы обратно пропорциональна длительности импульса; чем меньше импульсы, тем более широкая полоса с ними связана. Отметим также, что расстояние между спектральными линиями $\Delta f = 1/T_0$ обратно пропорционально периоду импульсов; при увеличении периода линии располагаются ближе друг к другу.

A.2.3. Представление в виде интеграла Фурье

В системах связи часто встречаются непериодические сигналы, имеющие конечную энергию в конечном интервале и нулевую энергию за пределами этого интервала. Подобные сигналы удобно описывать, используя представление в виде интеграла Фурье, или просто *Фурье-образ*. Непериодический сигнал можно описать как периодический в предельном смысле. Рассмотрим, например, последовательность импульсов, показанную на рис. A.3. Если T_0 стремится к бесконечности, последовательность импульсов превращается в отдельный импульс $x(t)$, число спектральных линий стремится к бесконечности, а график спектра превращается в гладкий спектр частот $X(f)$. Для данного предельного случая можно определить пару интегральных преобразований Фурье.

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi i f t} dt \tag{A.26}$$

и

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{2\pi i f t} df\,, \tag{A.27}$$

где f — частота, измеряемая в герцах. Данную пару преобразований можно использовать при описании частотно-временных соотношений непериодических сигналов.

С этого момента применение преобразования Фурье будем обозначать $\mathfrak{F}\{\cdot\}$, а обратное преобразование — $\mathfrak{F}^{-1}\{\cdot\}$. Связь частотной и временной областей будем указывать с использованием знака "↔".

$$x(t) \leftrightarrow X(f)$$

Данная запись означает, что $X(f)$ получается в результате применения преобразования Фурье к $x(t)$, а $x(t)$ — в результате применения обратного преобразования Фурье к $X(f)$. В контексте систем связи $x(t)$ — вещественная функция, а $X(f)$ — комплексная функция, имеющая действительный и мнимый компоненты; в полярной форме спектр $X(f)$ можно задать через его амплитудную и фазовую характеристики.

$$X(f) = |X(f)|e^{i\theta(f)} \tag{A.28}$$

Свойства $X(f)$, спектра непериодического сигнала, подобны свойствам периодического сигнала, представленным в формулах (A.17)–(A.23); т.е. если $x(t)$ принимает вещественные значения,

$$X(-f) = X^*(f) = \tag{A.29}$$

$$= |X(f)|e^{-i\theta(f)}, \tag{A.30}$$

где X^* — комплексно сопряженное X. Амплитудный спектр $|X(f)|$ — это четная функция f, а фазовый спектр — нечетная функция f. Во многих случаях функция $X(f)$ имеет или только действительную часть, или только мнимую, так что для ее описания достаточно одного графика.

A.3. Свойства преобразования Фурье

Существует множество хороших справочников, в которых подробно рассмотрены преобразования Фурье и их свойства [1–4]. В данном приложении внимание акцентируется на свойствах, представляющих интерес в теории связи. Некоторыми ключевыми особенностями передач в системах связи являются временная задержка, сдвиг фазы, перемножение с другими сигналами, трансляция частоты, свертка сигнала и свертка спектра. Остановимся подробнее на свойствах преобразования Фурье (сдвиг и свертка), необходимых для описания данных особенностей.

A.3.1. Сдвиг во времени

Если $x(t) \leftrightarrow X(f)$, то

$$\mathfrak{F}\{x(t-t_0)\} = \int_{-\infty}^{\infty} x(t-t_0)e^{-2\pi ift}dt \tag{A.31}$$

Пусть $\mu = t - t_0$, тогда

$$\mathfrak{F}\{x(t-t_0)\} = \int_{-\infty}^{\infty} x(\mu)e^{-2\pi if(\mu+t)}d\mu =$$

$$= X(f)e^{-2\pi ift_0}$$

Если сигнал запаздывает во времени, амплитуда его частотного спектра не меняется, а фазовый спектр сдвигается по фазе. Сдвиг на время t_0 во временной области эквивалентен умножению на $e^{-2\pi ift_0}$ (сдвигу фазы на $-2\pi ft_0$) во временной области.

A.3.2. Сдвиг по частоте

Если $x(t) \leftrightarrow X(f)$, то

$$F\{x(t)e^{2\pi if_0t}\} = \int_{-\infty}^{\infty} x(t)e^{2\pi if_0t}e^{-2\pi ift}dt =$$

$$= \int_{-\infty}^{\infty} x(t) e^{-2\pi i(f - f_0)t} dt =$$

$$= X(f - f_0) \tag{A.32}$$

Выше приведено свойство *трансляции частоты*, которое описывает смещенный спектр, возникающий при умножении сигнала на $e^{2\pi i f_0 t}$. Используя формулу (A.32) вместе с формулой (A.9), можно получить выражения для Фурье-образа сигнала, умноженного на косинусоиду.

$$\begin{aligned} x(t)\cos 2\pi f_0 t &= \tfrac{1}{2}[x(t)e^{2\pi i f_0 t} + x(t)e^{-2\pi i f_0 t}] \\ x(t)\cos 2\pi f_0 t &\leftrightarrow \tfrac{1}{2}[X(f - f_0) + X(f + f_0)] \end{aligned} \tag{A.33}$$

Данное свойство также называется теоремой о *модуляции* (или *смешивании*). Умножение произвольного сигнала на синусоиду частоты f_0 приводит к трансляции исходного спектра сигнала на f_0 и $-f_0$.

A.4. Полезные функции

A.4.1. Дельта-функция

Полезной функцией в теории связи является так называемая *дельта-функция Дирака*, или единичный импульс, $\delta(t)$. Импульсную функцию можно определить из любой фундаментальной функции (например, прямоугольного или треугольного импульса). В любом случае импульсная функция определяется в пределе (амплитуда импульса стремится к бесконечности, длительность импульса — к нулю, а площадь импульса равна единице) [5]. Единичная импульсная функция имеет следующие свойства.

$$\int_{-\infty}^{\infty} \delta(t)\, dt = 1 \tag{A.34}$$

$$\delta(t) = 0 \quad \text{при } t \neq 0 \tag{A.35}$$

$$\delta(t) \text{ не ограничена в точке } t = 0 \tag{A.36}$$

$$\mathfrak{F}\{\delta(t)\} = \mathfrak{F}^{-1}\{\delta(f)\} = 1 \tag{A.37}$$

$$\int_{-\infty}^{\infty} x(t)\delta(t - t_0)\, dt = x(t_0) \tag{A.38}$$

Формула (A.38) представляет *просеивающее* (или *выборочное*) *свойство*; результат интегрирования функции $x(t)$ с дельта-функцией — выборка функции $x(t)$ в точке $t = t_0$.

В некоторых задачах полезными бывают следующие представления дельта-функции в частотной и временной областях.

$$\delta(t) = \int_{-\infty}^{\infty} e^{2\pi i f t}\, df \tag{A.39}$$

$$\delta(f) = \int_{-\infty}^{\infty} e^{-2\pi i f t}\, dt \tag{A.40}$$

A.4.2. Спектр синусоиды

Для нахождения Фурье-образа синусоидального сигнала необходимо предположить, что данный сигнал существует только в интервале ($-T_0/2 < t < T_0/2$). При таком условии функция будет иметь Фурье-образ, пока T_0 будет конечно. В пределе T_0 предполагается очень большим, но конечным. Спектр сигнала $x(t) = A\cos 2\pi f_0 t$ можно найти, используя формулы (A.9) и (A.26).

$$X(f) = \int_{-\infty}^{\infty} \frac{A}{2}(e^{2\pi i f_0 t} + e^{-2\pi i f_0 t})e^{-2\pi i f t} dt =$$

$$= \frac{A}{2}\int_{-\infty}^{\infty} [e^{-2\pi i (f - f_0)t} + e^{-2\pi i (f + f_0)t}]dt$$

Как видно из формулы (A.40), данное интегральное выражение можно записать через следующие единичные импульсные функции.

$$X(f) = \frac{A}{2}[\delta(f - f_0) + \delta(f + f_0)] \tag{A.41}$$

Подобным образом можно показать, что спектр синусоидального сигнала $y(t) = A\sin 2\pi f_0 t$ равен следующему.

$$Y(f) = \frac{A}{2i}[\delta(f - f_0) - \delta(f + f_0)] \tag{A.42}$$

Спектр косинусоидального сигнала показан на рис. A.6, а спектр синусоидального сигнала — на рис. A.7. Все дельта-функции на этих рисунках изображены как пики с весовыми коэффициентами $A/2$ или $-A/2$.

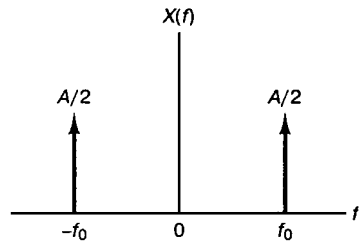


Рис. A.6. Спектр сигнала $x(t) = A\cos 2\pi f_0 t$

A.5. Свертка

В конце XIX века Оливер Хевисайд (Oliver Heaviside) использовал свертку для вычисления тока на выходе электрической схемы, на вход которой подан сигнал, описываемый сложной функцией напряжения. Использование методов Хевисайда предшествовало применению аналитических методов, разработанных Фурье и Лапласом (хотя публикации Фурье и Лапласа вышли раньше).

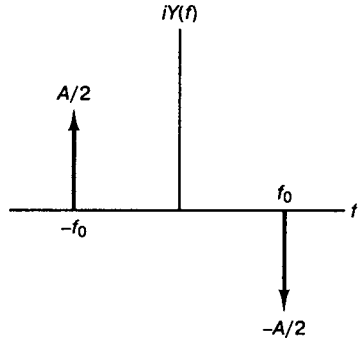


Рис. А.7. Спектр сигнала $y(t) = A \sin 2\pi f_0 t$

Отклик схемы на входное импульсное возмущение $v(t) = \delta(t)$ называется *импульсной характеристикой* и обозначается $h(t)$, как показано на рис. А.8, т.е. это просто выходное напряжение, полученное при подаче на вход дельта-функции. Хевисайд аппроксимировал произвольный сигнал, подобный показанному на рис. А.9, *а*, набором равноотстоящих импульсов. Подобные импульсы конечной высоты и ненулевой длительности показаны на рис. А.9, *б*. В пределе при длительности импульса $\Delta\tau \to 0$ каждый импульс стремится к дельта-функции с весовым коэффициентом, равным площади импульса. Далее будем считать, что данные равноотстоящие импульсы имеют нулевую длительность, хотя строго они являются такими *только в пределе*.

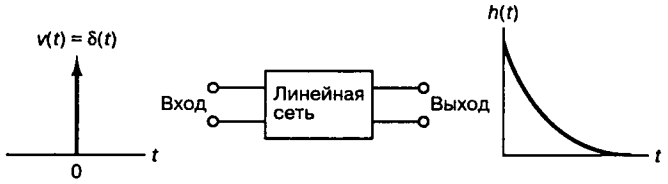


Рис. А.8. Импульсная характеристика линейной системы

Поскольку нас интересует как время подачи импульсов на вход, так и время наблюдения реакции на них на выходе, следует весьма аккуратно относиться к записи времени. Поэтому определим две различные временные последовательности; начнем с использования следующей формы записи.

1. Время на входе будем обозначать через τ, так что входные импульсы напряжения будут записываться как $v(\tau_1), v(\tau_2), \ldots, v(\tau_N)$.
2. Время на выходе будем обозначать через t, так что выходные функции тока будут записываться как $i(t_1), i(t_2), \ldots, i(t_N)$.

Хевисайд нашел отклик схемы (или ток на выходе) для каждого входного импульса; после этого он сложил эти токи и получил общий ток на выходе. Весовой коэффициент прямоугольного импульса, поданного в момент τ_1, — это произведение $v(\tau_1)\Delta\tau$. Если устремить $\Delta\tau$ к нулю, последовательность импульсов будет аппроксимировать произвольное входное напряжение настолько точно, насколько это нужно. Снова отметим, что момент подачи импульса на вход — это τ_i, а момент определения реакции системы — t_i, где τ — переменная входного времени, а t — переменная выходного времени, $i = 1, \ldots, N$.

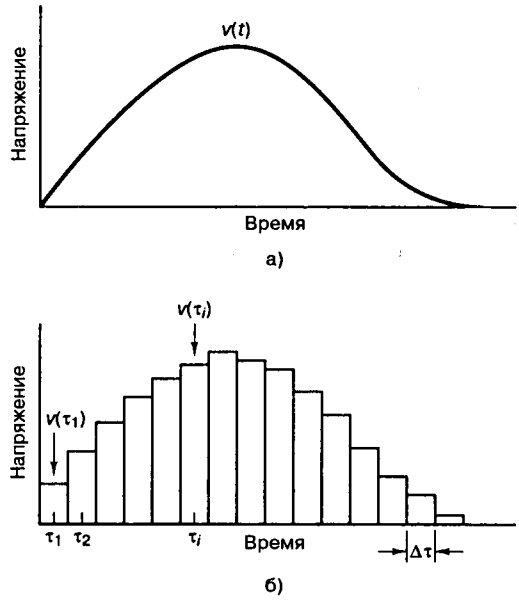


Рис. А.9. Аппроксимация произвольного входного сигнала: а) входной сигнал; б) аппроксимация входного сигнала

На рис. А.10 показана выходная реакция $i(t) = A_1 h(t - \tau_1)$ на импульс с весовым коэффициентом $v(\tau_1)$. Поскольку входной импульс в момент τ_1 *не* является *единичным*, он умножается на весовой коэффициент — интенсивность (или площадь) $A_1 = v(\tau_1)\,\Delta\tau$. В некоторый момент времени t_1, где $t_1 > \tau_1$, выходная реакция на импульс $v(\tau_1)$, как показано на рис. А.10, выражается следующим образом.

$$i(t_1) = A_1 h(t_1 - \tau_1) \quad \text{при } t_1 > \tau_1$$

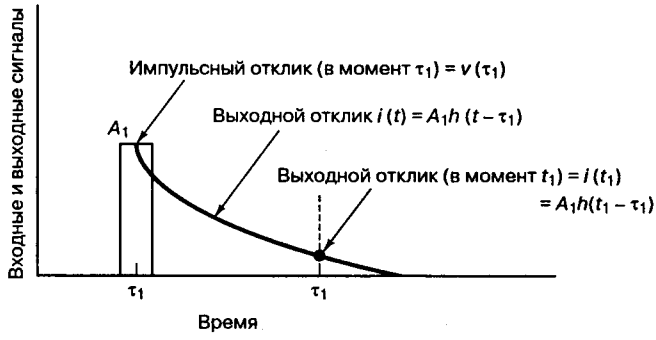


Рис. А.10. Реакция на импульс в момент τ_1

При наличии нескольких входных импульсов общий выходной отклик линейной системы — это просто сумма отдельных откликов. На рис. А.11 показан отклик сети на два единичных импульса. При N импульсах на входе ток на выходе, измеренный в момент времени t_1, можно записать следующим образом.

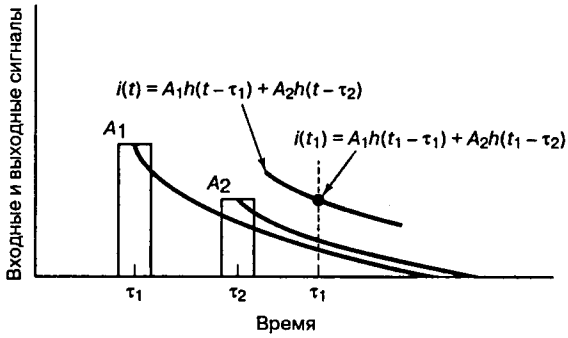


Рис. А.11. Реакция на два импульса

$$i(t_1) = A_1h(t_1 - \tau_1) + A_2h(t_1 - \tau_2) + \ldots + A_Nh(t_1 - \tau_N),$$

где импульсы подаются в моменты $\tau_1, \tau_2, \ldots, \tau_N$ и где $t_1 > \tau_N$.

Все импульсы, поданные на вход после момента t_1, не учитываются, поскольку они не дают вклада в $i(t_1)$. Это согласуется с требованием *причинности* физически реализуемых систем — отклик системы должен быть нулевым до применения возмущения. Итак, можно записать ток на выходе в любой момент времени t следующим образом.

$$i(t) = A_1h(t - \tau_1) + A_2h(t - \tau_2) + \ldots + A_Nh(t - \tau_N)$$

или, поскольку весовой коэффициент импульса в момент времени τ_j равен $v(\tau_j)$,

$$i(t) = \sum_{j=1}^{N} v(\tau_j)\Delta\tau h(t - \tau_j) \qquad \text{(A.43)}$$

Когда $\Delta\tau$ стремится к нулю, сумма входных импульсов — к действительному напряжению $v(\tau)$, $\Delta\tau$ можно заменить $d\tau$, при этом сумма переходит в *интеграл свертки.*

$$i(t) = \int_{-\infty}^{\infty} v(\tau)h(t - \tau)\, d\tau \qquad \text{(A.44,а)}$$

или

$$i(t) = \int_{-\infty}^{\infty} v(t - \tau)h(\tau)\, d\tau \qquad \text{(A.44,б)}$$

В сокращенной записи

$$i(t) = v(t) * h(t) \qquad \text{(A.45)}$$

Итак, $i(t)$ — это сумма реакций на отдельные импульсные возмущения, произведенные в некоторый входной момент τ, причем каждый импульс умножается на весовой коэффициент — интенсивность.

А.5.1. Графическая иллюстрация свертки

Рассмотрим квадратный импульс $v(t)$, подаваемый на вход линейной сети, импульсная характеристика которой равна $h(t)$ (рис. А.12, *а*). Отклик на выходе описывается интегралом свертки, представленным в формуле (А.44).

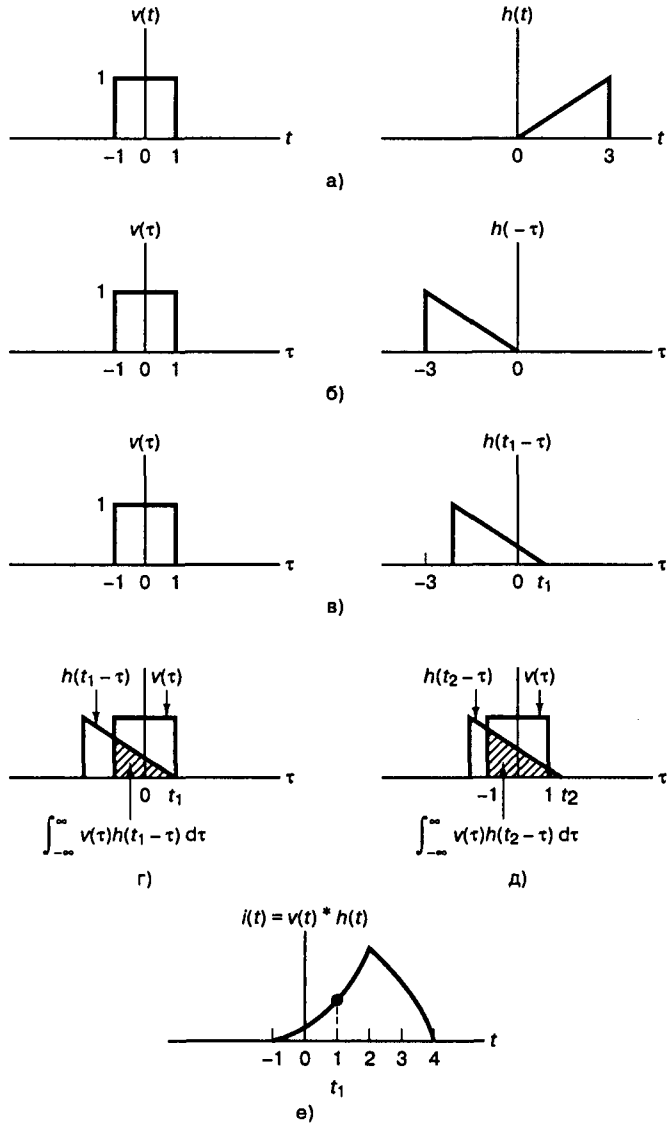


Рис. А.12. Графическая иллюстрация свертки

Независимой переменной в интеграле свертки является τ. На рис. А.12, *б* показаны функции $v(\tau)$ и $h(-\tau)$. Отметим, что $h(-\tau)$ получается отображением $h(\tau)$ относительно оси $\tau = 0$. Член $h(t - \tau)$ представляет функцию $h(-\tau)$, смещенную на t секунд вдоль положительного направления оси τ. На рис. А.12, *в* показана функция $h(t_1 - \tau)$. Значение

интеграла свертки в момент времени $t = t_1$ получается из формулы (A.44), в которой положено $t = t_1$. Это просто площадь под кривой произведения $v(\tau)$ на $h(t_1 - \tau)$, показанного на рис. A.12, *г*. Подобным образом интеграл свертки, взятый в момент $t = t_2$, равен заштрихованной области на рис. A.12, *д*. На рис. A.12, *е* приведен график отклика на выходе схемы при квадратном импульсе на входе, показанном на рис. A.12, *а*. Каждое вычисление интеграла свертки для некоторого момента времени t_i дает одну точку $i(t_i)$ графика на рис. A.12, *е*.

A.5.2. Свертка по времени

Если $x_1(t) \leftrightarrow X_1(f)$ и $x_2(t) \leftrightarrow X_2(f)$, то

$$x_1(t) * x_2(t) = \int_{-\infty}^{\infty} x_1(\tau) x_2(t - \tau)\, d\tau$$

$$\mathfrak{F}\{x_1(t) * x_2(t)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1(\tau) x_2(t - \tau)\, d\tau\, e^{-2\pi i f t}\, dt$$

Для линейных систем порядок интегрирования можно изменить.

$$\mathfrak{F}\{x_1(t) * x_2(t)\} = \int_{-\infty}^{\infty} x_1(\tau) d\tau \int_{-\infty}^{\infty} x_2(t - \tau) e^{-2\pi i f t} dt \tag{A.46}$$

С помощью *свойства сдвига во времени* второе интегральное выражение правой части можно заменить на $X_2(f) e^{-2\pi i f \tau}$.

$$\begin{aligned} \mathfrak{F}\{x_1(t) * x_2(t)\} &= X_2(f) \int_{-\infty}^{\infty} x_1(\tau) e^{-2\pi i f \tau} d\tau = \\ &= X_1(f) X_2(f) \end{aligned} \tag{A.47}$$

Следовательно, операцию *свертки* во временной области можно заменить *умножением* в частотной области.

A.5.3. Свертка по частоте

Можно показать, что, вследствие симметрии пары преобразований Фурье (формулы (A.26) и (A.27)), умножение во временной области переходит в свертку в частотной области.

$$x_1(t) x_2(t) \leftrightarrow X_1(f) * X_2(f) \tag{A.48}$$

Данный переход умножения в одной области в свертку в другой весьма удобен, поскольку, как правило, одну из этих операций выполнить значительно проще, чем другую. Например, ранее говорилось, что Хевисайд использовал свертку для нахождения тока на выходе линейной системы при подаче на вход произвольного переменного напряжения. Подобные методы часто включают вычисление (иногда трудоемкое) свертки входного сигнала с импульсной характеристикой системы. Поскольку, как видно из формулы (A.47), свертка во временной области переходит в умножение в частотной, для линейной системы входной сигнал можно просто умножить на переда-

точную функцию системы. Выходной сигнал затем получается путем применения к произведению обратного преобразования Фурье.

$$i(t) = \mathfrak{F}^{-1}\{V(f)H(f)\} \tag{A.49}$$

Вычислить выражение (A.49) часто намного проще, чем (A.45). В то же время, при определенных обстоятельствах, операция свертки настолько проста, что ее можно выполнить графически, просто внимательно изучив соответствующий график. Предположим, что некоторый произвольный сигнал необходимо умножить на косинусоиду фиксированной частоты, например несущую (если речь идет о модуляции). С помощью формулы (A.48) спектр произвольного сигнала можно свернуть со спектром косинусоиды, что, как показывается в следующем разделе, выполняется довольно просто.

A.5.4. Свертка функции с единичным импульсом

При использовании свойства, представленного в формуле (A.47), очевидно, что если

$$x(t) \leftrightarrow X(f)$$

и

$$\delta(t) \leftrightarrow 1,$$

то

$$x(t) * \delta(t) \leftrightarrow X(f). \tag{A.50}$$

Также должно быть очевидно, что

$$x(t) * \delta(t) = x(t) \tag{A.51}$$

и

$$X(f) * \delta(f) = X(f). \tag{A.52}$$

Следовательно, можно сделать вывод, что свертка функции с единичным импульсом дает исходную функцию. Простое развитие формулы (A.52) дает следующее.

$$X(f) * \delta(f - f_0) = X(f - f_0) \tag{A.53}$$

На рис. A.13 показано, насколько просто производится свертка спектра произвольного сигнала со спектром косинусоиды. На рис. A.13, *а* представлен спектр $X(f)$ произвольного узкополосного сигнала. На рис. A.13, *б* показан спектр $Y(f) = \delta(f - f_0) + \delta(f + f_0) = \mathfrak{F}\{2\cos 2\pi f_0 t\}$. Выход $Z(f) = X(f) * Y(f)$ на рис. A.13, *в* получается при свертке спектра сигнала с импульсной функцией $Y(f)$, согласно формуле (A.53), где импульсы действуют как стробирующие функции. Следовательно, в данном простом примере свертку можно выполнить графически, заметая стробирующие импульсы через спектр сигнала. Умножение на импульсные функции на каждом шаге заметания приводит к повторению спектра сигнала. Результат, показанный на рис. A.13, *в*, — это версия исходного спектра $X(f)$, смещенная к месторасположению импульсных функций, изображенных на рис. A.13, *б*.

A.5.5. Применение свертки при демодуляции

В разделе A.5.4 рассматривался сигнал, умноженный на $2\cos 2\pi f_0 t$. Было показано, как в частотной области выглядит свертка спектра сигнала со спектром косинусоиды. В данном разделе рассматривается обратный процесс. Необходимо демодулировать сигнал, умноженный на $2\cos 2\pi f_0 t$ (сигнал нужно восстановить в его изначальном диапазоне частот).

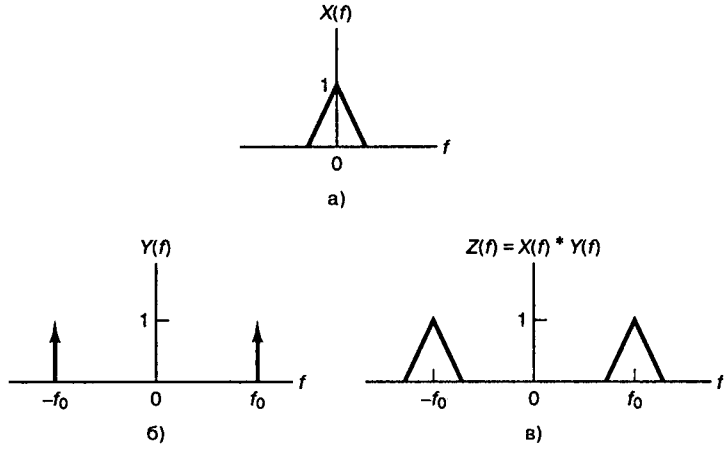


Рис. А.13. Свертка спектра сигнала со спектром косинусоиды

На рис. А.14, *а* представлен спектр, $Z(f)$, сигнала, смещенного вверх по частоте. Можно демодулировать данный смещенный сигнал и восстановить исходный сигнал, умножив данный сигнал на $2\cos 2\pi f_0 t$. Вместо этого мы можем проиллюстрировать процесс обнаружения в частотной области, свернув $Z(f)$ со спектром несущей, $Y(f) = \delta(f - f_0) + \delta(f + f_0)$, показанным на рис. А.14, *б*.

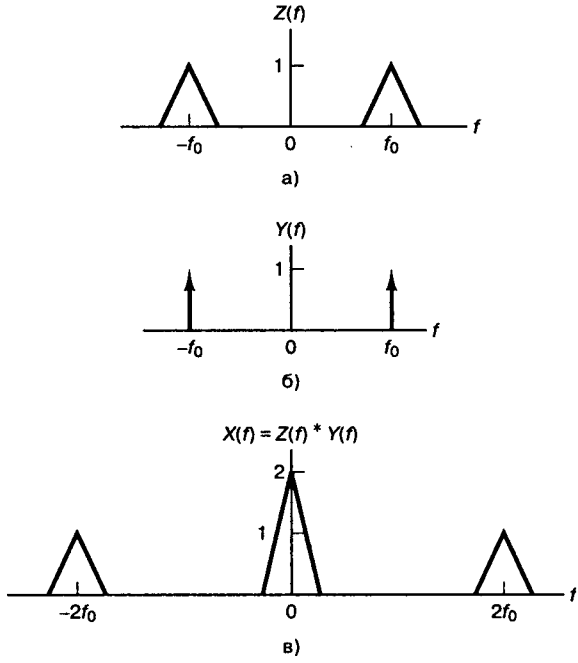


Рис. А.14. Применение демодуляции

Использование формул (А.52) и (А.53) позволяет записать следующее.

$$X(f - f_0) * \delta(f - f_1) = X(f - f_0 - f_1) \qquad \text{(А.54)}$$

Следовательно, результат демодуляции $X(f) = Z(f) * Y(f)$ получаем в результате применения формулы (A.54). Получающийся спектр сигнала — это спектр исходного сигнала плюс компоненты, центрированные на частотах $\pm 2f_0$, как показано на рис. A.14, *в*. Как и в предыдущем разделе, свертку можно выполнить графически. На рис. A.14, *в* отображены следующие члены.

$$\begin{aligned}&[Z(f-f_0)+Z(f+f_0)] * [\delta(f-f_0)+\delta(f+f_0)] = \\ &= Z(f-f_0) * \delta(f-f_0) + Z(f-f_0) * \delta(f+f_0) + \\ &\quad + Z(f+f_0) * \delta(f-f_0) + Z(f+f_0) * \delta(f+f_0) = \\ &= 2Z(f) + Z(f-2f_0) + Z(f+2f_0)\end{aligned} \tag{A.55}$$

Отметим, что результат — это спектр исходного сигнала плюс члены, связанные с высокочастотными компонентами. Данный результат типичен для процесса обнаружения; высокочастотные члены отфильтровываются и отбрасываются, оставляя спектр демодулированного исходного сигнала.

A.6. Таблицы Фурье-образов и свойств преобразования Фурье

В табл. A.1 и A.2 приведены Фурье-образы наиболее часто встречающихся функций и некоторые свойства преобразования Фурье.

Таблица A.1. Фурье-образы

$x(t)$	$X(f)$
1. $\delta(t)$	1
2. 1	$\delta(t)$
3. $\cos 2\pi f_0 t$	$\frac{1}{2}[\delta(f-f_0)+\delta(f+f_0)]$
4. $\sin 2\pi f_0 t$	$\frac{1}{2}[\delta(f-f_0)-\delta(f+f_0)]$
5. $\delta(t-t_0)$	$e^{(2\pi i f t_0)}$
6. $e^{(2\pi i f_0 t)}$	$\delta(f-f_0)$
7. $e^{(-a\|t\|)}, a>0$	$\frac{2a}{a^2+(2\pi f)^2}$
8. $\exp\left[-\pi\left(\frac{t}{T}\right)^2\right]$	$T\exp[-\pi(fT)^2]$
9. $u(t)=\begin{cases}1 & \text{при } t>0\\ 0 & \text{при } t<0\end{cases}$	$\frac{1}{2}\delta(f)+\frac{1}{2\pi i f}$
10. $\exp(-at)\,u(t), a>0$	$\frac{1}{a+2\pi i f}$

$x(t)$	$X(f)$
11. $t\exp(-at)\,u(t),\ a>0$	$\dfrac{1}{(a+2\pi if)^2}$
12. $\operatorname{rect}\left(\dfrac{t}{T}\right)$	$T\operatorname{sinc} fT$
13. $\cos 2\pi f_0 t\left[\operatorname{rect}\left(\dfrac{t}{T}\right)\right]$	$\dfrac{T}{2}[\operatorname{sinc}(f-f_0)T+\operatorname{sinc}(f+f_0)T]$
14. $W\operatorname{sinc} Wt$	$\operatorname{rect}\left(\dfrac{f}{W}\right)$
15. $\begin{cases}1-\dfrac{\lvert t\rvert}{T} & \text{при } \lvert t\rvert\le T\\ 0 & \text{при } \lvert t\rvert>T\end{cases}$	$T\operatorname{sinc}^2 fT$
16. $\sum\limits_{m=-\infty}^{\infty}\delta(t-mT_0)$	$\dfrac{1}{T}\sum\limits_{m=-\infty}^{\infty}\delta(f-\dfrac{m}{T_0})$

Примечание: $\operatorname{rect}(f/2W)=1$ для $-W<f<W$ и 0 для $|f|>W$; $\operatorname{sinc} x=(\sin\pi x)/\pi x$.

Таблица А.2. Свойства преобразования Фурье

Действие	$x(t)$	$X(f)$
1. Изменение масштаба	$x(at)$	$\dfrac{1}{\lvert a\rvert}X(\dfrac{f}{a})$
2. Сдвиг во времени	$x(t-t_0)$	$X(f)e^{2\pi ift_0}$
3. Сдвиг по частоте	$x(t)e^{2\pi ift_0}$	$X(f-f_0)$
4. Дифференцирование по времени	$\dfrac{d^n x}{dt^n}$	$(2\pi if)^n X(f)$
5. Дифференцирование по частоте	$(-it)^n x(t)$	$\dfrac{d^n X}{df^n}$
6. Интегрирование по времени	$\int\limits_{-\infty}^{t}x(\tau)d\tau$	$\dfrac{1}{2\pi if}X(f)+\dfrac{1}{2}X(0)\delta(f)$
7. Свертка по времени	$x_1(t)*x_2(t)$	$X_1(f)X_2(f)$
8. Свертка по частоте	$x_1(t)x_2(t)$	$X_1(f)*X_2(f)$

Литература

1. Papoulis A. *Signal Analysis.* McGraw-Hill Book Company, New York, 1977.
2. Panter P. F. *Modulation, Noise, and Spectral Analysis.* McGraw-Hill Book Company, New York, 1965.
3. Bracewell R. *The Fourier Transfer and Its Applications.* McGraw-Hill Book Company, New York, 1978.
4. Haykin S. *Communications Systems.* John Wiley & Sons, Inc., New York, 1983.
5. Schwartz M. *Information, Transmission, Modulation, and Noise.* McGraw-Hill Book Company, New York, 1980.

ПРИЛОЖЕНИЕ Б

Основы теории принятия статистических решений

Основными элементами задачи статистического принятия решений являются (1) набор гипотез, описывающих возможные истинные состояния природы, (2) тест, дающий данные, из которых мы можем сделать логический вывод, (3) правило принятия решения, применяемое к данным и определяющее, какая гипотеза наилучшим образом описывает состояние природы, и (4) критерий оптимальности. Все они рассматриваются ниже. *Критерий оптимальности* для правила принятия решения выбирается так, чтобы минимизировать вероятность принятия ошибочного решения, хотя возможны и другие критерии [1].

Предмет теории принятия статистических решений и проверки гипотез основывается на математической дисциплине *теория вероятностей и случайных переменных*. Предполагается, что читатель знаком с этим; в противном случае рекомендуется работа [2].

Б.1. Теорема Байеса

Математические основы проверки гипотез базируются на теореме Байеса, которая следует из определения отношения между условной вероятностью и совместной вероятностью случайных переменных A и B.

$$P(A|B)P(B) = P(B|A)P(A) = P(A, B) \tag{Б.1}$$

Теорема формулируется следующим образом.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{Б.2}$$

Теорема Байеса позволяет выводить условную вероятность $P(A|B)$ из условной вероятности $P(B|A)$.

Б.1.1. Дискретная форма теоремы Байеса

Теорему Байеса можно записать в дискретной форме следующим образом.

$$P(s_i|z_j) = \frac{P(z_j|s_i)P(s_i)}{P(z_i)} \quad i = 1, \ldots, M \qquad j = 1, \ldots, \tag{Б.3}$$

где

$$P(z_j) = \sum_{i=1}^{M} P(z_j|s_i)P(s_i)$$

В приложениях связи s_i — это i-й класс сигнала из набора M классов, а z_j — j-я выборка принятого сигнала. Уравнение (Б.3) можно рассматривать как описание эксперимента, в котором задействована принятая выборка и некоторые статистические знания о классах сигнала, к которым может принадлежать эта принятая выборка. До эксперимента вероятность появления i-го класса сигнала $P(s_i)$ называется *априорной*. В результате изучения конкретной принятой выборки z_j из плотности условной вероятности $P(z_j|s_i)$ можно найти статистическую меру *правдоподобия* принадлежности z_j к классу s_i. *После* эксперимента можно вычислить *апостериорную вероятность* $P(s_i|z_j)$, которую можно рассматривать как "уточнение" наших априорных знаний. Таким образом, к эксперименту мы приступаем, имея некоторые априорные знания, касающиеся вероятности состояния природы, а после изучения выборочного сигнала получаем апостериорную ("после свершения") вероятность. Параметр $P(z_j)$ — это вероятность принятой выборки z_j во всем пространстве классов сигналов. Этот термин, $P(z_j)$, можно рассматривать как масштабный множитель, поскольку его значение одинаково для *всех* классов сигнала.

Пример Б.1. Использование (дискретной формы) теоремы Байеса

Имеется два ящика деталей. Ящик 1 содержит 1000 деталей, 10% из которых неисправны, а ящик 2 — 2000 деталей, из которых неисправными являются 5%. Если в результате случайного выбора ящика и детали из него деталь оказывается исправной, то чему равна вероятность того, что данная деталь взята из ящика 1?

Решение

$$P(\text{ящик 1}|\text{ИД}) = \frac{P(\text{ИД}|\text{ящик 1})P(\text{ящик 1})}{P(\text{ИД})},$$

где ИД означает "исправная деталь".

$$P(\text{ИД}) = P(\text{ИД}|\text{ящик 1})P(\text{ящик 1}) + P(\text{ИД}|\text{ящик 2})P(\text{ящик 2}) =$$

$$= (0{,}90)(0{,}5) + (0{,}95)(0{,}5) =$$

$$= 0{,}450 + 0{,}475 = 0{,}925$$

$$P(\text{ящик 1}|\text{ИД}) = \frac{0{,}450}{0{,}925} = 0{,}486$$

До эксперимента априорные вероятности выбора ящика 1 или 2 равны. После получения исправной детали вычисления, проведенные согласно теореме Байеса, могут рассматриваться как способ "точной подстройки" нашего представления о том, что $P(\text{ящик 1}) = 0{,}5$, в результате которой возникает апостериорная вероятность 0,486. Теорема Байеса — это просто формализация здравого смысла. Если была получена исправная деталь, то не кажется ли вам

(интуитивно), что она с большей вероятностью могла быть взята из ящика с более высокой концентрацией исправных деталей и с меньшей — из ящика с меньшей концентрацией? Теорема Байеса уточняет априорную статистику выбора ящиков, порождая апостериорную статистику.

Пример Б.2. Применение теории принятия решений в теории игр

В ящике находится три монеты: обычная, с двумя орлами и с двумя решками. Вам предлагается случайным образом вытянуть одну монету, взглянуть на одну ее сторону и угадать, что находится на другой стороне. Какой стратегии лучше всего придерживаться?

Решение

Данную задачу можно рассматриваться как задачу обнаружения сигнала. Сигнал передается, но вследствие шума канала принятый сигнал не совсем отчетлив. Невозможность взглянуть на обратную сторону монеты равносильна приему сигнала, возмущенного шумом. Пусть H_i представляет гипотезу ($i = П, O, P$), где индексы $П$, O и P обозначают правильную монету, монету с двумя орлами и монету с двумя решками.

$$H_П = O, P \text{ (правильная монета)}$$

$$H_O = O, O \text{ (монета с двумя орлами)}$$

$$H_P = P, P \text{ (монета с двумя решками)}$$

Пусть z_j представляет принятую выборку ($j = O, P$), где z_O — орел, а z_P — решка. Пусть априорные вероятности гипотез равновероятны, так что $P(H_П) = P(H_O) = P(H_P) = 1/3$. Используем теорему Байеса.

$$P(H_i|z_j) = \frac{P(z_j|H_i)P(H_i)}{\sum_i P(z_j|H_i)P(H_i)}$$

Нам необходимо вычислить вероятности всех гипотез для всех классов сигнала. Следовательно, нам нужно изучить результаты *шести* вычислений, после чего мы сможем установить оптимальную стратегию принятия решения. В каждом случае значение $P(z_j|H_i)$ можно вычислить из условных вероятностей, изображенных на рис. Б.1. Пусть мы выбрали монету и увидели орел (z_O), тогда вычисление трех апостериорных вероятностей дает следующие результаты.

$$P(H_П|z_O) = \frac{\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)}{\left(\frac{1}{2}\right)\left(\frac{1}{3}\right) + (1)\left(\frac{1}{3}\right) + 0} = \frac{1}{3}$$

$$P(H_O|z_O) = \frac{(1)\left(\frac{1}{3}\right)}{\left(\frac{1}{2}\right)\left(\frac{1}{3}\right) + (1)\left(\frac{1}{3}\right) + 0} = \frac{2}{3}$$

$$P(H_P|z_O) = 0$$

Если принятой выборкой является решка (z_P), вычисления дают следующее.

$$P(H_n \mid z_m) = \frac{1}{3}$$

$$P(H_O|z_P) = 0$$

$$P(H_m \mid z_m) = \frac{2}{3}$$

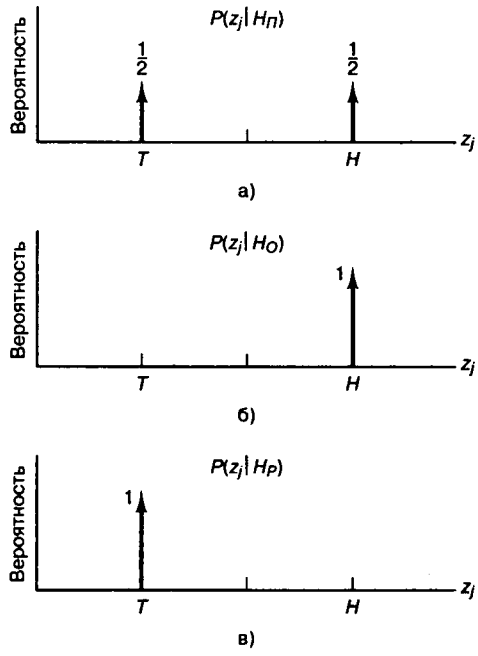


Рис. Б.1. Условная вероятность $P(z_j|H_i)$: а) для правильной монеты; б) для монеты с двумя орлами; в) для монеты с двумя решками

Таким образом, оптимальной стратегией принятия решения является следующая: если принят орел (z_O), выбрать гипотезу H_O (соответствующую монете с двумя орлами); если принята решка (z_P), выбрать гипотезу H_P (соответствующую монете с двумя решками).

Б.1.2. Теорема Байеса в смешанной форме

Для большинства приложений связи, представляющих практический интерес, возможные значения принятой выборки принадлежат *непрерывному* диапазону (причина — наличие в канале связи аддитивного гауссового шума). Следовательно, наиболее полезная форма теоремы Байеса содержит плотность вероятности с непрерывными, а не дискретными значениями. Изменим соответствующим образом формулу (Б.3).

$$P(s_i|z) = \frac{P(z|s_i)P(s_i)}{P(z)} \quad i = 1, \ldots, M$$

$$p(z) = \sum_{i=1}^{M} p(z|s_i)P(s_i) \qquad \text{(Б.4)}$$

Здесь $p(z|s_i)$ — плотность условной вероятности принятой выборки z (принимающей значения из непрерывного диапазона) при условии принадлежности к классу s_i.

Пример Б.3. Наглядное представление теоремы Байеса

Даны два класса сигнала s_1 и s_2, которые описываются треугольными функциями плотности условной вероятности $p(z|s_1)$ и $p(z|s_2)$, показанными на рис. Б.2. Принят некоторый сигнал;

Б.2. Теория принятия решений

Б.2.1. Элементы задачи теории принятия решений

После того как мы описали проверку гипотез на основе статистики Байеса, перейдем к рассмотрению элементов задачи теории принятия решений в контексте системы связи, как показано на рис. Б.3. Источник сигнала в передатчике состоит из множества $\{s_i(t)\}$, $i = 1, \ldots, M$ сигналов (или гипотез). Принимается сигнал $r(t) = s_i(t) + n(t)$, где $n(t)$ — присутствующий в канале аддитивный белый гауссов шум (additive white Gaussian noise — AWGN). В приемнике сигнал сокращается до единственного числа $z(t = T)$, которое может принимать любое значение. Поскольку шум является гауссовым процессом и приемник предполагается линейным, выход $z(t)$ также есть гауссовым процессом [1], а число $z(T)$ — *случайной переменной, принимающей значения из непрерывного диапазона.*

$$z(T) = a_i(T) + n_0(T) \tag{Б.5}$$

Выборка $z(T)$ составляется из сигнального компонента $a_i(T)$ и шумового компонента $n_0(T)$. Время T — это длительность символа. В каждый момент времени kT, где k — целое, приемник использует правило принятия решения для определения принадлежности принятого сигнала к определенному классу сигнала. Для простоты записи выражение (Б.5) иногда используют в виде $z = a_i + n_0$, где функциональная зависимость от T не выражается явно.

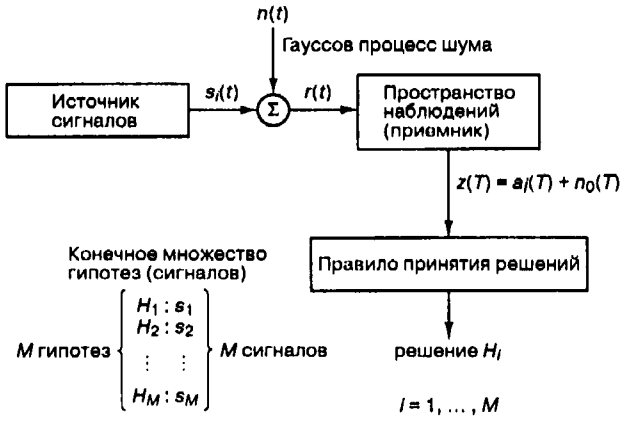


Рис. Б.3. Элементы задачи теории принятия решений в контексте системы связи

Б.2.2. Проверка методом отношения правдоподобий и критерий максимума апостериорной вероятности

При определении правила принятия решения для *двух* классов сигналов разумно начать со следующего соотношения.

$$P(s_1 \mid z) \underset{H_2}{\overset{H_1}{\gtrless}} P(s_2 \mid z) \tag{Б.6}$$

Выражение (Б.6) — это сокращенная запись следующего утверждения: "выбрать гипотезу H_1, если апостериорная вероятность $P(s_1|z)$ больше апостериорной вероятности $P(s_2|z)$; в противном случае выбрать гипотезу H_2".

Апостериорные вероятности в формуле (Б.6) можно заменить эквивалентными выражениями, полученными вследствие использования теоремы Байеса (уравнение (Б.4)), что дает следующее.

$$P(z|s_1)P(s_1) \underset{H_2}{\overset{H_1}{\gtrless}} P(z|s_2)P(s_2) \tag{Б.7}$$

Итак, у нас есть правило принятия решения, выраженное через плотности вероятности (правдоподобия). Если переписать выражение (Б.7) и привести его к следующему виду

$$\frac{P(z|s_1)}{P(z|s_2)} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{P(s_2)}{P(s_1)}, \tag{Б.8}$$

то отношение в левой части будет называться *отношением правдоподобий*, а все выражение часто именуют *критерием отношения правдоподобий*. Выражение (Б.8) — это принятие решений на основе сравнения принятого сигнала с порогом. Поскольку проверка опирается на выбор класса сигналов с максимальной апостериорной вероятностью, критерий принятия решения часто называется критерием *максимума апостериорной вероятности* (maximum a posteriori — MAP). Другое название — *критерий минимума ошибки*, поскольку в среднем он дает минимальное количество неверных решений. Стоит отметить, что данный критерий является оптимальным, только если ошибки всех типов наносят одинаковый вред (или имеют равную цену). Если ошибки некоторых типов обходятся дороже других, необходимо применять критерий, который учитывал бы относительные стоимости ошибок [1].

Б.2.3. Критерий максимального правдоподобия

Довольно часто сведения об априорных вероятностях гипотез или классов сигналов отсутствуют. Даже при наличии такой информации ее точность иногда вызывает сомнения. В таких случаях решения обычно принимаются исходя из предположения о возможности наиболее выгодной априорной вероятности; иными словами, значения априорных вероятностей выбираются так, чтобы классы были *равновероятными*. Если выбран такой подход, то критерий принятия решения является критерием максимального правдоподобия, и выражение (Б.8) записывается в следующем виде.

$$\frac{P(z|s_1)}{P(z|s_2)} \underset{H_2}{\overset{H_1}{\gtrless}} 1 \tag{Б.9}$$

Отметим, что критерий максимального правдоподобия, приведенный в выражении (Б.9), аналогичен правилу максимального правдоподобия, описанному в примере Б.3.

Б.3. Пример обнаружения сигнала

Б.3.1. Двоичное решение по принципу максимального правдоподобия

В наглядном представлении процесса принятия решения (пример Б.3) фигурировали треугольные функции плотности вероятности. На рис. Б.4 приведены функции плотностей условных вероятностей для двоичных выходных сигналов, искаженных шумом: $z(T) = a_1 + n_0$ и $z(T) = a_2 + n_0$. Сигналы a_1 и a_2 взаимно независимы и равновероятны. Шум n_0 предполагается независимой гауссовой случайной переменной с нулевым средним, дисперсией σ_0^2 и плотностью вероятности, описываемой следующей формулой.

$$p(n_0) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{n_0^2}{\sigma_0^2}\right)\right] \tag{Б.10}$$

Следовательно, отношение правдоподобий, выраженное в формуле (Б.8), можно записать следующим образом.

$$\begin{aligned}
\Lambda(z) &= \frac{p(z|s_1)}{p(z|s_2)} \\
&= \frac{\frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z-a_1}{\sigma_0}\right)^2\right]}{\frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z-a_2}{\sigma_0}\right)^2\right]} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{P(s_2)}{P(s_1)} \\
&= \frac{\exp\left(-\frac{z^2}{2\sigma_0^2}\right)\exp\left(-\frac{a_1^2}{2\sigma_0^2}\right)\exp\left(\frac{2za_1}{2\sigma_0^2}\right)}{\exp\left(-\frac{z^2}{2\sigma_0^2}\right)\exp\left(-\frac{a_2^2}{2\sigma_0^2}\right)\exp\left(\frac{2za_2}{2\sigma_0^2}\right)} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{P(s_2)}{P(s_1)} \\
&= \exp\left[\frac{z(a_1 - a_2)}{\sigma_0^2} - \frac{a_1^2 - a_2^2}{2\sigma_0^2}\right] \underset{H_2}{\overset{H_1}{\gtrless}} \frac{P(s_2)}{P(s_1)} .
\end{aligned} \tag{Б.11}$$

Здесь a_1 — сигнальный компонент на выходе приемника при переданном $s_1(t)$, а a_2 — сигнальный компонент на выходе приемника при переданном $s_2(t)$. Неравенство (Б.11) сохраняется при любом монотонно возрастающем (или убывающем) преобразовании.

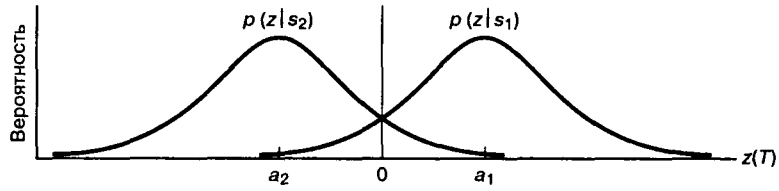


Рис. Б.4. Плотности условных вероятностей для типичного двоичного приемника

Следовательно, для упрощения выражения (Б.11) от его обеих частей можно взять натуральный логарифм, что даст логарифмическое отношение правдоподобий.

$$L(z) = \frac{z(a_1 - a_2)}{\sigma_0^2} - \frac{a_1^2 - a_2^2}{2\sigma_0^2} \underset{H_2}{\overset{H_1}{\gtrless}} \ln \frac{P(s_2)}{P(s_1)} \tag{Б.12}$$

Если классы равновероятны, то

$$\ln \frac{P(s_2)}{P(s_1)} = 0,$$

так что

$$\begin{gathered} z \underset{H_2}{\overset{H_1}{\gtrless}} \frac{a_1^2 - a_2^2}{2(a_1 - a_2)} \\ z \underset{H_2}{\overset{H_1}{\gtrless}} \frac{a_1 + a_2}{2} = \gamma_0 \end{gathered} \tag{Б.13}$$

Для *антиподных сигналов* $s_1(t) = -s_2(t)$ и $a_1 = -a_2$, так что можем записать следующее.

$$z \underset{H_2}{\overset{H_1}{\gtrless}} 0 \tag{Б.14}$$

Следовательно, правило максимального правдоподобия для равновероятных антиподных сигналов заключается в сравнении принятой выборки с нулевым порогом, что равносильно выбору $s_1(t)$, если выборка положительна, и выбору $s_2(t)$ — если она отрицательна.

Б.3.2. Вероятность битовой ошибки

Для двоичного примера, приведенного в разделе Б.3.1, рассчитаем вероятность битовой ошибки P_B с помощью правила принятия решений из формулы (Б.13). Вероятность ошибки вычисляется путем суммирования вероятностей различных возможностей появления ошибки.

$$P_B = P(H_2|s_1)P(s_1) + P(H_1|s_2)P(s_2) \tag{Б.15}$$

Другими словами, при переданном сигнале $s_1(t)$ ошибка произойдет, если будет выбрана гипотеза H_2; или ошибка произойдет, если при переданном сигнале $s_2(t)$ будет выбрана гипотеза H_1. Для частного случая симметричных функций плотности вероятности и для $P(s_1) = P(s_2) = 0{,}5$ можем записать следующее.

$$P_B = P(H_2|s_1) = P(H_1|s_2) \tag{Б.16}$$

Вероятность ошибки P_B равна вероятности принятия неверной гипотезы H_1 при переданном сигнале $s_2(t)$ или принятия неверной гипотезы H_2 при переданном сигнале $s_1(t)$. Следовательно, P_B численно равна площади под хвостом любой функции плотности вероятности, $p(z|s_1)$ или $p(z|s_2)$, "заползающим" на *неверную сторону* порога. Таким образом, P_B мы можем вычислить, проинтегрировав $p(z|s_1)$ от $-\infty$ до γ_0 или $p(z|s_2)$ от γ_0 до ∞.

$$P_B = \int_{\gamma_0 = (a_1 + a_2)/2}^{\infty} p(z|s_2)\,dz = \tag{Б.17}$$

$$= \int_{(a_1 + a_2)/2}^{\infty} \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - a_2}{\sigma_0}\right)^2\right] dz$$

Пусть

$$u = \frac{z - a_2}{\sigma_0}$$

Тогда $\sigma_0 du = dz$ и

$$P_B = \int_{u = (a_1 - a_2)/2\sigma_0}^{u = \infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du = Q\left(\frac{a_1 - a_2}{2\sigma_0}\right), \tag{Б.18}$$

где $Q(x)$, именуемая *гауссовым интегралом ошибок*[1], протабулирована в табл. Б.1.

Таблица Б.1. Гауссов интеграл ошибок $Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$

	$Q(x)$									
x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2168	0,2148
0,8	0,2169	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0599
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294

[1]Отметим, что гауссов интеграл ошибок определяется по-разному; впрочем, все определения, по сути, эквивалентны.

	$Q(x)$									
x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002

Еще одной часто используемой формой гауссова интеграла ошибок является следующая.

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} \exp(-u^2)\, du \tag{Б.19}$$

Функции $Q(x)$ и $\operatorname{erfc}(x)$ связаны следующим образом.

$$\operatorname{erfc}(x) = 2Q(x\sqrt{2}) \tag{Б.20}$$

$$Q(x) = \frac{1}{2}\operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) \tag{Б.21}$$

Литература

1. Van Trees H. L. *Detection, Estimation, and Modulation Theory.* Part 1, John Wiley & Sons. Inc., New York, 1968.
2. Papoulis A. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill Book Company, New York, 1965.

ПРИЛОЖЕНИЕ В

Отклик корреляторов на белый шум

На вход группы из N корреляторов подается белый гауссов процесс шума $n(t)$ с нулевым средним и двусторонней спектральной плотностью мощности $N_0/2$. Выходом каждого коррелятора в момент времени $t = T$ является *гауссова случайная переменная*, определяемая следующим образом.

$$n_j = \int_0^T n(t)\psi_j(t)\,dt \quad j = 1, \ldots, N \tag{B.1}$$

Здесь сигналы $\{\psi_j(t)\}$ формируют ортонормированное множество. Поскольку переменная n_j является гауссовой, она полностью определяется средним и дисперсией. Среднее n_j равно

$$\overline{n_j} = \mathbf{E}\{n_j\} = \mathbf{E}\left\{\int_0^T n(t)\psi_j(t)\,dt\right\}, \tag{B.2}$$

где $\mathbf{E}\{\cdot\}$ — оператор математического ожидания. Дисперсия n_j равна

$$\sigma_j^2 = \mathbf{E}\{n_j^2\} - \overline{n}_j^2 = \tag{B.3}$$

$$= \mathbf{E}\left\{\int_0^T n(t)\psi_j(t)\,dt \int_0^T n(s)\psi_j(s)\,ds\right\} - \overline{n}_j^2 = \tag{B.4}$$

$$= \int_0^T \int_0^T \mathbf{E}\{n(t)n(s)\psi_j(t)\psi_j(s)\}\,dt\,ds - \overline{n}_j^2. \tag{B.5}$$

Поскольку $n(t)$ — это процесс с нулевым средним,

$$\mathbf{E}\{n(t)\} = 0. \tag{B.6}$$

Отсюда следует

$$\bar{n}_j = \mathbf{E}\{n_j\} = 0 \tag{B.7}$$

Автокорреляционная функция процесса $n(t)$ равна следующему.

$$R_n(t, s) = \mathbf{E}\{n(t)n(s)\} \tag{B.8}$$

Если шум $n(t)$ предполагать стационарным, то $R_n(t, s)$ зависит только от разности времен $\tau = t - s$. Из уравнения (B.5) получаем следующее.

$$\sigma_j^2 = \mathrm{var}\{n_j\} = \int_0^T \int_0^T R_n(\tau)\psi_j(t)\psi_j(s)\, dt\, ds \tag{B.9}$$

Для стационарного случайного процесса спектральная плотность мощности $G_n(f)$ и автокорреляционная функция $R_n(\tau)$ являются Фурье-образами друг друга. Таким образом, можем записать следующее.

$$R_n(\tau) = \int_{-\infty}^{\infty} G_n(f) e^{2\pi i f \tau}\, df \tag{B.10}$$

Поскольку $n(t)$ — это белый шум, его спектральная плотность мощности $G_n(f)$ равна $N_0/2$ для всех f, и предыдущее выражение можно переписать следующим образом.

$$R_n(\tau) = \int_{-\infty}^{\infty} \frac{N_0}{2} e^{2\pi i f \tau} df = \frac{N_0}{2}\delta(\tau), \tag{B.11}$$

где $\delta(\tau)$ — единичная импульсная функция, определенная в разделе A.4.1. Подставляя выражение (B.11) в (B.9), получаем следующее.

$$\sigma_j^2 = \frac{N_0}{2} \int_0^T \int_0^T \delta(t - s)\psi_j(t)\psi_j(s)\, dt\, ds = \tag{B.12}$$

$$= \frac{N_0}{2} \int_0^T \psi_j^2(t) dt = \frac{N_0}{2} \quad j = 1, \ldots, N \tag{B.13}$$

Здесь было использовано *просеивающее свойство* единичной импульсной функции (см. раздел A.4.1) и то, что функции $\{\psi_j(t)\}$, $j = 1, \ldots, N$, составляют ортонормированное множество. Таким образом, для белого гауссова шума с двусторонней спектральной плотностью мощности $N_0/2$ Вт/Гц, мощность шума на выходе каждого из N корреляторов равна $N_0/2$ Вт.

ПРИЛОЖЕНИЕ Г

Полезные соотношения

$$\cos x \cos y = \tfrac{1}{2}\cos(x+y) + \tfrac{1}{2}\cos(x-y) \tag{Г.1}$$

$$\sin x \sin y = -\tfrac{1}{2}\cos(x+y) + \tfrac{1}{2}\cos(x-y) \tag{Г.2}$$

$$\sin x \cos y = \tfrac{1}{2}\sin(x+y) + \tfrac{1}{2}\sin(x-y) \tag{Г.3}$$

$$\cos x \sin y = \tfrac{1}{2}\sin(x+y) - \tfrac{1}{2}\sin(x-y) \tag{Г.4}$$

$$\sin(x \pm y) = \sin x \cos y \pm \cos x \sin y \tag{Г.5}$$

$$\cos(x \pm y) = \cos x \cos y \mp \sin x \sin y \tag{Г.6}$$

$$\cos^2 x = \tfrac{1}{2}(1 + \cos 2x) \tag{Г.7}$$

$$\sin^2 x = \tfrac{1}{2}(1 - \cos 2x) \tag{Г.8}$$

$$\sin x \cos x = \tfrac{1}{2}\sin 2x \tag{Г.9}$$

$$\sin x + \sin y = 2\sin\tfrac{1}{2}(x+y)\cos\tfrac{1}{2}(x-y) \tag{Г.10}$$

$$\sin x - \sin y = 2\cos\tfrac{1}{2}(x+y)\sin\tfrac{1}{2}(x-y) \tag{Г.11}$$

$$\cos x + \cos y = 2\cos\tfrac{1}{2}(x+y)\cos\tfrac{1}{2}(x-y) \tag{Г.12}$$

$$\cos x - \cos y = -2\sin\tfrac{1}{2}(x+y)\sin\tfrac{1}{2}(x-y) \tag{Г.13}$$

$$\sin x = \frac{e^{ix} - e^{-ix}}{2i} \tag{Г.14}$$

$$\cos x = \frac{e^{ix} + e^{-ix}}{2} \tag{Г.15}$$

$$P_B = \frac{1}{n}\sum_{j=2}^{n} j\binom{n}{j} p^j (1-p)^{n-j} = p - p(1-p)^{n-1} \tag{Г.16}$$

Доказательство

$$j\binom{n}{j} = j\frac{n!}{j!(n-j)!} = \frac{n!}{(j-1)!(n-j)!} = n\frac{(n-1)!}{(j-1)![(n-1)-(j-1)]!} = n\binom{n-1}{j-1}$$

$$P_B = p\sum_{j=2}^{n}\binom{n-1}{j-1}p^j(1-p)^{n-j} = p\sum_{j=2}^{n}\binom{n-1}{j-1}p^{j-1}(1-p)^{(n-1)-(j-1)}$$

Замена $i = (j-1)$

Таким образом ($j = 2$) переходит в ($i = 1$), а ($j = n$) — в ($i = n-1$).

$$P_B = p\sum_{i=1}^{n-1}\binom{n-1}{i}p^i(1-p)^{(n-1)-i} =$$

$$= p\sum_{i=0}^{n-1}\left[\binom{n-1}{i}p^i(1-p)^{(n-1)-i} - \binom{n-1}{0}p^0(1-p)^{(n-1)-0}\right] =$$

$$= p[1-(1-p)^{n-1}] =$$

$$= p - p(1-p)^{n-1}$$

ПРИЛОЖЕНИЕ Д

s-область, *z*-область и цифровая фильтрация

Роберт Стюарт (Robert W. Stewart)
Отдел электроники и электротехники
Университет Стратклайда, Глазго, Шотландия, Великобритания

В формулах (А.26) и (А.27) приложения А были определены прямое и обратное преобразования Фурье. Хотя преобразования Фурье и полезны для стационарного частотного анализа системы, они не всегда подходят для анализа переходных процессов. Для некоторых функций не существует интеграла Фурье, тогда как существует Лапласс-образ, рассматриваемый в данном приложении. Следовательно, для более глубокого анализа линейной системы часто выбирается именно преобразование Лапласа. Используя определения преобразований Фурье и Лапласа, легко показать, что последнее является расширением первого. Если анализируемая система — это система дискретного, а не непрерывного времени, можно использовать более простое (с точки зрения записи) *z*-преобразование (дискретное преобразование Лапласа), выводимое непосредственно из преобразования Лапласа. Еще одной причиной применения преобразования Лапласа (для анализа систем непрерывного времени) и *z*-преобразования (для анализа систем дискретного времени) является то, что операции, громоздкие во временной области (например, свертка), могут легче выполняться в *s*- или *z*-области.

Таким образом, в данном приложении рассматривается преобразование Лапласа, дискретное преобразование Лапласа и дискретное частотное преобразование, после чего описываются распространенные цифровые фильтры и представляется литература по названным преобразованиям.

Д.1. Преобразование Лапласа

Напомним преобразование Фурье, приведенное в формуле (А.26) приложения А.

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-2\pi ift}dt \text{ или } X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t}dt, \qquad (Д.1)$$

где $\omega = 2\pi f$.

Определим новую функцию $v(t)$, равную $x(t)$, умноженному на $e^{-\sigma t}$, где σ — вещественное число, т.е. $v(t) = x(t)e^{-\sigma t}$. Фурье-образ функции $v(t)$ будет выглядеть следующим образом.

$$V(\omega) = \int_{-\infty}^{\infty} v(t)e^{-i\omega t}dt = \int_{-\infty}^{\infty} x(t)e^{-\sigma t}e^{-i\omega t}dt = \int_{-\infty}^{\infty} x(t)e^{-(\sigma + i\omega)t}dt \qquad (Д.2)$$

Таким образом, можно переписать формулу (Д.1).

$$X(\sigma + i\omega) = \int_{-\infty}^{\infty} x(t)e^{-(\sigma + i\omega)t}dt \qquad (Д.3)$$

Пусть s — комплексная частота, $s = \sigma + i\omega$, тогда *Фурье-образ* временного сигнала $x(t)$ можно определить следующим образом.

$$X(s) = \int_{-\infty}^{\infty} x(t)e^{-st}dt, \qquad (Д.4)$$

где s — переменная Лапласа. Перепишем обратное преобразование Фурье, приведенное в формуле (А.27), через угловую частоту $\omega = 2\pi f$; тогда $d\omega/df = 2\pi$ и

$$x(t) = \int_{-\infty}^{\infty} X(\omega)e^{i\omega t}\frac{d\omega}{2\pi} \qquad (Д.5)$$

Поскольку $s = \sigma + i\omega$, из этого следует, что $ds/d\omega = i$, и мы можем определить *обратное преобразование Лапласа* следующим образом.

$$x(t) = \frac{1}{2\pi i}\int_{\sigma - i\infty}^{\sigma + i\infty} X(s)e^{st}ds \qquad (Д.6)$$

Формулы (Д.4) и (Д.6) представляют *пару преобразований Лапласа* $[x(t) \leftrightarrow X(s)]$, или, более точно, *пару двусторонних преобразований Лапласа*. Если (разумно) предположить, что до момента $t = 0$ сигнал не существует (т.е. является причинным), то преобразование можно назвать *односторонним*, что записывается следующим образом.

$$X(s) = \int_{0}^{\infty} x(t)e^{-st}dt \qquad (Д.7)$$

Обратное одностороннее преобразование Лапласа аналогично преобразованию, приведенному в формуле (Д.6). Таким образом, формулы (Д.6) и (Д.7) можно называть *парой односторонних преобразований Лапласа*.

Д.1.1. Стандартное преобразование Лапласа

В табл. Д.1 приведены некоторые стандартные односторонние преобразования Лапласа. Отметим, что (двустороннее) преобразование Лапласа, приведенное в формуле (Д.4), идентично преобразованию Фурье, приведенному в формуле (А.26), при $s = i\omega$, где $\omega = 2\pi f$. Для создания преобразования Лапласа $x(t)$ умножается на "множитель сходимости" $e^{-\sigma t}$, где σ — любое вещественное число. Таким образом, при фактическом вычислении значений интегралов преобразование Лапласа может существовать для многих функций, для которых отсутствует соответствующее преобразование Фурье. Одним из ключевых преимуществ преобразования Лапласа является возможность преобразования функций, не являющихся абсолютно интегрируемыми.

Таблица Д.1. Преобразования Лапласа

Тип сигнала	Временная функция	Преобразование Лапласа
Импульс	$\delta(t)$	1
Единичная ступенчатая функция (Хевисайда)	$u(t)$	$\frac{1}{s}$
Линейно растущая функция	$tu(t)$	$\frac{1}{s^2}$
Экспоненциальные функции	$e^{at}u(t)$	$\frac{1}{s-a}$
	$te^{at}u(t)$	$\frac{1}{(s-a)^2}$
Синусоида	$\sin(\omega t)u(t)$	$\frac{\omega}{(s^2+\omega^2)}$
Косинусоида	$\cos(\omega t)u(t)$	$\frac{s}{(s^2+\omega^2)}$
Затухающая синусоида	$e^{at}\sin(\omega t)u(t)$	$\frac{\omega}{(s-a)^2+\omega^2}$
Затухающая косинусоида	$e^{at}\cos(\omega t)u(t)$	$\frac{(s-a)}{(s-a)^2+\omega^2}$

Д.1.2. Свойства преобразования Лапласа

Можно показать, что если известна пара преобразований Лапласа $y(t) \leftrightarrow Y(s)$, то для запаздывающей версии сигнала, которая записывается как $y(t-t_0)$, справедливо следующее.

$$y(t-t_0) \leftrightarrow e^{-st_0}Y(s) \tag{Д.8}$$

Данное свойство называется свойством смещения во времени. Другие свойства преобразования Лапласа приведены в табл. Д.2. Их справедливость можно проверить путем простой подстановки в интегральное выражение, описывающее соответствующее преобразование. Отметим, что соотношение $s = i\omega$ между преобразованиями Фурье и Лапласа означает, что существует простой эквивалентный переход между преобразованиями, приведенными в табл. Д.1 и А.1, и свойствами, указанными в табл. Д.2 и А.2.

Таблица Д.2. Свойства преобразования Лапласа

Свойство	Временная функция	Преобразование Лапласа
Произвольная функция	$x(t)$	$X(s)$
Произвольная функция	$y(t)$	$Y(s)$
Линейность	$ax(t) + by(t)$	$aX(s) + bY(s)$
Сдвиг во времени ($\tau > 0$)	$x(t-\tau)$	$e^{-s\tau}X(s)$
Масштабирование времени	$x(at)$	$\frac{1}{a}X\left(\frac{s}{a}\right)$
Модуляция	$e^{-at}x(t)$	$X(s-a)$
Дифференцирование	$\frac{dx(t)}{dt}$	$sX(s) - x(0)$
Интегрирование	$\int_{-\infty}^{t} x(\tau)\,d\tau$	$\frac{X(s)}{s}$
Свертка	$x(t) * y(t)$	$X(s)Y(s)$

Д.1.3. Использование преобразования Лапласа

Преобразования Лапласа полезны, когда требуется решать дифференциальные (по времени) уравнения или выполнять операцию свертки. Например, для нахождения тока $i(t)$ простой RC-цепи, показанной на рис. Д.1, отметим, что сумма напряжений на конденсаторе и сопротивлении равна входному напряжению.

$$v_{\text{in}}(t) = i(t)R + \frac{q}{C} = i(t)R + \frac{1}{C}\int_0^t i(t)\,dt \tag{Д.9}$$

Если входное напряжение — это единичная ступенчатая функция $v_{\text{in}}(t) = u(t)$, а q — заряд конденсатора (в кулонах), то, применяя к обеим частям формулы (Д.9) преобразование Лапласа и используя табл. Д.1 и Д.2, получаем следующее.

$$V_{\text{in}}(s) = RI(s) + \frac{I(s)}{sC} \text{ откуда следует } I(s) = \frac{V_{\text{in}}(s)}{R + 1/(sC)} = \frac{1/R}{s + 1/(RC)} \tag{Д.10}$$

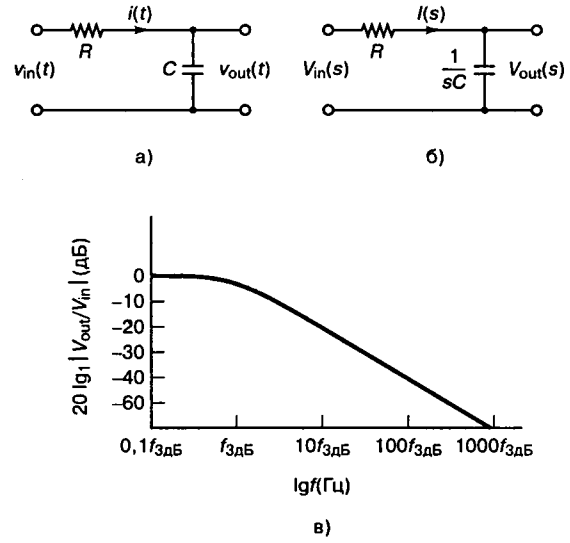


Рис. Д.1. Использование преобразования Лапласа: а) RC-контур; б) представление с помощью преобразования Лапласа; в) амплитудная характеристика

(Для единичной ступенчатой функции $V_{in}(s) = 1/s$.) Затем, возвращаясь во временную область (и снова используя таблицы свойств преобразования Лапласа), получаем следующее.

$$i(t) = \frac{1}{R} e^{-t/(RC)} \tag{Д.11}$$

Д.1.4. Передаточная функция

С помощью преобразования Лапласа можно определить (через переменную s) передаточную функцию линейной системы. Из уравнения (Д.10) при нулевом сопротивлении $R = 0$ импеданс конденсатора можно вычислить следующим образом.

$$Z_c = \frac{V_{in}(s)}{I(s)} = \frac{1}{sC} \tag{Д.12}$$

Входное и выходное напряжения (в s-области) можно записать следующим образом.

$$V_{in}(s) = I(s)R + \frac{I(s)}{sC} \text{ и } V_{out}(s) = \frac{I(s)}{sC} \tag{Д.13}$$

Таким образом, (в s-области) передаточную функцию можно определить следующим образом.

$$H(s) = \frac{V_{out}(s)}{V_{in}(s)} = \frac{\frac{I(s)}{sC}}{I(s)R + \frac{I(s)}{sC}} = \frac{1}{sRC + 1} \tag{Д.14}$$

Д.1.5. Фильтрация нижних частот в RC-цепи

Пусть на вход *RC*-цепи подается комплексная синусоида $v_{in}(t) = e^{i\omega t}$. Используя сказанное выше, можем перейти к преобразованию Фурье, положив $s = i\omega$, где $\omega = 2\pi f$. Таким образом, из передаточной функции можно получить частотную характеристику цепи.

$$\frac{V_{out}(f)}{V_{in}(f)} = \frac{1}{i\omega RC + 1} = \frac{1}{2\pi i f RC + 1} = \frac{1}{\sqrt{(2\pi f RC)^2 + 1}} e^{-i[\text{arctg}(2\pi f RC)]} \qquad (Д.15)$$

Для малых значений f $|H(f)| \approx 1$; а для больших значений f $|H(f)| \approx 0$. Если $f = f_0 = 1/(2\pi RC)$, то $|H(f)| \approx 1/\sqrt{2}$. Отметим, что $20\lg(1/\sqrt{2}) = -3$ дБ; следовательно, f_0 — это частота по уровню –3 дБ, когда выходное напряжение вдвое меньше входного. Следовательно, формула (Д.15) задает тот же фильтр нижних частот, что и формула (1.63). Низкие частоты проходят через фильтр, а высокие — подавляются; данная ситуация показана на рис. Д.1, *в*.

Д.1.6. Полюсы и нули

Линейные системы, а следовательно и (линейные) аналоговые фильтры, можно представить через дифференциальные уравнения во временной области. Рассмотрим, например, следующее уравнение второго порядка.

$$y(t) = A\frac{d^2x(t)}{dt^2} + B\frac{dx(t)}{dt} + Cx(t) + D\frac{d^2y(t)}{dt^2} + E\frac{dy(t)}{dt} \qquad (Д.16)$$

Реализация дифференцирования и/или интегрирования различных порядков происходит с использованием емкостей и индуктивностей вместе с усилителями с обратной связью, имеющими нужный порядок [2]. Применяя преобразование Лапласа к обеим частям уравнения (Д.16), получаем более удобное (с точки зрения математики и формы записи) уравнение Лапласа.

$$Y(s) = As^2X(s) + BsX(s) + CX(s) + Ds^2Y(s) + EsY(s) \qquad (Д.17)$$

Передаточная функция записывается в следующем виде.

$$H(s) = \frac{Y(s)}{X(s)} = \frac{As^2 + Bs + C}{-Ds^2 - Es + 1} = \frac{A(s - a_0)(s - a_1)}{-D(s - b_0)(s - b_1)} \qquad (Д.18)$$

Корни числителя $\{a_0, a_1\}$ называются *нулями*, а корни знаменателя $\{b_0, b_1\}$ — *полюсами*. Отметим, что если A, B и C — вещественны, нули $\{a_0, a_1\}$ являются комплексно-сопряженными.

Д.1.7. Устойчивость линейных систем

Рассмотрим однополюсное уравнение, соответствующее некоторой линейной системе.

$$H(s) = \frac{1}{s - \sigma} \qquad (Д.19)$$

Импульсную характеристику данной системы можно (используя табл. Д.1) найти как обратное преобразование Лапласа выражения (Д.19); если $\sigma = \rho + i\zeta$, то импульсная характеристика выглядит следующим образом.

$$h(t) = e^{\sigma t} = e^{\rho t} e^{i\zeta t} \qquad (Д.20)$$

Видим, что $\mathrm{Re}[\sigma] = \rho$; если $\rho > 0$, импульсная характеристика расходится с увеличением t (времени). В то же время, если $\rho < 0$, импульсная характеристика сходится с увеличением t. Член $e^{i\zeta t}$ — это комплексная (осциллирующая) синусоида (см. раздел А.2.1). Используя формулировку, несколько отличающуюся от применяемых ранее, можно сказать, что *система устойчива, если все полюса в s-области имеют отрицательную действительную часть.*

Таким образом, если изобразить полюса на комплексной s-плоскости, все они должны располагаться в ее левой части. На рис. Д.2 показана область устойчивости и приведен пример *устойчивой* передаточной функции третьего порядка, все полюса которой попадают в левую часть комплексной s-плоскости, т.е. имеют отрицательную действительную часть. Отметим, что нули функции могут быть в левой или правой части s-плоскости, и это не влияет на устойчивость.

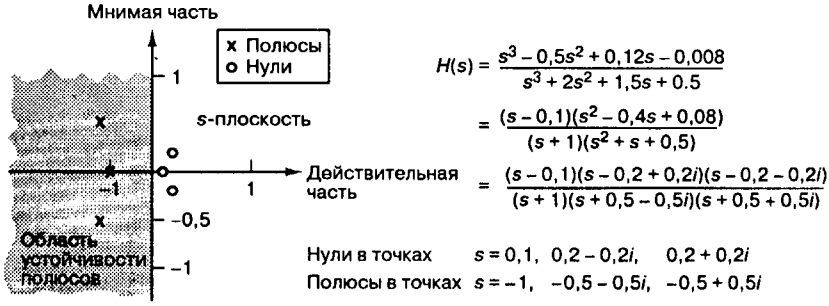


Рис. Д.2. Нули и полюса передаточной функции, изображенные в s-области

Если цепь имеет более одного полюса, передаточную функцию можно рассматривать как последовательность однополюсных функций.

$$H(s) = \frac{(s-a_0)(s-a_1)(s-a_2)}{(s-b_0)(s-b_1)(s-b_2)} = (s-a_0)(s-a_1)(s-a_2)\left[\frac{1}{s-b_0}\right]\left[\frac{1}{s-b_1}\right]\left[\frac{1}{s-b_2}\right] \qquad (Д.21)$$

Для устойчивости все полюсы должны находиться в левой части комплексной плоскости. Отметим, что для реальных схем с вещественными коэффициентами Лапласа (т.е. в уравнении (Д.16) A, B, C, D и E — вещественные) полюсы и нули будут вещественными или будут разбиты на пары комплексно-сопряженных величин, как показано на рис. Д.2.

Для нашего предыдущего примера RC-цепи передаточная функция в формуле (Д.14) является безусловно устойчивой, поскольку $2\pi RC$ — это всегда положительная величина, что, разумеется, является ожидаемым результатом. Неустойчивость в линейных системах возникает только при наличии в них обратной связи (рекурсии), например, при использовании фильтров с инвертирующими или неинвертирующими усилителями.

Д.2. z-преобразование

По сути, z-преобразование — это дискретный эквивалент преобразования Лапласа. Оно делает возможным удобный математический анализ (стационарный анализ и анализ переходных процессов) и манипулирование сигналами и спектрами. Возмож-

но, наиболее распространенным современным применением z-преобразования является описание дискретных систем и анализ их устойчивости.

z-преобразование позволяет вычислять свертку входного сигнала и характеристики дискретной линейной системы в математически удобном виде. Кроме того, могут определяться нули и полюса системы, что позволяет извлекать информацию о динамическом поведении и устойчивости дискретной системы. Следует отметить, что нули и полюса z-преобразования отличаются от нулей и полюсов преобразования Лапласа.

Д.2.1. Вычисление z-преобразования

z-преобразование можно вывести непосредственно из преобразования Лапласа, определенного в формуле (Д.4), рассмотрев для этого сигнал $x(t)$, выборка которого производится каждые T секунд. Таким образом, сигнал будет представлен как функция дискретного времени: $x(0)$, $x(T)$, $x(2T)$, ... = $\{x(kT)\}$. Дискретные данные представляют множество взвешенных и смещенных дельта-функций, применение к которым преобразования Лапласа дает следующий результат (использовано свойство сдвига во времени).

$$X(s) = \sum_{k=0}^{\infty} x(kT)e^{-skT} \qquad (Д.22)$$

Введем параметр $z = e^{sT}$ и заменим дискретное время kT номером выборки k. В результате получаем следующее.

$$X(z) = \sum_{k=0}^{\infty} x(k)z^{-k} \qquad (Д.23)$$

Приведем в качестве примера результат применения z-преобразования к единичной ступенчатой функции (Хевисайда).

$$U(z) = \sum_{k=0}^{\infty} u(k)z^{-k} = 1 + z^{-1} + z^{-2} + z^{-3} + \ldots = \frac{1}{1 - z^{-1}} \qquad (Д.24)$$

Выше при суммировании геометрической прогрессии было использовано предположение $|z| < 1$ (область сходимости). В табл. Д.3 и Д.4 приведены, соответственно, примеры применения z-преобразования к некоторым распространенным функциям и представлены полезные свойства данного преобразования.

Таблица Д.3. z-преобразование некоторых функций

Тип сигнала	Временная функция	z-преобразование
Импульс	$\delta(k)$	1
Задержанный импульс	$\delta(k - m)$	z^{-m}
Единичная ступенчатая функция (Хевисайда)	$u(k)$	$\frac{z}{z-1}$
Линейно растущая функция	$ku(k)$	$\frac{z}{(z-1)^2}$

Тип сигнала	Временная функция	z-преобразование
Экспоненциальная функция	$e^{ak}u(k)$	$\frac{z}{z-e^{a}}$
Синусоида	$\sin(\omega k)u(k)$	$\frac{z\sin\omega}{z^{2}-2z\cos\omega+1}$
Косинусоида	$\cos(\omega k)u(k)$	$\frac{z[z-\cos\omega]}{z^{2}-2z\cos\omega+1}$

Таблица Д.4. Свойства *z*-преобразования

Свойство	Временная функция	Преобразование Лапласа
Произвольная функция	$x(t)$	$X(z)$
Произвольная функция	$y(t)$	$Y(z)$
Линейность	$ax(t) + by(t)$	$aX(z) + bY(z)$
Сдвиг во времени	$x(k-m)$	$z^{-m}X(z)$
Модуляция	$e^{-i\omega k}x(k)$	$X(e^{i\omega}z)$
Экспоненциальное масштабирование	$a^{k}x(k)$	$X(z/a)$
Линейное масштабирование	$kx(k)$	$-z\frac{d}{dz}X(z)$
Свертка	$x(k) * h(k)$	$X(z)H(z)$

Д.2.2. Обратное *z*-преобразование

Переход из *z*-области во временную область выполняется посредством обратного *z*-преобразования [2].

$$x(k)=z^{-1}\{X(z)\}=\frac{1}{2\pi i}\oint_{C}\frac{X(z)z^{n}}{z}dz \qquad (Д.25)$$

Здесь интегрирование в комплексной области $\oint$ проводится по любому простому контуру в области сходимости $X(z)$, включающему точку $z=0$. Как правило, вычисление обратного *z*-преобразования сложнее вычисления прямого. Обычно приходится раскладывать подынтегральное выражение на сумму рациональных дробей, делить полиномы, использовать теорему о вычетах и составлять разностные уравнения. Поэтому большая часть *z*-преобразований и обратных *z*-преобразований вычисляется с использованием таблиц интегралов и их свойств, так что явного вычисления выражения (Д.25) обычно удается избежать. При современном анализе цифровых сигналов и систем используются программные пакеты, подобные SystemView [1], а *z*-преобразование большей частью представляет собой просто аналитическую форму записи, удобную для определения устойчивости дискретных сигналов и систем.

Д.3. Цифровая фильтрация

С помощью подходящих аналоговых и цифровых компонентов цифровой фильтр можно настроить на выполнение селекции желаемой частоты или модификации фазы. На рис. Д.3 показаны компоненты, необходимые для создания цифрового фильтра, дающего фильтрованную последовательность $y(k)$ при входной последовательности $x(k)$ [2]. Выходной сигнал фильтра $y(k)$ создается из взвешенной суммы предыдущих входных сигналов $x(k)$ и предыдущих выходных сигналов $y(k-n)$, где $n > 0$. На рис. Д.4 показан поточный граф сигнала (состоит только из сумматоров, умножителей и схем задержки выборки) для цифрового фильтра с четырьмя весовыми коэффициентами прямой связи и тремя весовыми коэффициентами обратной связи. (Задержка, длительность которой равна длительности одной выборки, обозначена символом Δ. Довольно часто подобные графы изображаются с использованием обозначений временной области и z-области, где для представления задержки применяется запись z^{-1}; несмотря на широкое распространение такой формы записи, она не является строгой.)

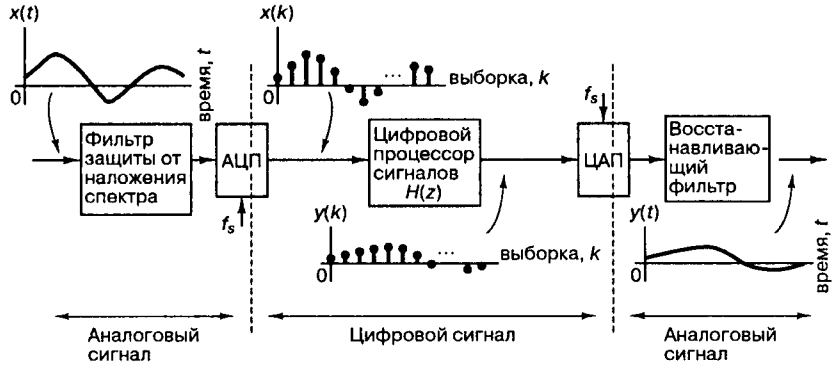


Рис. Д.3. Уравнения цифрового фильтра реализуются на устройстве цифровой обработки сигналов, преобразовывающем входной дискретный информационный сигнал в выходной дискретный информационный сигнал

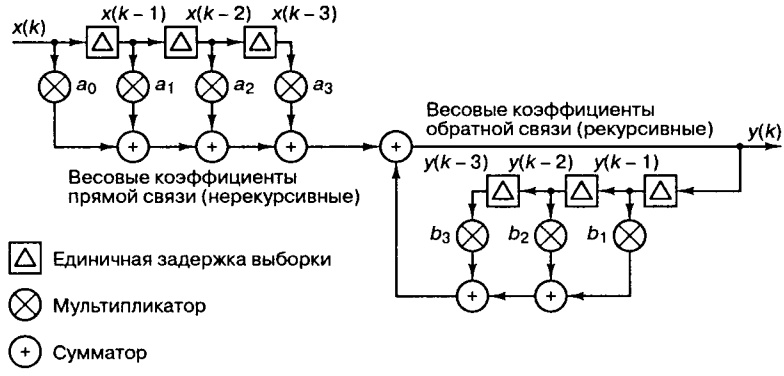


Рис. Д.4. Общая схема цифрового фильтра

Выход данного фильтра описывается следующим выражением.

$$\begin{aligned} y(k) &= a_0x(k) + a_1x(k-1) + a_2x(k-2) + a_3x(k-3) + \\ &+ b_1y(k-1) + b_2y(k-2) + b_3y(k+3) = \\ &= \sum_{n=0}^{3} a_n x(k-n) + \sum_{m=1}^{3} b_m y(k-m) \end{aligned} \tag{Д.26}$$

Применение z-преобразования к формуле (Д.26) дает следующий результат.

$$\begin{aligned} Y(z) &= a_0X(z) + a_1X(z)z^{-1} + a_2X(z)z^{-2} + a_3X(z)z^{-3} + \\ &+ b_1Y(z)z^{-1} + b_2Y(z)z^{-2} + b_3Y(z)z^{-3} \end{aligned} \tag{Д.27}$$

Д.3.1. Передаточная функция цифрового фильтра

Передаточная функция цифрового фильтра, изображенного на рис. Д.4, получается после преобразования выражения (Д.27) и выглядит следующим образом.

$$\begin{aligned} H(z) &= \frac{Y(z)}{X(z)} = \frac{a_0 + a_1z^{-1} + a_2z^{-2} + a_3z^{-3}}{1 - b_1z^{-1} + b_2z^{-2} + b_3z^{-3}} = \\ &= \frac{a_0(1-\alpha_1z^{-1})(1-\alpha_2z^{-1})(1-\alpha_3z^{-1})}{(1-\beta_1z^{-1})(1-\beta_2z^{-1})(1-\beta_3z^{-1})} = \\ &= \frac{a_0(z-\alpha_1)(z-\alpha_2)(z-\alpha_3)}{(z-\beta_1)(z-\beta_2)(z-\beta_3)} = \frac{A(z)}{B(z)} \end{aligned} \tag{Д.28}$$

Здесь через α_i обозначены *нули*, а через β — *полюса* z-области, которые находятся как корни полинома числителя $A(z)$ и полинома знаменателя $B(z)$. Для цифрового фильтра, подобного изображенному на рис. Д.4, но имеющего N весовых коэффициентов прямой связи и $M-1$ коэффициентов обратной связи, полиномы числителя и знаменателя в передаточной функции, приведенной в формуле (Д.28), будут иметь, соответственно, порядок N и M.

Д.3.2. Устойчивость однополюсного фильтра

Вследствие наличия в потоковом графе множественных обратных связей, цифровой фильтр может быть (численно) неустойчивым. Рассмотрим, например, фильтр с одним весовым коэффициентом обратной связи, изображенный на рис. Д.5.

$$y(k) = x(k) + by(k-1) \tag{Д.29}$$

Импульсная характеристика данного фильтра (т.е. подача на вход единичного импульса $\delta(k)$ плюс применение принципов свертки, описанных в разделе А.5) имеет следующий вид.

$$h(k) = b^k \tag{Д.30}$$

Если $|b| < 1$, импульсная характеристика фильтра сходится (устойчива); если $|b| > 1$, импульсная характеристика фильтра расходится (неустойчива). На рис. Д.5 показана сходящаяся импульсная характеристика с $|b| < 1$; более точно, $-1 < b < 1$. Применение z-преобразования к выражению (Д.29) дает следующее.

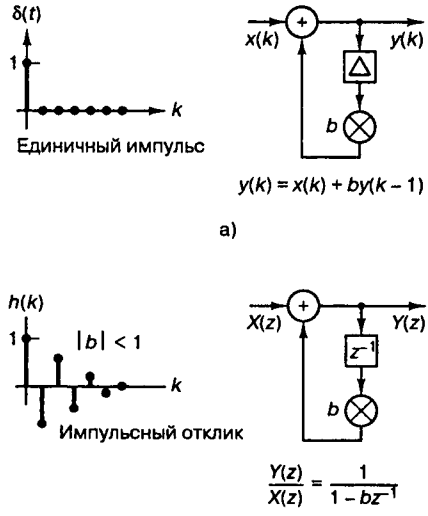


Рис. Д.5. Потоковый граф фильтра с одной обратной связью: а) во временной области; б) в z-области

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{1 - bz^{-1}} = \frac{z}{z - b} \qquad (Д.31)$$

Используя формулу (Д.31), получаем потоковый граф в z-области (рис. Д.5, *б*), соответствующий потоковому графу во временной области, изображенному на рис. Д.5, *а*. Элемент задержки (который на рис. Д.5, *а* обозначен через Δ) теперь представляется как z^{-1}, а вход и выход заданы как z-образы $X(z)$ и $Y(z)$. Отметим, впрочем, что общая топология двух графов одинакова. (Это частично объясняет то, что потоковые графы цифровых фильтров часто изображаются с использованием обозначений временной области и z-области.) Критерий устойчивости ($|b| < 1$) можно сформулировать следующим образом: *система устойчива, если полюсы (или корни полинома знаменателя) передаточной функции цифрового фильтра меньше единицы.*

Д.3.3. Устойчивость произвольного фильтра

При изучении факторизованной передаточной функции, приведенной в формуле (Д.28), поточный граф, представленный на рис. Д.4 для временной области, можно преобразовать в поточный граф в z-области (рис. Д.6). Последний граф — это, фактически, графическое представление формулы (Д.28), переписанной в следующем виде.

$$H(z) = a_0(1 - \alpha_1 z^{-1}) \cdot (1 - \alpha_2 z^{-1}) \cdot (1 - \alpha_3 z^{-1}) \cdot \left[\frac{1}{1 - \beta_1 z^{-1}}\right] \cdot \left[\frac{1}{1 - \beta_2 z^{-1}}\right] \cdot \left[\frac{1}{1 - \beta_3 z^{-1}}\right] \qquad (Д.32)$$

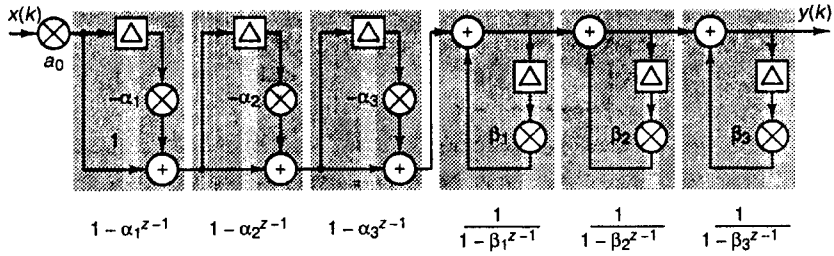


Рис. Д.6. Цифровой фильтр как последовательность каскадов прямой и обратной связи первого порядка

В данном выражении (и на рисунке) обособлены все блоки первого порядка, описываемые нулями и полюсами фильтра. Чтобы фильтр был устойчивым, модули всех полюсов $\{\beta_1, \beta_2, \beta_3\}$ каскада должны быть меньше 1. Если хотя бы один блок первого порядка неустойчив (или расходится), неустойчивым является и весь каскад. Как отмечалось для преобразования Лапласа, полюса (и нули) z-области могут быть комплексными, поэтому в качестве критерия устойчивости используется их абсолютная величина, а не амплитуда. (Стоит сказать, что реализация поточного графа, представленная на рис. Д.6, — это всего лишь иллюстрация принципов анализа; реальный цифровой фильтр никогда не реализуется в подобной факторизованной форме, поскольку в этом случае некоторые множители могут быть комплексными, а это может повлечь за собой ненужное усложнение вычислительных требований фильтров.)

Д.3.4. Диаграмма полюсов-нулей и единичная окружность

Если комплексные нули и полюса фильтра или линейной системы изобразить на плоскости с действительной и мнимой осями, данную плоскость можно будет назвать z-плоскостью (или комплексной плоскостью). Система является устойчивой, если все ее полюса находятся внутри *единичной окружности*. На рис. Д.7 показан вид z-плоскости для следующей передаточной функции.

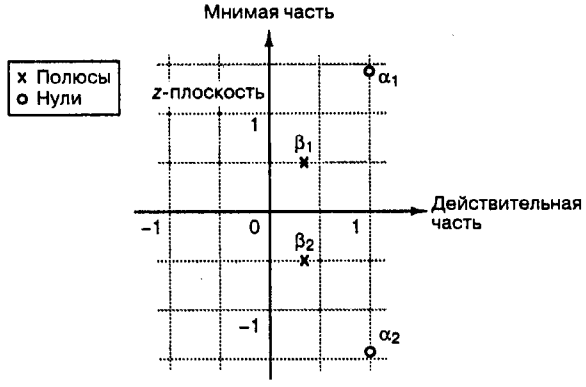


Рис. Д.7. Полюсы и нули, изображенные на z-плоскости

$$H(z)=\frac{1-2z^{-1}+3z^{-2}}{1-\frac{2}{3}z^{-1}+\frac{1}{3}z^{-2}}=$$
$$=\frac{(1-(1+i\sqrt{2})z^{-1})(1-(1-i\sqrt{2})z^{-1})}{(1-(1/3+i\sqrt{2}/3)z^{-1})(1-(1/3-i\sqrt{2}/3)z^{-1})}= \qquad (Д.33)$$
$$=\frac{(1-\alpha_1 z^{-1})(1-\alpha_2 z^{-1})}{(1-\beta_1 z^{-1})(1-\beta_2 z^{-1})}$$

Нули данной функции — $z=1+i\sqrt{2}$ и $z=1-i\sqrt{2}$, полюсы — $z=1/3+i\sqrt{2}/3$ и $z=1/3-i\sqrt{2}/3$. Поскольку все полюсы лежат внутри единичной окружности, данный фильтр является устойчивым.

Д.3.5. Дискретное преобразование Фурье импульсной характеристики цифрового фильтра

Частотная характеристика цифрового фильтра вычисляется из дискретного преобразования Фурье (discrete Fourier transform — DFT, ДПФ) импульсной характеристики фильтра. Напомним вид преобразования Фурье, приведенного в формуле (А.26).

$$X(f)=\int_{-\infty}^{\infty}x(t)e^{-2\pi ift}dt \qquad (Д.34)$$

Данную формулу можно использовать для вычисления Фурье-образа импульсной характеристики фильтра. Ее можно упростить, полагая, что используется дискретная версия сигнала $x(t)$, причем выборка сигнала производится каждые $T_s=1/f_s$ секунд.

$$X(f)=\int_{-\infty}^{\infty}x(kT_s)e^{-2\pi ifkT_s}d(kT_s)=\sum_{k=-\infty}^{\infty}x(kT_s)e^{-2\pi ifkT_s}=\sum_{k=-\infty}^{\infty}x(kT_s)e^{-(2\pi ifk)/f_s} \qquad (Д.35)$$

Разумеется, импульсная характеристика цифрового фильтра является причинной, и первая выборка импульсной характеристики производится в момент $k=0$, а последняя — в момент $k=N-1$, что в сумме дает N выборок на одно преобразование. Таким образом, для данного конечного числа выборок можно переписать формулу (Д.25), использовав не явное время kT_s, а число выборок k.

$$X(f)=\sum_{k=0}^{N}x(k)e^{-(2\pi ifk)/f_s} \qquad (Д.36)$$

Отметим, что значение выражения (Д.36) вычисляется для непрерывной частотной переменной f. В действительности же нам требуется знать это значение для некоторых определенных частот — нулевой частоты (постоянной составляющей) и гармоник "собственной" частоты; всего N дискретных частот: $0, f_0, 2f_0$ и так до f_s, где $f_0=1/NT_s$.

$$X\left(\frac{nf_s}{N}\right) = \sum_{k=0}^{N-1} x(k)e^{-(2\pi ikf_s n)/Nf_s} \quad \text{для } n \text{ от } 0 \text{ до } N-1 \qquad (Д.37)$$

Выражение выше можно упростить, использовав только временной индекс k и частотный индекс n. В результате получаем *дискретное преобразование Фурье* (discrete Fourier transform — DFT, ДПФ).

$$X(n) = \sum_{k=0}^{N-1} x(k)e^{-(2\pi ikn)/N} \quad \text{для } n \text{ от } 0 \text{ до } N-1 \qquad (Д.38)$$

Поскольку частота дискретизации сигнала $x(k)$ равна f_s выборок/с, сигнал включает налагающиеся (или дублирующиеся) компоненты на частотах свыше $f_s/2$. Следовательно, при вычислении значения выражения (Д.38) достаточно ограничиться частотами до $f_s/2$. Отметим, что формула (Д.38) аналогична формуле (Д.23), если положить $z = e^{(2\pi in)/N}$ для последовательности длиной N выборок.

Д.4. Фильтры с конечной импульсной характеристикой

На настоящий момент наиболее распространенный тип цифровых фильтров — это фильтры с конечной импульсной характеристикой (КИХ), имеющие, как понятно из названия, импульсную характеристику конечной длительности. Данные фильтры не имеют весовых коэффициентов обратной связи (см. рис. Д.4); следовательно, можно сделать вывод о их безусловной устойчивости. Выход фильтра с конечной импульсной характеристикой, приведенного на рис. Д.8, описывается следующим выражением.

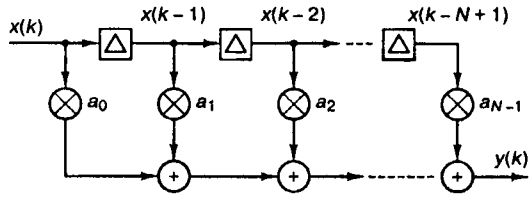


Рис. Д.8. Цифровой фильтр с конечной импульсной характеристикой

$$y(k) = a_0x(k) + a_1x(k-1) + a_2x(k-2) + a_3x(k-3) + \ldots + a_{N-1}x(k-N+1) = \sum_{n=0}^{N-1} a_n x(k-n) \qquad (Д.39)$$

Таким образом, передаточная функция фильтра имеет только нули и не имеет полюсов.

$$\begin{aligned} H(z) &= a_0 + a_1z^{-1} + a_2z^{-2} + a_3z^{-3} + \ldots + a_{N-1}z^{-N+1} = \\ &= a_0(1-\alpha_1z^{-1})(1-\alpha_2z^{-1})(1-\alpha_3z^{-1})\ldots(1-\alpha_Nz^{-1}) \end{aligned} \qquad (Д.40)$$

По сути, работа фильтра с конечной импульсной характеристикой — это вычисление текущего среднего, когда выход — это взвешенное среднее N последних входных выборок. Таким образом, фильтры данного типа часто называются *фильтрами скользящего среднего* (moving average filter). Кроме того, их еще называют *линиями задержки с отводами* (tapped delay line) и *трансверсальными фильтрами* (transversal filter).

Д.4.1. Структура фильтра с конечной импульсной характеристикой

В настоящее время цифровые фильтры с конечной импульсной характеристикой разрабатываются с использованием современного программного обеспечения, такого как SystemView [1]. При этом в основе разработки лежит график амплитудной характеристики, на котором указываются допустимые отклонения и пользовательские требования (рис. Д.9). Затем используются классические методы разработки фильтров, такие как метод Паркса-Мак-Леллана (Parks, McLellan), замена Ремеза (Remez Exchange), окно Кайзера и др. [4], в результате чего создается фильтр с подходящей частотной характеристикой, имеющей минимальное число весовых коэффициентов. Если не оговорено противное, большинство фильтров с конечной импульсной характеристикой разрабатывается в расчете на линейное изменение фазы или постоянную групповую задержку (что соответствует симметричной импульсной характеристике).

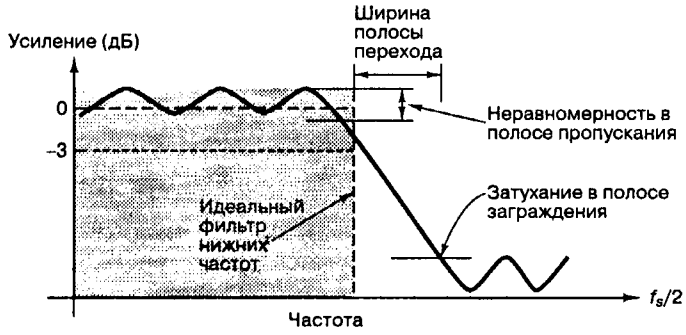


Рис. Д.9. Типичная амплитудная характеристика фильтра нижних частот. Чем строже требования к затуханию в полосе заграждения и полосе перехода и чем ниже допустимая неравномерность в полосе пропускания, тем больше требуется весовых коэффициентов

На рис. Д.10 показаны импульсная и частотная характеристики цифрового фильтра со следующими параметрами: частота среза = 1000 Гц, затухание в полосе заграждения = 20 дБ, неравномерность в полосе пропускания = 3 дБ, полоса перехода = 500 Гц, частота дискретизации, f_s = 10 000 Гц. Если нужен фильтр с более строгими требованиями к частотной характеристике (например, нужно более сильное затухание в полосе заграждения), то скорее всего на стадии проектирования фильтра с конечной импульсной характеристикой обнаружится, что требуется больше весовых коэффициентов [4].

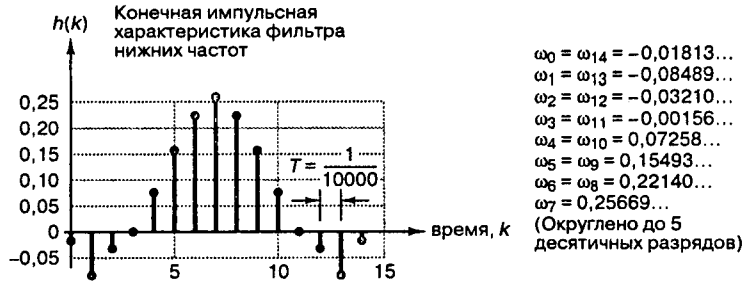
Д.4.2. Дифференциатор с конечной импульсной характеристикой

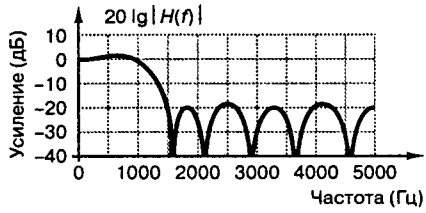
Рассмотрим простой цифровой *дифференциатор*, показанный на рис. Д.11. После изучения выхода для входных синусоид с высокой и низкой частотами, интуитивно можно предположить, что данный дифференциатор — это фильтр верхних частот. Выходная последовательность данного фильтра описывается следующим выражением.

$$y(k) = [x(k) - x(k-1)] \qquad (Д.41)$$

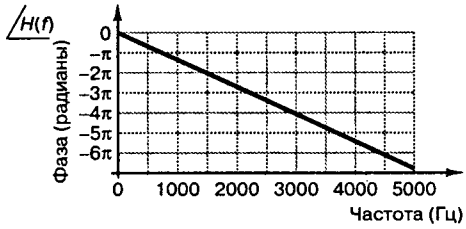
Применение z-преобразования к обеим частям формулы (Д.41) приводит к следующему результату.

$$Y(z) = [X(z) - X(z)z^{-1}] \qquad (Д.42)$$





Логарифмическая амплитудная характеристика



Фазовая характеристика

Рис. Д.10. Импульсная характеристика $h(n) = w_n$ и частотная характеристика $H(f)$ фильтра нижних частот с 15 весовыми коэффициентами; частота дискретизации = 10 000 Гц, частота среза = 1000 Гц

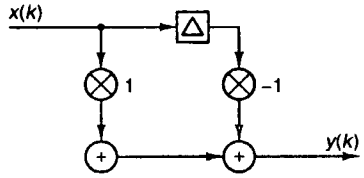


Рис. Д.11. Дифференциатор/фильтр верхних частот

Следовательно, передаточная функция имеет следующий вид.

$$\frac{Y(z)}{X(z)} = (1 - z^{-1}) \qquad (Д.43)$$

На рис. Д.12 показано, почему данная схема действует как фильтр верхних частот. По сути, выход фильтра — это разность двух последних выборок. Если разность последовательных выборок мала (как для случая низкой частоты), выход будет небольшим. Если разность велика (как для высоких частот), выход будет большим. Если на вход подать сигнал постоянного тока, то выходная амплитуда будет нулевой, т.е. будет

происходить бесконечное затухание. Частотную характеристику также можно найти как Фурье-образ импульсной характеристики.

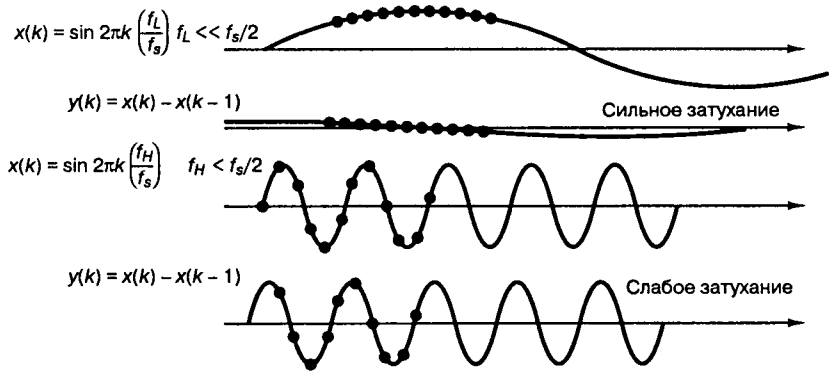


Рис. Д.12. Цифровой дифференциатор, действующий как фильтр верхних частот

Если весовые коэффициенты фильтра изменить с $\{1, -1\}$ на $\{1/T, -1/T\}$, где частота дискретизации $f_s = 1/T$, то для входных низкочастотных сигналов выход, $y(k)$, — это приблизительно дифференциал входа.

$$y(k) \approx \frac{x(k) - x(k-1)}{T} \approx \frac{dx(t)}{dt} \quad \text{и} \quad \frac{Y(z)}{X(z)} = \frac{1}{T}(1 - z^{-1}) \tag{Д.44}$$

Д.5. Фильтры с бесконечной импульсной характеристикой

Фильтры с бесконечной импульсной характеристикой (infinite impulse response — IIR, БИХ) обычно создаются из аналоговых прототипов с использованием отображения из *s*-плоскости в *z*-плоскость. Как понятно из названия, импульсная характеристика таких фильтров (предполагая арифметику бесконечной точности) может иметь бесконечную длительность. Данные фильтры имеют весовые коэффициенты и прямой, и обратной связи, подобно тому, как показано на рис. Д.4. Вследствие рекурсивной природы поточного графа, данные фильтры могут иметь весьма длительные импульсные характеристики (до нескольких весовых коэффициентов). Следовательно, фильтры с бесконечной импульсной характеристикой могут создаваться с меньшим числом весовых коэффициентов, чем фильтры с конечной импульсной характеристикой при аналогичных функциональных амплитудных характеристиках. В общем случае в цифровых фильтрах с бесконечной импульсной характеристикой фаза изменяется нелинейно.

Д.5.1. Оператор левосторонней разности

Уравнение (Д.44) позволяет связать переменную преобразования Лапласа s (непрерывное время) и переменную z-преобразования z (дискретное время). Известно, что при преобразовании Лапласа дифференцирование по времени (d/dt) переходит в умножение на переменную s.

$$y(t) = \frac{dx(t)}{dt} \quad \Rightarrow Y(s) = sX(s) \tag{Д.45}$$

Возьмем, например, следующую характеристику фильтра Баттерворта.

$$H(s)=\frac{1}{s^2+\sqrt{2}s+1} \tag{Д.46}$$

Данную аналоговую схему (фильтр нижних частот) можно аппроксимировать дискретно, подставив приближение

$$s\approx\frac{1}{T}(1-z^{-1}) \tag{Д.47}$$

в уравнение (Д.46). Это дает следующее уравнение в z-области.

$$\begin{aligned}H(z)=H(s)\big|_{s=\frac{1}{T}(1-z^{-1})}&=\frac{1}{\frac{1}{T^2}(1-z^{-1})^2+\sqrt{2}\frac{1}{T}(1-z^{-1})+1}\\&=\frac{T^2}{(1-2z^{-1}+z^{-2})+\sqrt{2}T(1-z^{-1})+T^2}\\&=\frac{T^2}{z^{-2}-(\sqrt{2}T+2)z^{-1}+(1+\sqrt{2}T+T^2)}\end{aligned} \tag{Д.48}$$

При низких частотах, когда приближение (Д.47) является "хорошим", данное преобразование может давать "разумный" цифровой эквивалент аналогового фильтра нижних частот. (Уравнение (Д.47) иногда называется "оператором левосторонней разности".) К сожалению, данное отображение является очень плохим при высоких частотах, а следовательно, оно не может использоваться при создании фильтров верхних частот. Таким образом, на практике оно применяется редко.

Д.5.2. Использование билинейного преобразования для создания фильтров с бесконечной импульсной характеристикой

Билинейное преобразование получается при замене s следующим приближением.

$$s\approx\frac{2}{T}\frac{(1-z^{-1})}{(1+z^{-1})} \tag{Д.49}$$

Данная подстановка приводит к отображению, сохраняющему устойчивость аналогового прототипа и дающему фильтры, значительно лучшие по своим характеристикам, чем в предыдущем случае (уравнение (Д.47)) [2]. В SystemView [1] билинейное преобразование используется для создания цифровых фильтров из стандартных аналоговых прототипов, таких как фильтры Баттерворта, эллиптические фильтры и фильтры Чебышева. Отметим, что билинейное преобразование всегда дает фильтр, имеющий нули и полюсы; следовательно, данные фильтры имеют бесконечную импульсную характеристику (БИХ).

Д.5.3. Интегратор с бесконечной импульсной характеристикой

Цифровой *интегратор* — это, по сути, БИХ-фильтр с одним весовым коэффициентом.

$$y(k)=x(k)+y(k-1)=\sum_{i=0}^{k}x(i) \tag{Д.50}$$

В z-области передаточная функция дискретного интегратора получается из соотношения

$$Y(z) = X(z) + z^{-1}Y(z), \tag{Д.51}$$

которое дает следующее.

$$\frac{Y(z)}{X(z)} = \frac{1}{1 - z^{-1}} \tag{Д.52}$$

Реализация простого цифрового интегратора и графическое представление связи с интегрированием по непрерывному времени показаны на рис. Д.13.

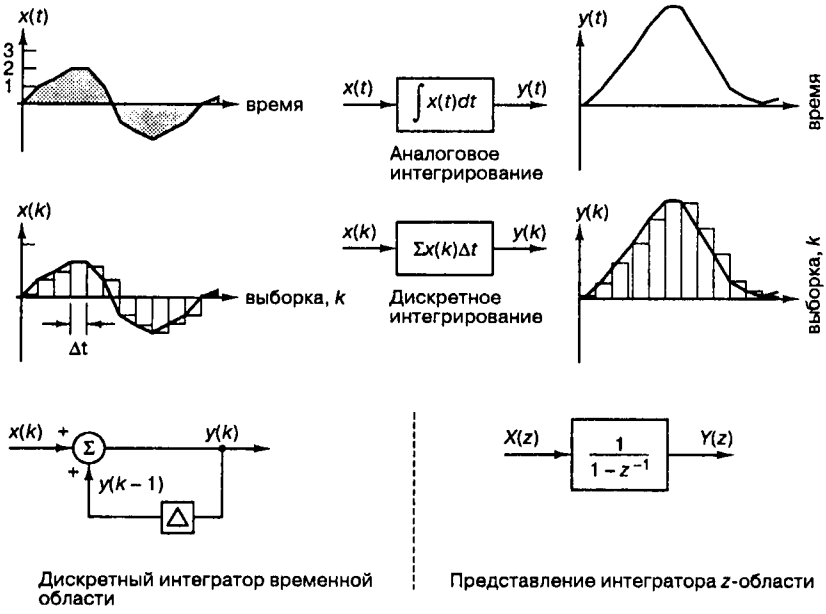


Рис. Д.13. Однополюсный фильтр, действующий как интегратор. В контур обратной связи часто вводится весовой коэффициент, немного меньший 1, который обеспечивает "забывание" интегратора

Если в контур обратной связи вводится весовой коэффициент, меньший 1 (скажем, 0,99), интегратор часто называют *квазиинтегратором* (leaky integrator). При рассмотрении в частотной области, характеристики (квази)интегратора и фильтра нижних частот не отличаются.

Литература

1. *SystemView DSP Communications Software*. Elanix, Westlake Village, CA, 2000.
2. Porat. B. *A Course in Digital Signal Processing*. John Wiley & Cons, 1997.
3. Moon T. K., Stirling W. C. *Mathematical Methods and Algorithms for Digital Signal Processing*. Prentice Hall, 2000.
4. Stewart R. W. *The DSPedia: A Multimedia Resource for DSP*. BlueBox Multimedia, UK, 2000.

ПРИЛОЖЕНИЕ Е

Перечень символов

a_{ij}	Коэффициент j-й базисной функции
a_i	Сигнальный компонент на выходе j-го коррелятора
A	Максимальная амплитуда сигнала
A_e	Эффективная площадь поверхности (антенны)
B_L	Односторонняя ширина полосы контура
c	Скорость света $\approx 3 \times 10^8$ м/с
C	Пропускная способность канала
C	Электрическая емкость
C/N_0	Отношение средней мощности несущей к спектральной плотности мощности шума
d	Расстояние
d_0	Эталонное расстояние
d_f	Просвет
$d_{\min}$	Минимальное расстояние
D	Время задержки (сообщения)
D	Избыточность языка
D	Преобразование дешифрования
e	Основание натурального логарифма $\approx 2{,}7183$
$\mathbf{e}$	Вектор ошибочной комбинации
$e(t)$	Сигнал ошибки
$\mathbf{e}(X)$	Полином ошибочной комбинации
E	Преобразование шифрования
E_x	Энергия сигнала $x(t)$
$\mathbf{E}\{X\}$	Математическое ожидание случайной переменной X

EIRP	Эффективная излученная мощность относительно изотропного источника
E_b/J_0	Отношение энергии бита к спектральной плотности мощности станции-постановщика помех
E_b/N_0	Отношение энергии бита к спектральной плотности мощности шума
E_c/N_0	Отношение энергии канального символа к спектральной плотности мощности шума
f	Частота (герц)
f_c	Частота несущей волны
f_m	Максимальная частота
f_0	Ширина полосы когерентности
f_d	Доплеровское расширение полосы частот
f_l	Нижняя частота среза фильтра
f_u	Верхняя частота среза фильтра
F	Шум-фактор
$\mathfrak{F}\{x\}$	Фурье-преобразование функции $x(t)$
$\mathfrak{F}^{-1}\{x\}$	Обратное Фурье-преобразование функции $X(f)$
F	Поле
F^*	Конечное поле
$g(t)$	Псевдослучайная кодовая функция
$\mathbf{g}(X)$	Полиномиальный генератор (для циклического кода)
G	Коэффициент усиления антенны
G	Эффективность кодирования
$\mathbf{G}$	Матрица генератора (для линейных блочных кодов)
G	Нормированный объем информации
G_p	Коэффициент расширения спектра
$G_x(f)$	Спектральная плотность мощности сигнала $x(t)$
$h(t)$	Импульсная характеристика сети
$h_c(t)$	Импульсная характеристика канала
$\mathbf{H}$	Матрица проверки четности для кода
H_i	i-я гипотеза
$\mathbf{H}_k$	Матрица Адамара
$H(f)$	Частотная передаточная функция сети
$H_0(f)$	Оптимальная частотная передаточная функция
$H(X)$	Энтропия информационного источника X
$H(X\|Y)$	Условная энтропия (энтропия X при условии Y)
$i(t)$	Форма кривой электрического тока
I	Электрический ток
$I_0(x)$	Модифицированная функция Бесселя первого рода нулевого порядка
$I(X)$	Самоинформация информационного источника X
J	Средняя принятая мощность станции-постановщика помех
J_0	Спектральная плотность мощности станции-постановщика помех
J/S	Отношение средней принятой мощности станции-постановщика помех к средней мощности сигнала
k	Число бит в M-арном множестве сигналов

k/n	Степень кодирования (отношение длины исходного блока информации к длине его кодированного представления)
K	Длина кодового ограничения сверточного кодера
K	Ключ, определяемый схемой шифрования или преобразованием дешифрования
l	Число бит квантования
$l(d_k)$	Правдоподобие информационного бита d_k
L	Длина упреждения в сверточном декодировании с обратной связью
L	Число ответвляющихся слов в последовательности
L	Число уровней квантования
$L(d_k)$	Логарифмическое отношение правдоподобий информационного бита d_k
L_e	Внешнее логарифмическое отношение правдоподобий
L_s	Потери в свободном пространстве
L_o	Другие потери
L_p	Потери в тракте связи
L_o	Время наблюдения
L_c	Время наблюдения для ISI, введенной каналом
L_{CISI}	Время наблюдения для контролируемой ISI
L_c	Канальное логарифмическое отношение правдоподобий
$\mathbf{m}$	Вектор сообщения
$\mathbf{m}(X)$	Полином сообщения
m_i	Бит данных
M	Резерв
M	Размер множества сигналов
(n, k)	Маркировка кода, где n — общее число бит, а k — число бит в кодовом слове
$\bar{n}$	Среднее число бит на символ
n_0	Переменная случайного шума на выходе коррелятора в момент $t = T$
$n(t)$	Гауссов процесс шума
N	Мощность шума
N	Расстояние единственности
N_0	Уровень односторонней спектральной плотности мощности белого шума
NSR	Отношение средней мощности шума к средней мощности сигнала
p_c	Вероятность ошибки в канальном символе
p_i	Бит четности
$p(t)$	Мгновенная мощность
$p(x)$	Функция плотности вероятности непрерывной случайной переменной
$p(x\|y)$	Функция плотности вероятности x при условии y
$\mathbf{P}$	Массив четности
P_B	Вероятность битовой ошибки
P_E	Вероятность символьной ошибки
P_{FA}	Вероятность ложной тревоги
P_m	Вероятность несоответствия
P_M	Вероятность блочной ошибки или ошибки сообщения
P_{nd}	Вероятность необнаруженной ошибки

P_r/N_0	Отношение средней принятой мощности сигнала к спектральной плотности мощности шума
$\mathbf{p}(X)$	Полином остатка
$P(X)$	Вероятность дискретной случайной переменной
P_x	Средняя мощность сигнала $x(t)$
q	Шаг квантования
$\mathbf{q}(X)$	Полином частного
$Q(x)$	Гауссов интеграл ошибок
r	Коэффициент сглаживания фильтра
r	Истинная интенсивность языка
r'	Абсолютная интенсивность языка
$r(t)$	Принятый сигнал
R	Скорость передачи данных (бит/с)
$R(\Delta f)$	Корреляционная функция разнесения частоты
$R(\Delta t)$	Корреляционная функция разнесения времени
R_c	Скорость передачи кодовых или канальных бит (кодовых бит/с)
R_{ch}	Скорость передачи элементарных сигналов (элементарных сигналов/с)
R_s	Скорость передачи символов (символов/с)
$R_x(\tau)$	Автокорреляционная функция сигнала $x(t)$
$\Re$	Электрическое сопротивление
$s(t)$	Сигнал
$\hat{s}(t)$	Оценка сигнала
$S(\nu)$	Доплеровская спектральная плотность мощности
$S(\tau)$	Профиль интенсивности при многолучевом распространении
$\mathbf{s}$	Вектор сигнала
$\operatorname{sgn} x$	Знаковая функция x
S_k	Состояние в момент k
S	Мощность сигнала
$\mathbf{S}$	Вектор синдрома
SJB	Отношение средней мощности сигнала к средней мощности помех
SNR	Отношение средней мощности сигнала к средней мощности шума
S/N	Отношение мощности сигнала к мощности шума
$S(f)$	Фурье-образ сигнала $s(t)$
$\mathbf{S}(X)$	Полином синдрома
t	Число исправимых ошибок в коде коррекции ошибок
t	Независимая переменная времени
t_0	Временная задержка
t_{ij}	Объем информации, переданной от i к j
T	Длительность импульса
T	Длительность символа
$T(D)$	Передаточная функция или производящая функция сверточного кода
T_{ch}	Время передачи на одной частоте
T_{hop}	Длительность перехода
T_s	Интервал дискретизации
$T°$	Температура

$T°_A$	Температура антенны
$T°_L$	Эффективная температура линии связи
T_m	Задержка многолучевого распространения (максимальная)
T_0	Время когерентности
$T°_R$	Эффективная температура приемника
$T°_S$	Температура системы
T_{acq}	Время синхронизации
u_i	Кодовый символ
$u(t)$	Единичная ступенчатая функция
$\mathbf{U}$	Вектор кодового слова
$\mathbf{U}(X)$	Полином кодового слова
v	Относительная скорость
$v(t)$	Форма кривой электрического напряжения
$\text{var}(X)$	Дисперсия случайной переменной X
V	Скорость
$w(t)$	Сигнал станции-постановщика помех
W	Ширина полосы
W_f	Ширина полосы фильтра
W_{DSB}	Двусторонняя ширина полосы
W_{N}	Ширина полосы шумового эквивалента
W_{ss}	Ширина полосы расширения спектра
$z(t)$	Выход согласованного фильтра или коррелятора
α_k	Прямая метрика состояния в момент k
β_k	Обратная метрика состояния в момент k
Γ	Отношение SNR, усредненное по подъемам и спадам замирания
Γ_a	Метрика состояния для состояния a
γ	Порог (принятия) решения
γ_0	Оптимальный порог
$\gamma_{\mathbf{U}^{(m)}}$	Правдоподобие кодового слова $\mathbf{U}^{(m)}$
δ	Фракционный уход частоты за день
δ_k	Метрика ветви в момент k
δ_{mn}	Дельта-функция Кронекера
$\delta(t)$	Импульсная функция (дельта-функция Дирака)
ε	Ошибка
ζ	Характеристика демпфирования контура (контур второго порядка)
η	Эффективность антенны
$\theta(t)$	Переменная фаза
$\Theta(\omega)$	Фурье-образ $\theta(t)$
κ	Постоянная Больцмана, $1{,}38 \times 10^{-23}$ Дж/К
$\Lambda(d_k)$	Отношение правдоподобий бита данных d_k
λ	Совместная вероятность
λ	Длина волны
λ	Скорость поступления пакетов
π	Число “пи”, 3,14159
ρ	Часть полосы частот, подвергающаяся воздействию помех

ρ	Часть времени, в течение которого "включены" помехи
ρ	Нормированное отношение сигнал/шум контура
ρ	Нормированная пропускная способность сообщений
ρ	Число исправимых стираний в коде коррекции ошибок
ρ	Коэффициент временной корреляции
ρ_0	Значение ρ, максимизирующее вероятность битовой ошибки (наихудший случай, возможный при помехах)
σ_τ	Среднеквадратическое распространение задержки
σ_X	Среднеквадратическое отклонение случайной переменной X
σ_X^2	Дисперсия случайной переменной X
τ	Ширина импульса
τ	Сдвиг во времени (независимая переменная автокорреляционной функции)
$\psi_j(t)$	Базисная функция
$\Psi_x(t)$	Спектральная плотность энергии сигнала $x(t)$
ω	Угловая частота (радиан в секунду)

Предметный указатель

Л

М

Н

О

П

Р

С

Э

Научно-популярное издание

Бернард Скляр

Цифровая связь. Теоретические основы и практическое применение, 2-е издание

Литературный редактор *Е.Д. Давидян*
Верстка *О.В. Линник*
Художественный редактор *М.А. Смолина*
Обложка *С.А. Чернокозинский*
Корректоры *З.В. Александрова, Л.А. Гордиенко, Л.В. Коровкина, О.В. Мишутина*

Издательский дом "Вильямс".
101509, Москва, ул. Лесная, д. 43, стр. 1.
Изд. лиц. ЛР № 090230 от 23.06.99
Госкомитета РФ по печати.

Подписано в печать 06.12.2002. Формат 70×100/16.
Гарнитура Times. Печать офсетная.
Усл. печ. л. 89,0. Уч.-изд. л. 73,5.
Тираж 3000 экз. Заказ № 1938.

Отпечатано с диапозитивов в ФГУП "Печатный двор"
Министерства РФ по делам печати,
телерадиовещания и средств массовых коммуникаций.
197110, Санкт-Петербург, Чкаловский пр., 15.