Speech Distortion Minimized Noise Reduction Algorithm

Ekaterina Verteletskaya, Boris Simak

***Abstract*— In this paper, we propose a method for enhancing of speech corrupted by broadband noise. The method is based on the spectral subtraction technique. Besides reducing the noise conventional spectral subtraction introduces an annoying residual musical noise. To eliminate the musical noise we propose to introduce reduced varying scaling factor of spectral subtraction, with a following application of weighted function. Weighting function, used in the proposed algorithm, attenuates frequency spectrum components lying outside identified formants regions. Algorithm effects a substantial reduction of the musical noise without significantly distorting the speech. Listening tests were performed to determinate the subjective quality and intelligibility of speech enhanced by our method.**

***Index Terms*—musical noise, noise estimation, spectral subtraction, speech enhancement, VAD.**

I. Introduction

Broadband noise presented in speech signal recorded under the real conditions can impair the quality of the signal, reduce intelligibility, and increase listener fatigue. Since in practice many kinds of noise is presented in recording speech, the problem of noise reduction is essential in the world of telecommunications and has gained much attention in recent years. Various classes of noise reduction algorithms have been developed mostly based on transform domain techniques, adaptive filtering, and model-based methods. Amongst the speech enhancement techniques, DFT-based transforms domain techniques have been widely spread in the form of spectral subtraction [1]. Even though the algorithm has very low computational complexity, it can reduce the background noise effectively. However, experimental results show that there is some residual noise in the processed signal, which affects the hearing effect. To reduce the influence of the background noise and increase the definition of the speech, the algorithm based on the modified spectral subtraction algorithm is introduced in this paper.

Manuscript received May 24, 2010. This work was supported by Czech Technical University grant SGS No. OHK3-108/10 and the Ministry of Education, Youth and Sports of Czech Republic by the research program MSM6840770014.

E. Verteletskaya is with the Department of Telecommunication Engineering Czech Technical University in Prague, Prague, Czech Republic (e-mail: verteeka@fel.cvut.cz).

B. Simak , is with the Department of Telecommunication Engineering Czech Technical University in Prague, Prague, Czech Republic (e-mail: simak@fel.cvut.cz).

II. Spectral Subtraction Method

A. *The principle of spectral subtraction*

The spectral subtraction is based on the principle that the enhanced speech can be obtained by subtracting the estimated spectral components of the noise from the spectrum of the input noisy signal. Assuming that noise *w(n)* is additive to the speech signal *x(n)*, the noisy speech *y(n)* can be written as,

$$y\ (n) = x\ (n) + w\ (n),\ \ for\ \ 0 \le n \le N-1 \qquad (1)$$

Where *n* is the time index, N is a number of samples. The objective of speech enhancement is to find the enhanced speech $\hat{x}\ (n)$ from given *y(n),* with the assumption that *w(n)* is uncorrelated with *x(n).* Input signal *y(n)* is segmented into K segments of the same length. The time-domain signals can be transformed to the frequency-domain as,

$$Y_k = X_k + W_k,\ \ for\ \ 0 \le k \le K-1 \qquad (2)$$

Where k is the segment index, Y_k, X_k and W_k denote the short-time DFT magnitudes taken of *y(n)*, *x(n)*, and *w(n)*, respectively, and raised to a power *a* (*a=1* corresponds to magnitude spectral subtraction, *a=2* corresponds to power spectrum subtraction). If an estimate of the noise spectrum $\hat{W}_k$ can be obtained, then an approximation of speech $\hat{X}_k$ can be obtained from Y_k

$$\hat{X}_k = Y_k - \hat{W}_k \qquad (3)$$

The noise spectrum cannot be calculated precisely, but can be estimated during period when no speech is present in the input signal. Most single channel spectral subtraction methods use a voice activity detector (VAD) to determine when there is silence in order to get an accurate noise estimate. The noise is assumed to be short-term stationary, so that noise from silent frames can be used to remove noise from speech frames.

Fig. 1 shows a block diagram of the spectral subtraction method. The harshness of the subtraction can be varied by applying a scaling factor α [1]. The values of scaling factor α higher than 1 result in high SNR level of denoised signal, but too high values may cause distortion in perceived speech quality. Subtraction process with applying scaling factor α can be expressed as:

$$\hat{X}_k = Y_k - \alpha \cdot \hat{W}_k \qquad (4)$$

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

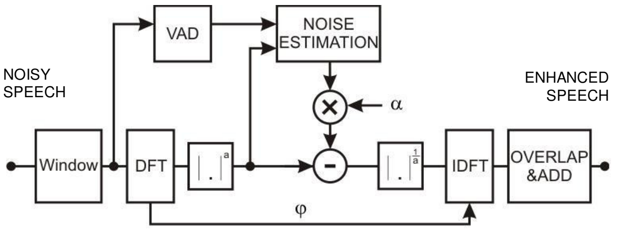

Figure1. Spectral subtraction algorithm block-diagram

After subtraction, the spectral magnitude is not guaranteed to be positive and some possibilities to remove the negative components are by half-wave rectification (setting the negative portions to zero), or full wave rectification (absolute value). Half-wave rectification is commonly used, but introduces musical tone artifacts in the processed signal. Full wave rectification avoids the creation of musical noise, but less effective at reducing noise. After subtraction, *a* root of the $\hat{X}_k$ is extracted to provide corresponding Fourier amplitude components. An inverse Fourier transform, using phase components directly from Fourier transform unit, and overlap add is then done to reconstruct the speech estimate in the time-domain.

B. *Musical noise*

Although spectral subtraction method provide an improvement in terms of noise attenuation, it often produce a new randomly fluctuating type of noise, referred to as musical noise due to their narrow band spectrum and presence of tone-like characteristics. This phenomenon can be explained by noise estimation errors leading to false peaks in the processed spectrum. When the enhanced signal is reconstructed in the time-domain, these peaks result in short sinusoids whose frequencies vary from frame to frame. Musical noise although very different from the original noise, can sometimes be very disturbing. A poorly designed spectral subtraction, which caused musical noise, can sometime results in the signal that has lower perceived quality and lower information content, than the original noisy signal. Most of the research at the present time is focused in ways to combat the problem of musical noise [2]. It is almost impossible to minimize musical noise without affecting the speech, and hence there is a tradeoff between the amount of noise reduction and speech distortion. Due to this fact several perceptual based approaches were introduced, wherein instead of completely eliminating the musical noise (and introducing distortion), the noise is masked taking advantage of the masking properties of the auditory system [3].

III. VAD Algorithm

The decision about voice activity presence is the sensitive part of the whole spectral subtraction algorithm as the noise power estimation can be significantly degraded by the errors in voice activity detection. VAD accuracy dramatically affects the noise suppression level and amount of speech distortion that occurs. Many different techniques have been applied to the art of VAD. In the early VAD algorithms, short-time energy, zero-crossing rate, and linear prediction coefficients were among the common feature used in the detection process. Energy-based VADs [4] are frequently used because of their low computation costs. They work on principle, that the energy of the signal is compared with the threshold depending on the noise level. Speech is detected when the energy estimation lies over the threshold. Dynamical energy-based VAD described in [5] is used in proposed enhanced spectral subtraction method. In classical energy-based algorithms, detector cannot track the threshold value accurately, especially when speech signal is mostly voice-active and the noise level changes considerably before the next noise level re-calibration instant. The dynamical VAD was proposed to provide its classification more accurately in comparison with other energy-based techniques.

IV. Proposed Method

A. *Weighting function*

As already discussed above, the spectral subtraction technique employed in the apparatus of Fig.1 has the disadvantage that output, though less noisy than the input signal, contains musical noise. The majority of information in a segment of noise-free speech is contained within one or more high energy frequency bands, known as formants. Within the formant regions themselves, the musical noise is largely masked out by the speech itself. Proposed spectral subtraction algorithm aims to reduce the audible musical noise by attenuating the signal in the regions of the frequency spectrum lying between the formant regions. Attenuation of the regions between the formants has little effect on the perceived quality of the speech itself, so that this approach is able to effect a substantial reduction in the musical noise without significantly distorting the speech. This attenuation is performed by weight function $H(\omega)$ derived from the spectral envelope $L(\omega)$, which is obtained by means of linear prediction analysis.

B. *Implementation*

The overall block diagram of proposed enhanced spectral subtraction method is shown on Fig.2. The noisy speech is segmented into overlapping frames. Then Hamming window is applied on each segment and a set of Fourier coefficients using short-time fast Fourier transform (FFT) is generated. Fourier coefficients are raised to a power *a* (*a* =1 or 2). Noise spectrum is estimated during periods when no speech is present in the input signal. This condition is recognized by voice activity detector (VAD) to produce a control signal which permits the updating of store with spectrum Y_k when speech is absent from the current segment. This spectrum is smoothed by making each frequency samples of Y_k the average of adjacent frequency samples, given $\hat{Y}_k$. This smoothed spectrum then will be used to update a spectral estimate of noise, which consists of a proportion of the previous noise and a portion of the smoothed short-term spectrum of current segment. Thus the noise spectrum gradually adapts to changes in the actual spectrum noise. It can be defined as

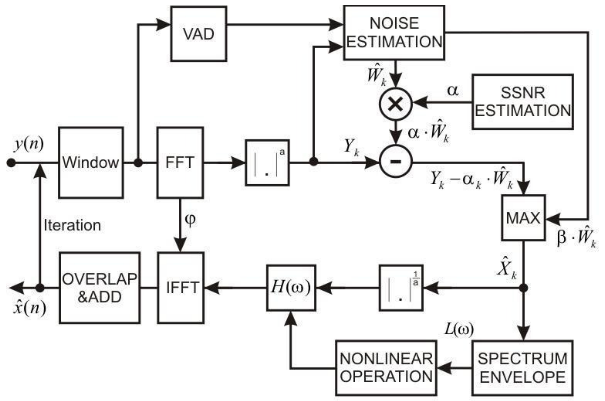

Figure 2. Block-diagram of proposed algorithm

$$\hat{W}_k = \lambda \cdot \hat{W}_{k-1} + (1-\lambda) \cdot \hat{Y}_k \tag{5}$$

Where $\hat{W}_k$ is the updated noise spectral estimate, k is a frame index, $\hat{W}_{k-1}$ is the old noise spectral estimate, $\hat{Y}_k$ is the smoothed noise spectrum from the present frame, and λ is a decay factor. Present noise spectral estimate is subtracted from the noisy speech power spectrum:

$$\hat{X}_k = Y_k - \alpha_k \cdot \hat{W}_k \tag{6}$$

The scaling factor α_k defines a subtraction dimension. When the spectral subtraction is followed by the weighting function *H(ω)* a lower value of the scaling factor can be used, therefore speech will be less distorted. The value of α_k in proposed method is the function of segmental SNR and varies from frame to frame within the same signal. Using the SSNR α_k can be determined for each frame as:

$$\alpha_k = \begin{cases} 3, & SSNR_k < 0dB \\ 3 - SSNR_k/10, & 0dB \le SSNR_k \le 20dB \\ 1, & SSNR_k > 20dB \end{cases} \tag{7}$$

Subtraction defined in (6) may result in negative terms. Since a frequency component cannot have a negative power, in proposed method a nonzero minimum noise power level $Z_k = \beta \cdot \hat{W}_k$ is defined, where β determinates the minimum power level or "spectral floor". Thereby output of spectral subtraction $\hat{X}_k$ is defined as the maximum of $Y_k - \alpha_k \cdot \hat{W}_k$ and $\beta \cdot \hat{W}_k$. A non zero value of *β* (*0<β<<1*) reduces the effect of musical noise by retaining a small amount of the original noise signal.

Musical noise in proposed method is also reduced by attenuating the signal in the regions of the frequency spectrum lying between the formant regions. This attenuation is performed by special unit, which multiplies the Fourier coefficients by respective terms of a weighting function *H(ω).* The weighting function is obtained from spectral envelope *L(ω)* which is obtained by means of a LPC analysis. The attenuation operation is such that any coefficient of the spectrally subtracted speech $\hat{X}_k$ is attenuated only if the corresponding frequency term of the spectral envelope is below a threshold value τ. Thus the response *H(ω)* is a nonlinear function of *L(ω)* and is obtained by nonlinear processing unit according to the rule:

$$\begin{aligned} & if\ L(\omega) \ge \tau,\ \ H(\omega) = 1 \\ & else\ H(\omega) = \left[\frac{L(\omega)}{\tau}\right]^{\gamma} \end{aligned} \tag{8}$$

The threshold value τ is a constant for all frequencies and for all speech segments. In a strongly voiced segment of speech, only small portions of the spectrum will be attenuated, whereas in quiet segments most of the spectrum may be attenuated. A value of the threshold about 10% of peak amplitude of the speech is found to work well. A lower value of τ will produce a more harsh filtering operation. Thus the value could be increased for higher SNR, and lowered for lower signal to noise ratios. The power term γ=2 were used. Larger value of γ will make the attenuation harsher. The value of γ may be used to vary the harshness of the attenuation.

After subtraction and nonlinear weighting, the a root of the output terms is taken to provide corresponding Fourier amplitude components, and the time-domain signal segments reconstructed by an inverse Fourier transform unit from these along with phase components φ directly from the FFT unit. The windowed speech segments are overlapped to provide the reconstructed output signal at an output.

We increase algorithm performance by iteration of whole process [6]. After the first spectral subtraction process, the type of additive noise is changed to that of musical noise. We estimate the noise signal from unvoiced segment parts using the VAD. Noise estimate accuracy increases due to VAD accuracy increasing after first spectral subtraction, which results in SNR improvement. We design a new spectral subtraction by using the new estimated noise (musical noise) and the new noisy speech (including the musical noise), which is the output signal by the first spectral subtraction. Therefore, musical noise also can be reduced by performing the iteration of spectral subtraction as shown on Fig. 2.

V. Experiments and Discussion

A. *Speech data and analysis parameters*

Proposed method has been tested on real speech data by computer simulation in MATLAB environment. Real speech signals from SpeechDat database were used for experiments. The sampling frequency of the signals is 8kHz. Utterances were pronounced by female and male speakers in Czech language. Three types of noise were used to generate noisy speech signals with the different SNR level (*SNR*= [0dB 5dB 10dB 15dB]). The noises used for experiments are following: AWGN noise generated by computer, office noise, and engine noise recorded inside the car taken from CAR2ECS database [7]. Parameters for computing are as follows. The frame length is 20ms (160 samples). Each frame is 50 % overlapped. Frames were windowed with Hamming window and 256 points FFT is applied to each frame. We have experimented with two values of power *a* (*a=1,a=2*), and with or without iteration of proposed algorithm. The value *a=2* and one iteration of whole algorithm were found to result best algorithm performance. Listening tests were held to evaluate quality and intelligibility of enhanced speech.

B. *Numerical results*

The results of the proposed noise reduction algorithm, as well as comparison to conventional spectral subtraction algorithm are shown on Fig. 3. Evaluation of the algorithm performance for different kinds of noise and different levels of SNR is indicated by an improvement of segmental SNR (SSNR). The SSNR is a mean value of SNR in each frame. The improvement of SSNR is obtained by the SSNR of the output signal minus the segmental SNR of the input signal.

Fig. 3b presents outcomes of subjective speech quality and intelligibility test. In the experiments, we adopt informal speech test. Five noisy speech sentences were processed by two methods (conventional spectrum subtraction (CSS), and the proposed method in this paper (PM) respectively). Five listeners were invited to test to assess quality and intelligibility of enhanced speech sentences for three times. As it shown on Fig.3 the best results in SNR improvement of noisy speech were obtained by proposed algorithm on AWGN noise. The difference in SNR improvement between conventional SS and proposed method increase considerably with the decreasing input signal SNR conditions. For car and office noise in 0dB SNR level the difference in SNR improvement between two concerned algorithms is 1.5-2dB. However the difference in speech quality and intelligibility enhancement between conventional SS and proposed algorithm is considerable.

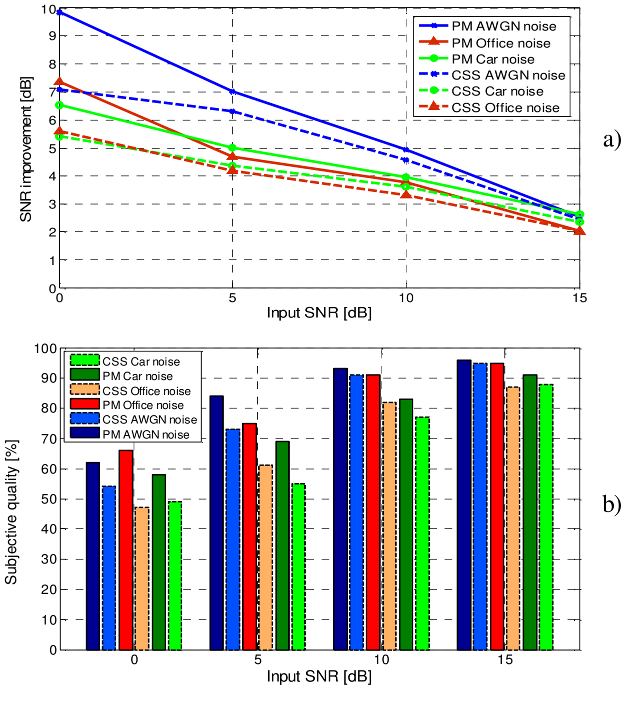

Figure 3. SNR improvement (a) and speech quality assesment (a) of conventional SS (CSS) and proposed method (PM) for various noise types and input signal SNR levels

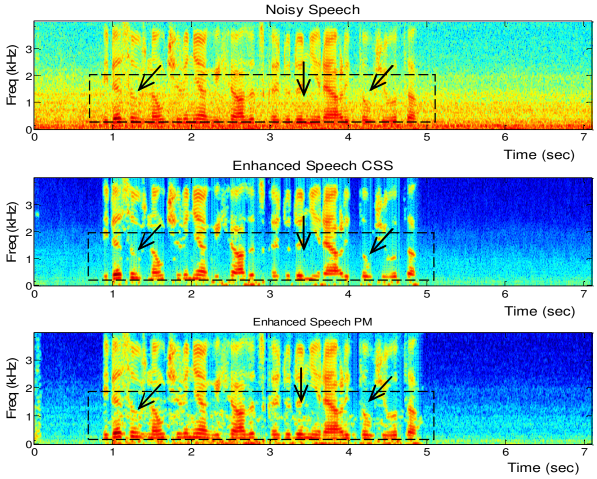

Figure 4. Spectrograms of speech enhanced by conventional SS (CSS) and proposed method (PM) . Car noise in 0dB SNR level.

In low SNR conditions conventional SS algorithm results in annoying musical noise and some speech distortions. In spite of high factor of SNR improvement, enhanced signal has a lower perceived quality and lower information content, than the original noisy signal. Proposed algorithm even in low input signal SNR levels result in high results of SNR improvement as well as in substantial speech quality enhancement with minimal distortions (see Fig.4).

VI. Conclusion

In this paper speech enhancing method based on improved SS algorithm was introduced. For effective noise reduction with minimal distortion proposed algorithm takes in account perceptual aspects of human ear. It can be seen from the experimental results that proposed method effectively reduces background noise in comparison with commonly used SS algorithm. Proposed method results in greater improvement of SNR and considerably improvement of perceptual speech quality in compassion to conventional spectral subtraction method.

References

[1] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise” in Processing of international Conference of Acoustic, Speech and Signal Processing,1979, pp. 208-211.

[2] T. Esch, P. Vary, ”Efficient musical noise suppression for speech enhancement system” IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 2009, pp. 4409 - 4412.

[3] E. Dong, X. Pu, “Speech denoising based on perceptual weighting filter” 9th International Conference on Signal Processing , Leipzig, Germany, 2008, pp 705-708.

[4] V. Prasad, R. Sangwan et al., “Comparison of voice activity detection algorithms for VoIP”, proc. of the Seventh International Symposium on Computers and Communications, Taormina, Italy, 2002, pp. 530-532.

[5] K.Sakhnov, E.Verteletskaya, B. Šimák, “Dynamical Energy-Based Speech/Silence Detector for Speech Enhancement Applications”. In World Congress of Engeneering 2009 Proceedings. Hong Kong:, 2009, pp.801-806.

[6] S. Ogata, T.Shimamura, ”Reinforced spectral subtraction method to enhance speech signal”, Proceedings of IEEE International Conference on Electrical and Electronic Technology, 2001, vol 1, pp 242 – 245.

[7] P. Pollák, ”Speech signals database creation for speech recognition and speech enchancement applications” [associate professor innagural dissertation] CTU in Prague, FEE, Prague, 2002.