Available online at www.sciencedirect.com

ScienceDirect

Digital Signal Processing 18 (2008) 581–587

Digital Signal Processing

www.elsevier.com/locate/dsp

An improved spectral subtraction method for speech enhancement using a perceptual weighting filter

Radu Mihnea Udrea *, Nicolae D. Vizireanu, Silviu Ciochina

Politehnica University of Bucharest, Telecommunications Department, Iuliu Maniu 1-3, Bucharest, Romania

Available online 15 August 2007

Abstract

We propose an improved spectral subtraction method for reducing acoustic noise added to speech in noisy environments like helicopter cockpit or car engine. This implementation uses oversubtraction method for spectral subtraction. Residual noise can be masked by exploiting the masking properties of the human auditory system. A psychoacoustically motivated weighting filter was included to eliminate residual musical noise. Simulations showed a better quality with no distortion for the enhanced speech and the musical noise was effectively masked.

Keywords: Speech enhancement; Spectral subtraction

1. Introduction

The speech signal is often accompanied by the background noise of the environment. There are many negative effects when processing the degraded speech for some applications like: voice command systems, voice recognition, speaker authentication, hands-free systems.

Enhancement techniques can be classified as single channel and dual channel or multi-channel enhancement techniques. Single channel enhancement techniques apply to situations in which only one acquisition channel is available.

The spectral subtraction method is a well-known single channel noise reduction technique [1,2]. Most implementations and variations of the basic technique apply subtraction of the noise spectrum estimate over the speech spectrum. The conventional power spectral subtraction method substantially reduces the noise levels in the noisy speech. However, it also introduces an annoying distortion in the speech signal called musical noise.

In this paper we used a modified spectral oversubtraction approach [3] that allows better and more suppression of the noise. The approach used was to estimate the power frequency spectrum of the clean speech by subtracting an overestimate of the noise power spectrum from the speech power spectrum. In addition, a psychoacoustically motivated spectral weighting rule was incorporated to find the best tradeoff between speech distortion and noise reduction. The existing residual noise can be masked by exploiting the masking properties of the human auditory system. A modified masking threshold estimation was used to eliminate noise influence when determine the speech masking threshold.

[*] Corresponding author.
E-mail address: mudrea@comm.pub.ro (R.M. Udrea).

1051-2004/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.dsp.2007.08.002

The proposed method gives a superior performance as compared to the conventional method of power spectral subtraction and largely reduces the musical noise.

2. Spectral oversubtraction method

The basic assumption of the method is treating the noise as uncorrelated additive noise. Assume that a speech signal $s(t)$ has been degraded by the uncorrelated additive noise signal $n(t)$:

$$x(t) = s(t) + n(t). \tag{1}$$

Short time power spectrum of the noisy speech can be approximated by

$$\left|X(e^{j\omega})\right|^2 \approx \left|S(e^{j\omega})\right|^2 + \left|N(e^{j\omega})\right|^2. \tag{2}$$

The power spectral subtraction estimator results by replacing noise square-magnitude $|N(e^{j\omega})|^2$ with its average value $|\bar{N}(e^{j\omega})|^2$ taken during nonspeech activity period.

$$\left|\hat{S}(e^{j\omega})\right|^2 = \left|X(e^{j\omega})\right|^2 - \left|\bar{N}(e^{j\omega})\right|^2. \tag{3}$$

Due to differences between estimated and effective noise a residual noise appears after applying spectral subtraction method. This is perceived as a distortion in the speech signal called musical noise. Berouti [1] proposed an important variation of spectral subtraction for reduction of residual musical noise. An overestimate of the noise power spectrum is subtracted and the resulted spectrum is limited from going below a preset minimum level (spectral floor). The proposed algorithm could be expressed as

$$\left|\hat{S}(e^{j\omega})\right|^2 = \begin{cases} |X(e^{j\omega})|^2 - \alpha|\bar{N}(e^{j\omega})|^2, & \text{if } |\hat{S}(e^{j\omega})|^2 > \beta|\bar{N}(e^{j\omega})|^2, \\ \beta|\bar{N}(e^{j\omega})|^2, & \text{otherwise}, \end{cases} \tag{4}$$

where α is the subtraction factor and β is the spectral floor parameter.

To reduce the speech distortion caused by large values of α, its value is adapted from frame to frame. The basic idea is to take into account that the subtraction process must depend on the segmental noisy signal to noise ratio (NSNR) of the frame, in order to apply less subtraction with high NSNRs and vice versa.

Segmental noisy signal to noise ratio NSNR is calculated for every frame with

$$\text{NSNR(dB)} = 10\log_{10}\frac{\sum_{k=0}^{N-1}|X(e^{j\omega})|^2}{\sum_{k=0}^{N-1}|\bar{N}(e^{j\omega})|^2}. \tag{5}$$

The oversubtraction factor α can be calculated [1] as

$$\alpha_i = \begin{cases} 1, & \text{NSNR}_i \geqslant 20\ \text{dB}, \\ \alpha_0 - \frac{3}{20}\text{NSNR}_i, & -6\ \text{dB} \leqslant \text{NSNR}_i < 20\ \text{dB}, \\ 4.9, & \text{NSNR}_i < -6\ \text{dB}, \end{cases} \tag{6}$$

where $\alpha_0 = 4$ is the desired value of α at 0 dB NSNR.

The enhanced speech spectrum is obtained using magnitude estimate $|\hat{S}(e^{j\omega})|$ and the phase $\theta_X(\omega)$ of the input signal:

$$\hat{S}(e^{j\omega}) = \left|\hat{S}(e^{j\omega})\right|e^{j\theta_X(\omega)}. \tag{7}$$

The phase of the input signal is used for reconstruction of the estimated signal spectrum based on the fact that for human perception the short time spectral amplitude is more important than the phase for intelligibility and quality [4].

3. Auditory masking model

In order to further enhance the quality of speech, a psychoacoustically motivated spectral weighting rule was incorporated. This approach was motivated from the algorithm proposed in [5]. Some musical residual noise remains in the estimated clean speech after spectral oversubtraction. The existing residual noise can be masked by exploiting the masking properties of the human auditory system.

There are three types of masking effects: simultaneous, forward and backward. We considered only the simultaneous effect of masking in frequency where a low level signal is masked (is inaudible) by a simultaneously, nearby-frequency, stronger signal.

First of all we compute the noise masking threshold that gives the maximum level of noise that is inaudible in the presence of speech. The noise masking threshold $T(k)$ is obtained through modeling the frequency selectivity of the human ear and its masking property. The different calculation steps taken from [6] are described below.

3.1. Frequency analysis along the critical band scale

In the frequency range from 0–4 kHz, there are 18 critical bands. The first step is the critical band analysis, wherein the FFT power spectrum $|\hat{S}(e^{j\omega})|^2$ of the estimated clean speech from the proposed spectral subtraction method is used, and the energies in each critical band are added up. This way we obtain the power spectral density on a Bark scale k:

$$B(k) = \sum_{w(k)} \left|\hat{S}(e^{j\omega})\right|^2, \quad k = 1, \ldots, K, \tag{8}$$

where k is the critical band number, $K = 18$ is the total number of critical bands and $w(k)$ is the frequency index depending on the lower and upper frequency boundary of the critical band k.

3.2. Convolution with a spreading function SF(k)

There is masking between the different critical bands and this interaction of the masking signals between different critical bands is modeled by application of a spreading function, which is asymmetric in frequency and operates on a Bark scale. An analytical expression for the spreading function is given by [7]

$$\mathrm{SF}(k) = 15.81 + 7.5(k + 0.474) - 17.5\sqrt{1 + (k + 0.474)^2}\ [\mathrm{dB}]. \tag{9}$$

This spreading function is convolved with the bark spectrum, to give the spread critical band spectrum $C(k)$:

$$C(k) = \mathrm{SF}(k) * B(k). \tag{10}$$

3.3. Relative threshold offset

A relative threshold offset $O(k)$ is subtracted from each critical band. For calculation of this threshold, it is required to distinguish between tone-like and noise-like components. There are two types of noise masking thresholds $T_{\mathrm{N}}(k)$ and $T_{\mathrm{T}}(k)$, depending on the noise-like or tone-like nature of the masking and masked signal:

$$\begin{aligned} &\text{tone masking a noise:} \quad T_{\mathrm{N}}(k) = C(k) - 14.5 - k, \\ &\text{noise masking a tone:} \quad T_{\mathrm{T}}(k) = C(k) - 5.5. \end{aligned}$$

To determine whether the signal is tone-like or noise-like, the spectral flatness measure (SFM) is used. The SFM (in dB) is defined as

$$\mathrm{SFM}_{\mathrm{dB}} = 10\log_{10}\left(\frac{G_{\mathrm{m}}}{A_{\mathrm{m}}}\right), \tag{11}$$

where G_{m} and A_{m} represent the geometric and arithmetic mean of the power spectrum respectively.

From this value, a tonality coefficient α_{SFM} is generated by

$$\alpha_{\mathrm{SFM}} = \min\left(\frac{\mathrm{SFM}_{\mathrm{dB}}}{\mathrm{SFM}_{\mathrm{dB\,min}}},\ 1\right), \tag{12}$$

where $\mathrm{SFM}_{\mathrm{dB\,min}} = -60$ dB represents the SFM of an entirely tone-like signal resulting in the tonality coefficient of $\alpha_{\mathrm{SFM}} = 1$. Conversely an entirely noise-like signal would have $\mathrm{SFM}_{\mathrm{dB}} = 0$ dB and thus $\alpha_{\mathrm{SFM}} = 0$.

Using α_{SFM}, the offset in dB for each band is calculated as

$$O(k) = \alpha_{\mathrm{SFM}}(14.5 + k) + (1 - \alpha_{\mathrm{SFM}}) \times 5.5. \tag{13}$$

This offset is then subtracted from the spread spectrum in the dB domain by

$$T(k) = 10^{\log_{10}|C(k)| - |O(k)/10|}. \tag{14}$$

The masking threshold computation for the speech signal is affected by residual noise. A method to estimate masking threshold for the clean speech signal from enhanced speech is presented in the next section.

4. Estimation of noise masking threshold for perceptual weighting

The estimate of the clean speech signal is computed with power spectral oversubtraction, resulting a reduction of background noise but introducing musical residual noise, especially at low SNR's. This residual noise modifies the tonality of the signal and the masking threshold is slightly different from the one obtained from the clean speech, especially for high frequencies.

Residual noise variance can be estimated during nonspeech activity [3], when the input signal is noise only. Maximum value of the difference between each frame noise power spectrum and its average can be taken over a number of L frames:

$$N_{\mathrm{R}}(e^{j\omega}) = \max_{L\ \text{frames}} \left| \left|N(e^{j\omega})\right|^2 - \left|\bar{N}(e^{j\omega})\right|^2 \right|. \tag{15}$$

Consequently, the relative threshold offset, has to be decreased with residual noise variance to take into account the tone-like nature of the musical residual noise. The residual noise variance is computed on every Bark scale k:

$$N_{\mathrm{R}}(k) = \sum_{w(k)} N_{\mathrm{R}}(e^{j\omega}). \tag{16}$$

The estimated masking threshold is computed by

$$\hat{T}(k) = 10^{\log_{10}|C(k)| - |O(k)/10| - |N_{\mathrm{R}}(k)/10|}. \tag{17}$$

From experiments results that this modification of threshold computation has to be made for high frequency domain (critical band $k > 12$).

5. Perceptual weighting filter

Let $\hat{S}(e^{j\omega}) = G(e^{j\omega})X(e^{j\omega})$ be the enhanced speech spectrum and $G(e^{j\omega})$ the perceptual weighting filter. The error spectrum between the spectrum of the clean signal and the estimated (enhanced) spectrum is given by

$$E(e^{j\omega}) = \hat{S}(e^{j\omega}) - S(e^{j\omega}) = G(e^{j\omega})Y(e^{j\omega}) - S(e^{j\omega}) = \left[G(e^{j\omega}) - 1\right]S(e^{j\omega}) + G(e^{j\omega})\bar{N}(e^{j\omega}). \tag{18}$$

The first term in the above equation describes the speech distortion caused by the spectral weighting which can be minimized using $G(e^{j\omega}) = 1$, the second term describes the noise distortion which can be minimized if the weighting filter $G(e^{j\omega}) = 0$. We can then compute a spectral weighting function $G(e^{j\omega})$ such that the noise and speech distortions fall below the masking threshold. We chose to estimate the weighting function that would minimize the noise distortion (in the sense of making it inaudible), while allowing a variable speech distortion.

Therefore the weighting function $G(e^{j\omega})$ was chosen to satisfy the following criteria:

$$\left|G(e^{j\omega})\right|\left|\bar{N}(e^{j\omega})\right| \leqslant \hat{T}(e^{j\omega}) \tag{19}$$

with the constraint $0 \leqslant G(e^{j\omega}) \leqslant 1$, where $\hat{T}(e^{j\omega})$ is the clean speech estimated masking threshold. Therefore the psychoacoustically motivated weighted filter is obtained as follows:

$$G(e^{j\omega}) = \min\left(\frac{\hat{T}(e^{j\omega})}{|\bar{N}(e^{j\omega})|}, 1\right). \tag{20}$$

Figure 1 presents the proposed method block diagram.

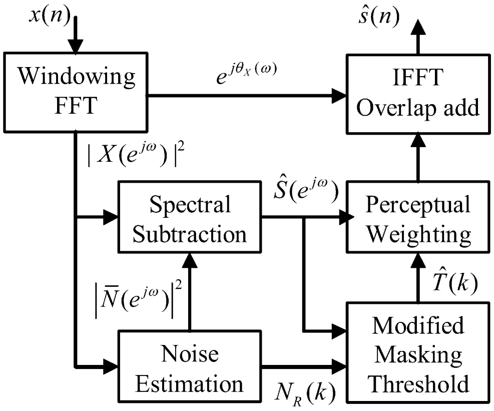

Fig. 1. Block diagram of the proposed spectral subtraction method with perceptual weighting.

Table 1
Objective measures obtained for different noise types for an input signal of 0 dB SNR

Noise type	Noisy signal			Enhanced signal		
	SNR	IS	MOS	SNR	IS	MOS
White Gaussian noise	0 dB	4.3	1.21	9.1 dB	2.3	2.84
Car noise	0 dB	3.8	1.68	8.9 dB	2.0	3.02
Aircraft cockpit noise	0 dB	3.6	1.03	7.3 dB	2.3	2.65
Helicopter cockpit noise	0 dB	2.6	1.26	7.1 dB	1.9	2.22
Factory noise	0 dB	4.1	1.80	7.8 dB	2.5	2.98

6. Performance evaluation

Input signal is sampled with a sampling frequency of 8 kHz and windowed with a 256 samples Hanning window with 50% overlap. Noise signals were taken from the Noisex-92 database, designed for speech recognition in noisy environments, and have different time-frequency distributions: white Gaussian noise, car noise, aircraft cockpit noise, helicopter cockpit noise, and factory noise. Noise has been added to the clean speech signal with a varying SNR. The power spectrum of the windowed data is calculated and subtracted by the average noise spectrum calculated during nonspeech activity.

The objective performance evaluation is based on the application of objective quality or intelligibility measures.

Itakura–Saito distortion (IS) is a objective quality measure that performs a comparison between spectral envelopes (all-pole parameters) and that is more influenced by a mismatch in formant location than in spectral valleys [8]. A typical range for the IS measure is 0–10, where the minimal value of IS corresponds to the best speech quality.

Also, to measure perceptual evaluation of speech quality (PESQ), we used the ITU-T P.862 standard [9]. PESQ is able to predict subjective quality with good correlation in a very wide range of conditions, which may include coding distortions, errors, noise, filtering, delay, and variable delay. The PESQ MOS as defined by the ITU recommendation P.862 ranges from 1.0 (worst) up to 4.5 (best).

Time-frequency distribution of the enhanced speech can be observed from Fig. 2 where spectrograms of input corrupted speech, enhanced speech using only over-subtraction method and enhanced speech using perceptual weighting are presented. This figures give a more accurate information about residual noise and speech distortion than the corresponding time waveforms.

7. Conclusions

This paper presents an improved spectral subtraction method combined with a perceptual weighting filter based on psychoacoustical properties. A modified masking threshold estimation was used to eliminate noise influence when de-

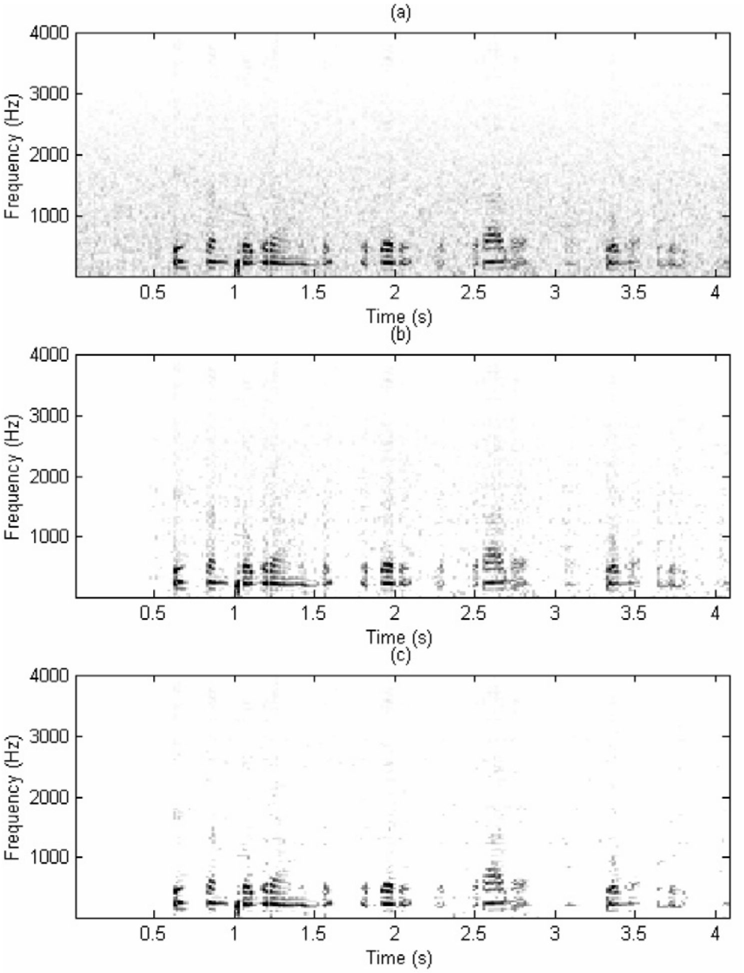

Fig. 2. Spectrograms of the (a) noise corrupted speech, (b) enhanced speech using oversubtraction method, (c) enhanced speech using oversubtraction and perceptual weighting.

termine the speech masking threshold. After applying perceptual weighting filter, simulations showed a better quality with no distortion for enhanced speech and that the musical noise was effectively masked.

References

[1] M. Berouti, R. Schwartz, J. Makhoul, Enhancement of speech corrupted by acoustic noise, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., April 1979, pp. 208–211.

[2] C.-T. Lin, Single-channel speech enhancement in variable noise-level environment, IEEE Trans. Syst. Man Cybernet. A 33 (1) (2003) 137–143.

[3] R.M. Udrea, S. Ciochină, Speech enhancement using spectral oversubtraction and residual noise reduction, in: Proc. of the Symposium SCS 2003, vol. II, Iaşi, Romania, July 2003, pp. 165–169.

[4] D.L. Wang, J.S. Lim, The unimportance of phase in speech enhancement, IEEE Trans. Acoust. Speech Signal Process. 30 (4) (1982) 679–681.

[5] S. Gustafsson, P. Jax, P. Vary, A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., vol. 1, 1998, pp. 397–400.

[6] J.D. Johnston, Transform coding of audio signals using perceptual noise criteria, IEEE J. Select. Areas Commun. 6 (2) (1988) 314–323.

[7] M.R. Schroeder, B.S. Atal, J.L. Hall, Optimizing digital speech coders by exploiting masking properties of the human ear, J. Acoust. Soc. Amer. 66 (6) (1979) 1647–1652.

[8] J. Deller, J. Proakis, J. Hansen, Discrete-Time Processing of Speech Signals, Prentice Hall, Englewood Cliffs, NJ, 1993.

[9] ITU-T Rec. P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, International Telecommunications Union, Geneva, Switzerland, February 2001.

Mihnea Radu Udrea received his MSEE and Ph.D. degrees in electronics, telecommunications and informations technology from the Politehnica University of Bucharest. He is an Associate Professor at the Politehnica University of Bucharest, Faculty of Electronics, Telecommunications and Informations Technology, Telecommunications Department. His work has primarily focused on digital signal processing, digital signal processors application on communications, and multimedia communications real time implementation.

Nicolae Dragos Vizireanu received his M.S. degree in electrical engineering from the Georgia Institute of Technology in 1996, and his Ph.D. degree in electronics and telecommunications from the Politehnica University of Bucharest in 1998. Since 2007 he is a Professor at the Politehnica University of Bucharest, Faculty of Electronics, Telecommunications and Informations Technology, Telecommunications Department. His work has primarily focused on digital/image and video processing, pattern recognition, real time implementation with focus on communications and multimedia systems, and hardware–software DSP processor systems.

Silviu Ciochina received his M.S. degree in electronics and communications engineering in 1971, and his Ph.D. degree in 1978, both from the Politehnica University of Bucharest. He is a Professor at the Politehnica University of Bucharest, Faculty of Electronics, Telecommunications and Information Technology, Telecommunications Department. His main areas of interest are digital signal processing, adaptive algorithms, spectrum and DOA estimation, and wireless communications.