Speech enhancement in quasi-periodic noises using improved spectral subtraction based on adaptive sampling

Elias Azarov, Maxim Vashkevich, and Alexander Petrovsky

Belarusian State University of Informatics and Radioelectronics,
Department of Computer Engineering, Minsk, Belarus
{azarov,vashkevich,palex}@bsuir.by

Abstract. The paper presents a speech processing method based on spectral subtraction that is effective for reduction of specific rate-dependent noises. Such noises are produced by a variety of different rotation sources such as turbines and car engines. Applicability of convenient spectral subtraction for such noises is limited since their power spectral density (PSD) is connected with rotation rate and therefore constantly changing. The paper shows that in some cases it is possible to compensate variation of PSD by adaptive sampling rate. The signal can be processed in warped time domain that makes noise parameters more stable and easy to estimate. Stabilization of PSD leads to more accurate evaluation of noise parameters and significantly improves result of noise reduction. For determination of current rotation rate the proposed method can either use external reference signal or the noisy signal itself applying pitch detector to it. Considering that the noise typically consists of deterministic and stochastic components narrow-band and wide-band components of the noise are removed separately. The method is compared to the recently proposed maximum a posteriori method (MAP).

Keywords: Noise reduction, spectral subtraction, time warping

1 Introduction

The paper addresses the problem of cleaning speech signals from time-varying quasi-periodic noises typically generated by rotating machines. The problem is characterized by the following tough points: unsteadiness of the noise that requires fast tracking of the noise parameters; extremely low signal-to-noise ratios (SNRs) that can be below -10 dB, the hybrid structure of the noise that combines deterministic (tonal or narrow-band) with stochastic (non-periodic wide-band components).

Among approaches that have been applied to the problem are spectral subtraction, Weiner filtering, MAP and others [1–5]. Majority of the methods use additional reference information such as engine or vehicle speed in order to find frequency locations of engine harmonic noise components. When rotation rate changes both deterministic and stochastic components shift their frequency, but

speed-dependent filtering is actually applied only to the deterministic part which is typically done by notch filtering. Notch filters are applied to suppress individual harmonics [1, 5] as long as it is not possible to obtain sufficient frequency resolution for such non-stationary periodic components using a filter bank, or short-time Fourier transform (STFT). The stochastic part is processed regardless of the rotation speed information that results in a high engine noise residue in the processed speech.

The paper presents a noise reduction technique based on spectral subtraction. Noise is processed in warped time domain which makes possible to process noise spectrum with high frequency resolution and apply speed-dependent filtering both to deterministic and stochastic noise components.

The method can use external reference information or estimate rotation rate from the noise using a pitch detector assuming that engine speed cannot change very rapidly. The obtained experimental results are rated in terms of objective and subjective values.

2 Method outline

The method can be shortly described by the following steps (figure 1): 1) acquisition of rotation rate using external source or noised signal; 2) warping of the signal accordingly to estimated fundamental frequency; 3) detection of tonal components and narrow-band filtering; 4) estimation of unvoiced regions and estimation of PSD of the noise; 5) spectral subtraction of the noise using gathered noise statistics; 6) inverse signal warping of the processed signal in order to return it into original time domain.

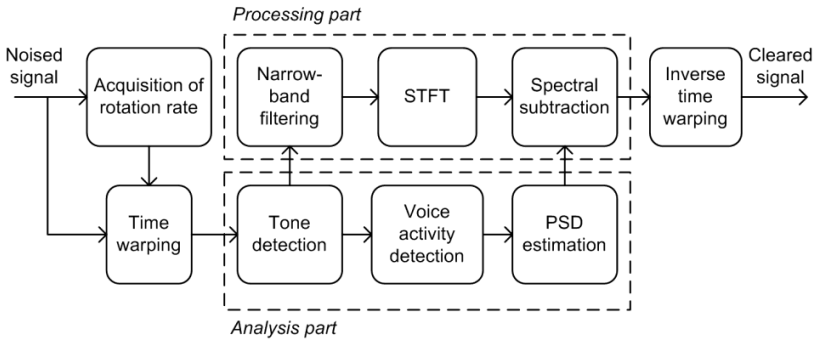

Fig. 1. Signal processing scheme

The steps of the algorithm are detailed and illustrated below.

2.1 Acquisition of rotation rate

If no reference signal available rotation rate is estimated using fundamental frequency of the noised signal. Fundamental frequency is estimated in three steps: subband signal decomposition, calculation of period candidate generation function and frequency tracking [6]. In order to make accurate estimation we use instantaneous harmonic parameters. First the signal is decomposed into overlapping subband analytic signals using Discrete Fourier Transform (DFT)-modulated filter bank. Then series of complex samples are transformed into instantaneous harmonic parameters [7, 8] and the values of period candidate generation function $\phi_{inst}(m,k)$ is calculated [6]:

$$\phi_{inst}(m,k) = \frac{\sum_{p=1}^{P} A_p^2(m)\cos(F_p(m)k)}{\sum_{p=1}^{P} A_p^2(m)},$$

where $A_p(m)$ – instantaneous amplitude, $F_p(m) \in [0,\pi]$ – instantaneous frequency, P – number of channels, m – signal sample number and k – lag in samples. The local maximum values of the function are traced from sample to sample using dynamic programming technique in order to impose constraints on fundamental frequency deviation speed. The estimation technique has inherent delay of 50 ms. An example of fundamental frequency contour estimation for speech degraded by a formula 1 car noise is given in figure 2.

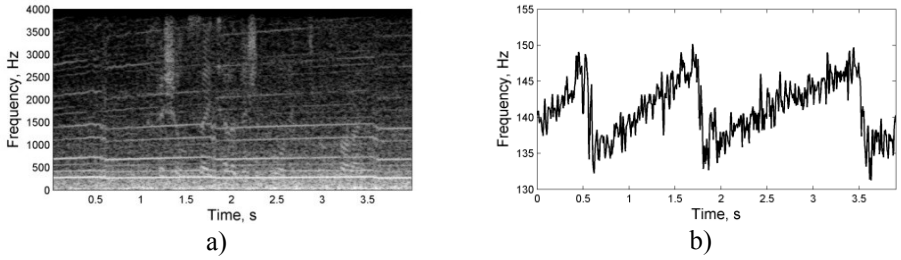

Fig. 2. Acquisition of rotation rate from noised signal using the fundamental frequency estimation algorithm. a – noised signal, b – estimated fundamental frequency

2.2 Time-warping

In order to stabilize frequencies of tonal components and PSD of the noise we eliminate pitch modulations using time-warping [6, 9]. Discrete signal $s(n)$ is interpolated in new time moments m so each rotation period corresponds to an equal number of samples N_{f_0}. Each time sample $s(n)$ is associated with a phase mark $\phi(n)$ using instantaneous pitch values $f_0(n)$:

$$\phi(n) = \sum_{i=0}^{n} f_0(i)$$

Interpolation moments m are obtained as:

$$m = \phi^{-1}(q/N_{f_0})$$

where q is sample index in warped time (phase) domain. The samples of the warped signal $s(q)$ are recalculated using sinc-interpolation. The result of time-warping of the same signal is given in figure 3.

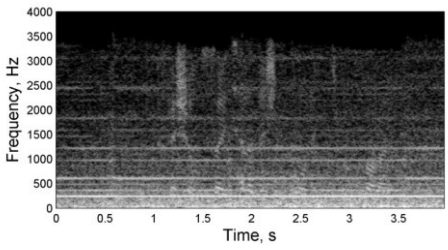

Fig. 3. Warped signal with constant fundamental frequency

2.3 Noise reduction

The time-warped signal is much more suitable for noise reduction. Noise reduction itself consists of the following operations: tone removing and noise subtraction. Both operations are implemented on short-time Fourier spectrum of the signal, however we use analysis windows of different lengths. For tone detection and removing an long-size (about 100 ms) STFT is used. Considering that the signal contains tonal noise components with constant frequencies their locations in the spectrum emerge as clear sharp peaks. The tone detector analyzes time-averaged amplitude spectrum to find frequency bins with relatively high energy (in comparison with adjacent bins). Speech harmonics are not detected as tonal noise because due to frequency variations its energy smoothed on the long analysis frame. Narrow-band components of the detected tones are removed from the spectrum. After removing tonal noise components wideband noise are attenuated using spectral subtraction technique [10]. We use short STFT windows (about 20 ms).

Considering that due to time-warping the noise spectral envelope becomes less de-pendent on pitch variations the noise reduction technique works very effectively. The PSD of noise is updated in silent regions where no voice is present. We use the voice activity detector described in [11]. After spectral subtraction the original time scale of the signal is restored by inverse time warping. The processing results for the time-warped signal are shown in figure 4.

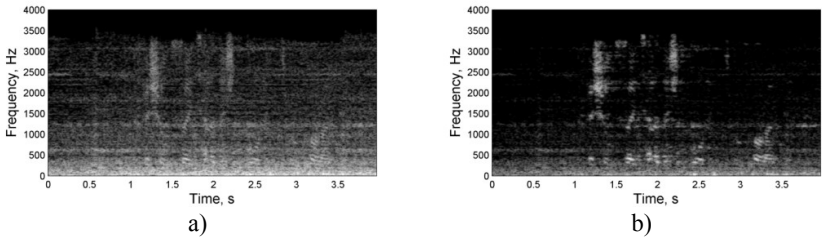

Fig. 4. Results of two-step noise reduction: a – tone removal, b – spectral subtraction

3 Experimental result

This section provides performance evaluation of the proposed algorithm. First, accuracy of fundamental frequency estimation is evaluated using synthetic quasi-periodical signals with different harmonic-to-noise ratios. Secondly spectrogram analysis is provided in order to compare obtained processing result with the method of maximum a posteriori estimation of noise from non-acoustic reference signals in very low signal-to-noise ratio environments [2]. Finally results of listening tests are presented.

3.1 Fundamental frequency estimation accuracy

In the context of the target application there are three main characteristics important for estimation of fundamental frequency: robustness against noise, robustness against tone modulations and frequency resolution. These characteristics are evaluated using three error measures: 1) gross pitch error (GPE); 2) mean fine pitch error (MFPE) and 3) root-mean square-error (RMSE) on signals with different amounts of noise and frequency modulations. Percentage of GPE is calculated as

$$\mathrm{GPE}(\%) = \frac{N_{\mathrm{GPE}}}{N_v} \times 100$$

where N_{GPE} – the quantity of frames with estimated fundamental frequency error more than $\pm 20\%$ of the true value, N_v – overall quantity of frames.

Mean fine pitch error is calculated on frames where no gross pitch errors occur.

$$\mathrm{MFPE}(\%) = \frac{1}{N_{\mathrm{FPE}}} \sum_{n=1}^{N_{\mathrm{FPE}}} \frac{|F_0^{true}(n) - F_0^{est}(n)|}{F_0^{true}(n)} \times 100,$$

where N_{FPE} – number of frames without GPE, $F_0^{true}(n)$ – true fundamental frequency and $F_0^{est}(n)$ estimated value. Root-mean square-error is evaluated in the following way:

$$\mathrm{RMSE}(\%) = \sqrt{\frac{1}{N_{\mathrm{FPE}}} \sum_{n=1}^{N_{\mathrm{FPE}}} \left[\frac{F_0^{true}(n) - F_0^{est}(n)}{F_0^{true}(n)} \times 100 \right]^2},$$

Test signals are generated using different fundamental frequency change rates and harmonic-to-noise rates (HNR)

$$\mathrm{HNR} = 10\mathrm{lg}\frac{\sigma_H^2}{\sigma_e^2},$$

σ_H^2 – the energy of the harmonic signal and σ_e^2 – the energy of the noise. Fundamental frequency is changes from 100 to 350 Hz. The performance of fundamental frequency evaluation is summarized in Table 1.

Table 1. Fundamental frequency evaluation GPE (%), MFPE(%) and RMSE(%)

	Fundamental frequency change rate, Hz/ms				
	0	0.5	1	1.5	2
	$HNR = 0$ dB				
GPE	0	0	0	0	0
MFPE	0.14	0.13	0.2	0.21	0.44
RMSE	0.15	0.16	0.24	0.32	0.69
	$HNR = -10$ dB				
GPE	0	0	0	1.31	1.52
MFPE	0.34	0.52	0.86	1	1.68
RMSE	0.58	0.64	1.12	1.77	2.42

The proposed algorithm showed a good robustness against noise and frequency modulations that makes possible to use it instead of the reference signal in motor noise processing.

3.2 Performance analysis

The following picture gives results of the proposed technique and the method of maximum a posteriori estimation of noise from non-acoustic reference signals (MAP) for the signal we used in the previous section – figure 5.

As can be seen the proposed algorithm provides a good average noise attenuation (18 dB) that is slightly better compared to MAP (17 dB). Moreover, in the high frequency band (higher than 2 kHz) the proposed method retains lesser amount of noise. The harmonic structure of the speech signal is a bit clearer perceived and less degraded than in MAP. However because of fully automatic harmonic noise detection due to absence of non-acoustic reference signals a small amount of harmonic noise has been left in the result.

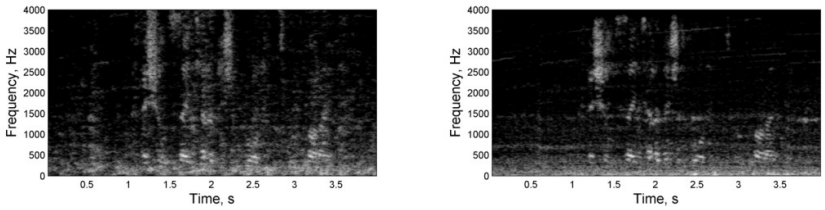

Fig. 5. Comparison of MAP and the proposed processing technique a – MAP output, b – output of the proposed algorithm

Listening tests were carried out using the same subjective measures as in [2], where Comparative Mean Opinion Score (CMOS) scale was used. The speech quality improvement is evaluated from source/processed speech pairs on a scale of -3 to $+3$, where -3 corresponds to a significant loss of quality, 0 to no perceived difference and $+3$ to significant quality improvement.

A group of 10 listeners participated in the listening. In each test every listener was listening to 3 pairs of speech samples and rated quality change. The averaged listening results for MAP filtering is 1.62 and for the proposed algorithm 1.71.

Conclusion

A method for racing car noise reduction has been proposed. The noise processing is made on a warped signal with constant fundamental frequency that makes it easier to estimate and remove noise components because of their stationarity. During noise reduction the narrow-band and wide-band noise components are processed separately. The method can be applied directly to one-channel noised signals and does not utilize any non-acoustic reference information. The method has provided good results with comparison to the MAP technique that utilize external engine speed measurements. The main drawback of the method is increased inherent delay which is about 100 ms.

Acknowledgments. This work was supported by Belarusian Republican Foundation for Fundamental Research (grant No F14MV-014).

References

1. Hadley, M., Milner, B., Harvey, R.: Noise reduction for driver to pit-crew communication in motor racing. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 165–168. IEEE Press, Toulouse, (2006).
2. Milner, B.: Maximum a posteriori estimation of noise from non-acoustic reference signals in very low signal-to-noise ratio environments. In: 12th Annual Conference of the International Speech Communication Association (Interspeech), pp. 357–360. Florence, (2011).

3. Gomez, P., Alvarez, A., Nieto, V., Martinez, R.: Speech enhancement for a car environment using LP residual signal and spectral subtraction. In: 8th European Conference on Speech Communication and Technology (Eurospeech), pp. 1373–1376. Geneva, (2003).
4. Vaseghi, S., Chen, A., McCourt, P.: State based sub-band LP Wiener filters for speech enhancement in car environments. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 213–216. IEEE Press, Istanbul, (2000).
5. Puder, H., Steffens, F.: Improved noise reduction for handsfree car phones utilizing information on vehicle and engine speeds. In: 10-th European Signal Processing Conference (EUSIPCO), pp. 1851–1854. Tampere, (2000).
6. Azarov, E., Vashkevich, M., Petrovsky, A.: Instantaneous Pitch Estimation Based on RAPT Framework. In: 20-th European Signal Processing Conference (EUSIPCO), pp. 1851–1854. Bucharest, (2012).
7. Petrovsky, Al., Azarov, E., Petrovsky, A.: Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding. Elsiver, Signal Processing, vol. 91, Issue 6, Fourier Related Transforms for Non-Stationary Signals, pp. 1489–1504 (2011).
8. Azarov, E., Petrovsky, A.: Instantaneous harmonic analysis: audio and speech processing in multimedia systems (in Russian). Lambert Academic Publishing (2011).
9. Petrovsky, A., Stankevich, A., Balunowski, J.: The order tracking front-end algorithms in the rotating machine monitoring systems based on the new digital low order filtering. In International Congresses on Sound and Vibration, pp. 2985-2992. Copenhagen, (1999).
10. Loizou, P. Speech enhancement: theory and practice. CRC press, Inc., (2007).
11. Puder, H. Single channel noise reduction using time-frequency dependent voice activity detection. In International Workshop on Acoustic Signal Enhancement, pp. 68-71, USA, Pocono Manor (1999).