Speech Emotion Recognition using Attention-based LSTM-Network with Residual Connection

D.V. Krasnoproshin, **M.I. Vashkevich**
vashkevich@bsuir.by

Белорусский государственный университет
информатики и радиоэлектроники
Кафедра электронных вычислительных средств
Минск, Беларусь

28-я конференция DSPA'2026
Цифровая обработка сигналов и её применение
Москва, Россия

25 марта, 2026

Содержание

Введение

Введение: распознавание эмоций в речи

Актуальность задачи

Распознавание эмоций в речи (SER) — играет важную роль в различных приложениях:

- Человеко-компьютерное взаимодействие
- Здравоохранение (диагностика, мониторинг)
- Обслуживание клиентов в колл-центрах

Тенденции и вызовы

- Исследования смещаются к **мультимодальным** системам (аудио + видео)
- Многие реальные приложения **ограничены только аудио-модальностью**:
  - Голосовые помощники (Siri, Alexa, Алиса)
  - Автоматический анализ удовлетворенности клиентов (колл-центры)
  - Системы контроля состояния водителя

Мотивация и цель работы

Проблема существующих подходов

- Предобученные модели (PANNs, Wav2Vec) достигают высокой точности, но имеют **значительную вычислительную сложность**
- Ограничивают применение в реальных условиях:
  - на периферийных устройствах (*edge devices*) с низкой производительностью;
  - в системы реального времени (в голосовых ассистентах с локальной обработкой);

Цель работы

Разработать **эффективную систему SER**, достигающую баланса между:

- Вычислительной сложностью
- Точностью распознавания

Предлагаемый подход

Легковесная LSTM-архитектура с улучшенным моделированием временных зависимостей за счет:

- Механизма мягкого внимания (*soft attention*)
- Комбинирования нескольких LSTM-ячеек с использованием остаточных связей (*residual connections*)

Извлечение акустических признаков

Извлечение акустических признаков

Речевой сигнал содержит богатую информацию об эмоциях, но для обработки моделями глубокого обучения необходимо преобразовать её в низкоразмерное представление.

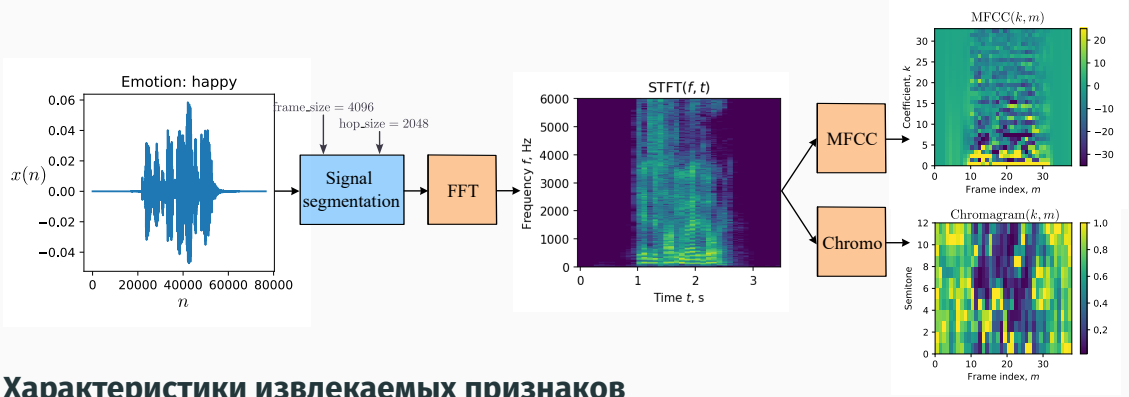
Мел-частотные кепстральные коэффициенты (MFCC)

- Основаны на психоакустических принципах
- Аппроксимируют нелинейное частотное разрешение человеческого слуха

Хромаграмма

- Проецирует спектральное содержание на 12 классов высоты тона
- Кодируют гармоническую структуру
- Захватывают просодические признаки, важные для распознавания эмоций

Процесс извлечения признаков



Характеристики извлекаемых признаков

- **MFCC:** 34 коэффициента на фрейм
- **Хромаграммы:** 12 коэффициентов на фрейм
- **Итого:** 46-мерный вектор признаков на фрейм

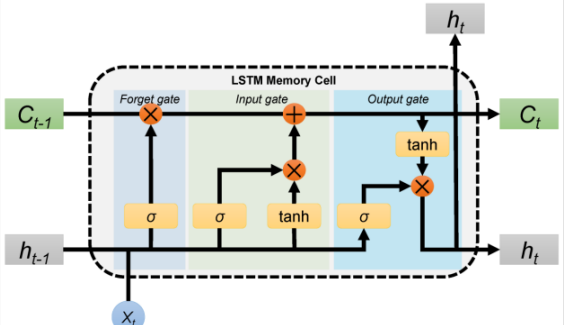
LSTM-сети с механизмом внимания для SER

LSTM-сети с механизмом внимания для распознавания эмоций

Базовая модель

Данная работа основана на архитектуре LSTM с механизмом мягкого внимания, предложенной в работе Mirsamadi et al. (2017).

LSTM (*Long Short-Term Memory*) – тип рекуррентной нейронной сети, способной обрабатывать последовательные данные и захватывать долгосрочные зависимости.



Применение для SER

Для классификации эмоций на уровне высказывания последовательность скрытых состояний $\mathbf{h}_t$ агрегируется с использованием механизма внимания и передается в полносвязный слой с функцией активации $\mathrm{softmax}$.

Уравнения LSTM-ячейки

Математическая модель

LSTM-ячейка обрабатывает последовательные входы через управляемые взаимодействия, описываемые следующими уравнениями:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{ii}\mathbf{x}_t + \mathbf{b}_{ii} + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_{hi}) \qquad \mathbf{o}_t = \sigma(\mathbf{W}_{io}\mathbf{x}_t + \mathbf{b}_{io} + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_{ho})$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{if}\mathbf{x}_t + \mathbf{b}_{if} + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_{hf}) \qquad \mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$

$$\mathbf{g}_t = \tanh(\mathbf{W}_{ig}\mathbf{x}_t + \mathbf{b}_{ig} + \mathbf{W}_{hg}\mathbf{h}_{t-1} + \mathbf{b}_{hg}) \qquad \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$

Обозначения

- $\mathbf{x}_t \in \mathbb{R}^d$ — вектор признаков на временном шаге t
- $\mathbf{h}_t$, $\mathbf{c}_t$ — скрытое состояние и состояние ячейки
- $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{g}_t$, $\mathbf{o}_t$ — входной, забывающий, кандидат на ячейку и выходной затворы

Механизм мягкого внимания

Вектор контекста

Аггрегация скрытых состояний $\mathbf{h}_t$ с использованием механизма внимания:

$$\mathbf{h}_{\text{context}} = \sum_{t=0}^{T-1} \alpha_t \mathbf{h}_t, \tag{1}$$

где α_t — вес внимания для временного шага t, представляющий относительную важность соответствующего скрытого состояния $\mathbf{h}_t$ в представлении высказывания.

Вычисление коэффициентов внимания

$$\alpha_t = \text{softmax}(e_t) = \frac{\exp(e_t)}{\sum_{t=0}^{T-1} \exp(e_t)}, \tag{2}$$

где $e_t = \mathbf{u}^T \mathbf{h}_t$ — оценка внимания, а $\mathbf{u}$ — вектор внимания (обучаемый параметр).

Базовая архитектура LSTM-SA

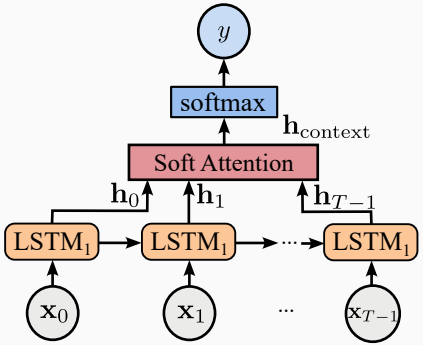
Обозначение архитектуры

Описанную архитектуру с размерностью скрытого состояния Y обозначим как LSTM-SA-hY, где ”SA” обозначает soft attention (мягкое внимание).

Характеристики базовой архитектуры

- Один слой LSTM с механизмом мягкого внимания
- Контекстный вектор $\mathbf{h}_{\text{context}}$ передается в полносвязный слой
- Функция активации softmax для классификации эмоций
- Обучаемые параметры: веса LSTM и вектор внимания $\mathbf{u}$

Базовая модель LSTM-SA



Архитектура ResLSTM-SA с остаточными связями

Архитектура ResLSTM-SA с остаточными связями

Мотивация и основная идея

В работе предлагается модификация базовой архитектуры LSTM-SA путем включения дополнительного слоя LSTM с остаточными связями (*residual connection*). Этот слой обогащает временное представление входных признаков перед их обработкой основной LSTM-сетью на основе механизма внимания, тем самым повышая способность модели улавливать долгосрочные зависимости.

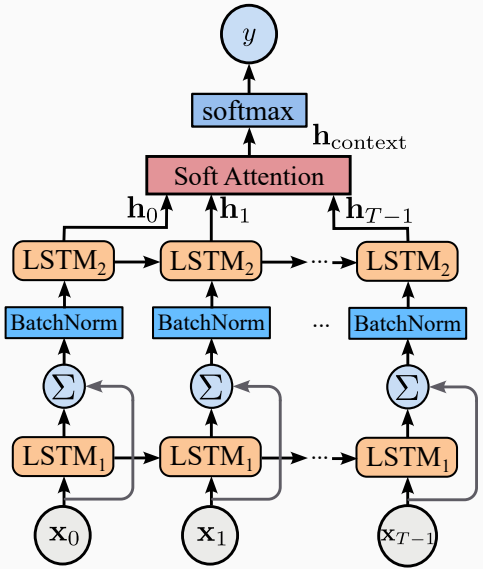
Обозначение архитектуры

Предлагаемую архитектуру обозначим как ResLSTM-SA, где ”Res” относится к остаточным связям. Для указания моделей с размерностью скрытого состояния Y используем обозначение ResLSTM-SA-hY (например, ResLSTM-SA-h64 для $Y = 64$).

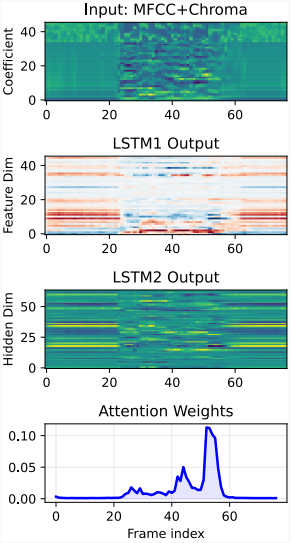
Модель ResLSTM-SA

Математическое представление

- Вход: $\mathbf{X} \in \mathbb{R}^{T \times d}$, где $d = 46$
- Первый слой LSTM_1 с выходной размерностью d: $\mathbf{H}_1 = \mathrm{LSTM}_1(\mathbf{X})$
- Обогащенное представление, вычисление суммы с остаточной связью: $\mathbf{X}' = \mathbf{X} + \mathbf{H}_1$
- Второй слой LSTM с вниманием: LSTM_2 с размерностью Y



Визуализация работы обученной ResLSTM-SA-h64



Анализ визуализации

- LSTM_1 генерирует контекстно обогащенные представления, которые добавляются к исходному входу через остаточные связи
- Механизм внимания фокусирует модель на эмоционально выразительных сегментах высказывания
- Сохранение входной размерности в LSTM_1 позволяет выполнять аддитивное слияние исходных признаков с их контекстно обогащенными аналогами

Преимущества ResLSTM-SA по сравнению с LSTM-SA

Улучшенное временное моделирование

- Первый слой LSTM захватывает локальные временные зависимости
- Остаточные связи сохраняют исходную информацию
- Второй слой LSTM с вниманием фокусируется на извлечении эмоциональных паттернов
- Комбинированное представление более информативно

Упрощение обучения

- Остаточные связи облегчают градиентный поток
- Предотвращают затухание градиентов
- Улучшают сходимость при обучении
- Позволяют эффективно обучать более глубокие архитектуры

Эксперименты и результаты

Набор данных RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

- 1,440 высококачественных аудиофайлов (16-бит, 48 кГц)
- 24 профессиональных актера (12 мужчин, 12 женщин)
- 2 фразы: ”Kids are talking by the door” и ”Dogs are sitting by the door”

Эмоции (8 классов)

- Нейтральность (*neutral*)
- Спокойствие (*calm*)
- Счастье (*happy*)
- Грусть (*sad*)
- Злость (*angry*)
- Испуганность (*fearful*)
- Удивление (*surprised*)
- Отвращение (*disgusted*)

Описание эксперимента

Основная цель эксперимента

Сравнительная оценка архитектур LSTM-SA и ResLSTM-SA для распознавания эмоций в речи с систематическим анализом:

- Влияния остаточных связей
- Влияния емкости модели (размерность скрытого состояния)

Процесс обучения

- Оптимизатор: Adam; планировщик скорости обучения – косинусный отжиг
- Функция потерь – категориальная перекрестная энтропия

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\,\log\bigl(\hat{y}_{n,c}\bigr), \tag{3}$$

где N — размер батча, C — число классов эмоций, $y_{n,c} \in \{0, 1\}$ — истинная метка, $\hat{y}_{n,c} \in (0, 1]$ — предсказанная вероятность класса.

Метрика оценки и валидация

Невзвешенная средняя полнота (UAR)

$$\mathrm{UAR} = \frac{1}{C}\sum_{c=1}^{C}\frac{A_{c,c}}{\sum_{i=1}^{C} A_{c,i}}, \tag{4}$$

где $A \in \mathbb{N}^{C\times C}$ — матрица ошибок, $A_{c,i}$ — количество образцов класса c, предсказанных как класс i.

Схема валидации

- Перекрестная проверка по 5 блокам
- Дикторонезависимое разделение на блоки (speaker-independent)

Данная схема валидации обеспечивает воспроизводимое сравнение с предыдущими исследованиями

Оптимизация гиперпараметров с помощью Optuna

Байесовская оптимизация гиперпараметров

- Алгоритм Tree-Structured Parzen Estimator (TPE)
- Динамическое построение пространства поиска во время оптимизации
- Число попыток (*trials*): 100

Оптимизируемые гиперпараметры

- Скорость обучения: $\eta \in [3 \cdot 10^{-5}, 2 \cdot 10^{-4}]$ (логарифмическая выборка)
- Затухание весов: $\lambda \in [2 \cdot 10^{-5}, 2 \cdot 10^{-2}]$ (логарифмическая выборка)
- Dropout в полносвязных слоях: $p_{drop} \in [0.1, 0.5]$ (равномерная выборка)
- Число циклов в косинусном отжиге: $T_0 \in \{1, 2, 3, 5, 10\}$
- Размер батча: $\text{batch\_size} \in \{8, 16, 32, 64\}$

Оценка производительности

Методология оценки

- Использование оптимальных гиперпараметров, найденных Optuna
- 10 независимых запусков обучения для каждой конфигурации
- Каждый запуск инициализирован уникальным случайным сидом (*seed*)

Цель методологии

- Снижение дисперсии, вызванной стохастической инициализацией весов
- Предотвращение переобучения на случайный сид, использованный при оптимизации гиперпараметров
- Получение статистически значимой оценки обобщающей способности модели

Результаты экспериментов с LSTM-сетями

Модель	# Параметров	UAR (среднее $\pm$ ст. откл.)	UAR (макс.)
LSTM-SA-h32	10.6 k	0.5352 $\pm$ 0.0123	0.5547
LSTM-SA-h64	28.3 k	0.5751 $\pm$ 0.0108	0.5996
LSTM-SA-h128	91.6 k	0.5895 $\pm$ 0.0076	0.6022
ResLSTM-SA-h32	28.0 k	0.6130 $\pm$ 0.0111	0.6315
ResLSTM-SA-h64	46.8 k	0.6232 $\pm$ 0.0119	**0.6517**
ResLSTM-SA-h128	108.9 k	0.6107 $\pm$ 0.0134	0.6348

Основные наблюдения

- ResLSTM-SA постоянно превосходит базовую LSTM-SA
- Улучшение до 7.7 процентных пунктов по UAR
- ResLSTM-SA-h32 (28.0k параметров) превосходит LSTM-SA-h128 (91.6k параметров)
- Оптимальная модель: ResLSTM-SA-h64 (46.8k параметров, UAR=0.6517)

Анализ результатов ResLSTM-SA

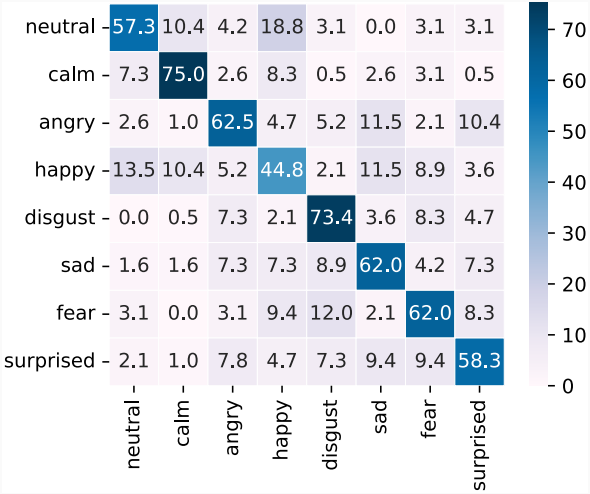
Эффективность остаточных связей

- ResLSTM-SA-h32 (28.0k параметров) превосходит LSTM-SA-h128 (91.6k параметров)
- Относительное улучшение: 4.0 процентных пункта по UAR
- В 3.3 раза меньше параметров при лучшей производительности

Влияние размера скрытого состояния

- Для LSTM-SA: монотонное улучшение UAR с ростом h (от 0.5352 до 0.5895)
- Для ResLSTM-SA: оптимальное значение при $h = 64$ (0.6232 среднее, 0.6517 макс.)
- При $h = 128$: небольшое снижение UAR, вероятно из-за переобучения

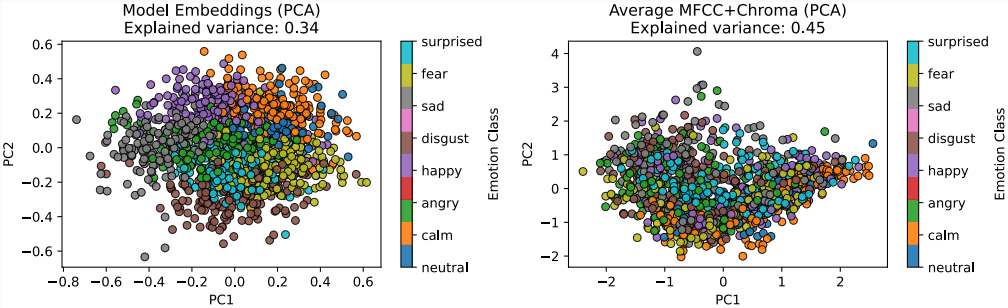
Матрица ошибок для ResLSTM-SA-h64



Анализ матрицы ошибок

- **Счастье:** Наименьшая полнота (44.8%), часто путается с **нейтральность**
- **Нейтральность:** 18.8% ошибочно классифицируются как **счастливый**
- **Систематическая путаница** между этими состояниями происходит по причине перекрывающиеся просодические паттерны в пространстве валентности-возбуждения

Визуализация эмбеддингов (PCA)



- **Исходные акустические признаки (MFCC + Chroma):**
  - Значительное перекрытие между классами эмоций
  - Внутриклассовая дисперсия ↑
- **Эмбеддинги ResLSTM-SA-h64:**
  - Компактные, хорошо разделимые кластеры
  - Внутриклассовая дисперсия ↓
  - Улучшенные межклассовые границы

Сравнение с известными работами

Модель	**# Параметров**	**UAR**
AlexNet embeddings + SVM **Luna-2021**	61.0 M	0.4580
CNN+LSTM **Dissanayake-2020**	-	0.5671
GResNet+S **Zeng-2019**	-	0.5970
Fine-tuned AlexNet **Luna-2021**	61.0 M	0.6167
ResLSTM-SA-h64 [предлагаемая]	**0.05 M**	**0.6517**
Fine-tuned CNN14 **Luna-2021**	81.0 M	0.7658
Fine-tuned xlsr-wav2vec 2.0 **Luna-2022**	317.0 M	0.8182
wav2vec 2.0 с аугментацией **Ibrahim-2024**	317.0 M	0.8229

Анализ сравнения с state-of-the-art

Классические и гибридные модели

- CNN и CNN-RNN модели: UAR в диапазоне 0.56–0.62
- Производительность обычно растет с увеличением сложности модели
- Предлагаемая ResLSTM-SA-h64 превосходит все не-самообучающиеся подходы

Самообучающиеся модели большого масштаба

- PANNs (CNN14) и wav2vec 2.0: Более высокие абсолютные значения UAR
- Требуют предобучения на больших речевых корпусах
- 81–317 миллионов параметров (на 3 порядка больше)
- Значительные вычислительные затраты на обучение и инференс

Выводы

Основные результаты

- Представлена **ResLSTM-SA** — легковесная архитектура для распознавания эмоций
- Интеграция остаточных связей и механизма мягкого внимания
- Эффективный компромисс между точностью и вычислительной эффективностью
- **ResLSTM-SA-h64**: Средний UAR 0.6232 ± 0.0119, максимальный 0.6517
- 46.8k параметров — в 3.3 раза меньше, чем у базовой модели LSTM-SA-h128 при лучшей производительности