IMPROVED SPECTRAL SUBTRACTION FOR SPEECH ENHANCEMENT

Y. Malca and D. Wulich
Department of Electrical & Computer Engineering, Ben-Gurion University of the Negev.
Beer-Sheva 84105, POB 635, Israel.
Tel: ++972-7-461537, Fax: ++972-7-472949, e-mail: dov@bguee.bgu.ac.il

ABSTRACT

The spectral subtraction approach has become almost standard in speech enhancement because it is relatively easy to understand and implement.

The major drawback of the spectral subtraction method is that it leaves residual noise with annoying noticeable tonal characteristics referred to as *musical noise*. For low SNR the perceived effect of the "musical noise" is close to that of the additive noise.

In the present work we propose to reduce the musical noise by applying the output of a standard spectral subtractor to a constrained high order notch filter which suppresses the "musical noise". The filtration process distorts the speech signal. It is possible to reduce the level of distortion if the speech signal is preprocessed properly before it is contaminated by the noise.

It will be demonstrated that the proposed method is superior to the standard spectral subtraction specially for low SNR. A comprehensive listening test indicated that for segmental SNR= -12dB, 77% of the listeners strongly preferred the proposed approach over the usual spectral subtraction approach.

1. INTRODUCTION

The spectral subtraction approach has become almost standard in speech enhancement. It has gained popularity because it is relatively easy to understand and implement, i.e., it only requires DFT of the noisy signal, application of a gain function, and inverse DFT [1,2,3].

In the Standard Spectral Subtraction (SSS) estimation approach the power spectral density of the clean signal is estimated by subtracting an estimate of the power spectral density of the noise process from the estimate of the power spectral density of the degraded signal [1,2,3]. The estimation is performed on a frame-by-frame basis, where each frame consists of T=10-30msec of speech samples. A block diagram of the SSS approach is shown in Fig. 1. The noisy vector z_t is given by $z_t = y_t + v_t$, where y_t denotes a K-dimensional vector of the clean signal, v_t denotes a K-dimensional vector of the noise process, and y_t and v_t are assumed uncorrelated.

The method of SSS is based on the following stages:
(1) Discrete Fourier Transform (DFT) of order K on the degraded signal z_t;
(2) Subtraction of the estimate of the power spectral density of the noise, which is normally estimated from a portion of the noisy signal during which speech is absent;
(3) Half-wave rectification: the negative result of subtracting is replaced by zero;
(4) The phase of the noisy signal is combined with the result of half-wave rectification and then the Inverse DFT is applied.

The major drawback of the SSS method is that it leaves residual noise with annoying noticeable tonal characteristics referred to as *musical noise*. For low SNR[1] (about -5dB) the perceived effect of the musical noise is close to that of the additive noise [1,2].

The elimination of musical noise has been considered in a number of publications.

In [1] a few modifications of the SSS method described above are proposed which may be used to reduce the effects of the musical noise.
(1) *Magnitude averaging*: The noisy vector available at the input is replaced by an averaged vector obtained as an average of M neighboring vectors. The total time duration of the M vectors is less then 38msec. The spectral subtraction is performed on the average vector. Such averaging reduces the spectral error between the average noise and the estimate value and consequently the level of the musical noise is also reduced. This averaging process, however degrades the quality of the speech signal.
(2) *Residual noise reduction*: The musical noise may be reduced if the vector at the output of the half-wave

[1] When the speech signal is considered the value of SNR depends on a specific record of the speech. In this paper we will use a Segmental SNR (SSNR) which is less sensitive to the specific record of the speech signal used and is defined as: $SSNR \stackrel{\Delta}{=} \frac{1}{M}\sum_{m=1}^{M} 10\log_{10}\frac{P_s(m)}{P_n(m)}$, where $P_s(m)$ and $P_n(m)$ are the powers of the speech signal and noise in the segment (frame) number m.

rectification process is replaced by another vector which contains, at each frequency, the minimum magnitude chosen from two neighboring frames (vectors). The duration of the two neighboring vectors must be less than 30msec. The drawback of this method is that the speech at this duration is only partially stationary which leads to magnitude errors and therefore to degradation of the speech quality.

(3) *Additional signal attenuation during silence intervals*: The musical noise is a random process which is active all time, while the speech is active about 40%-60% of the time. Therefore by attenuating the enhanced signal $\hat{y}_t$ during the silent intervals the average power of the musical noise may be reduced by 40%-60%. Such attenuation of the nonspeech frames (intervals) however, leads to a "flash of silence" between the speech. Unfortunately, the perceived effect of such a prepared signal is worse than if signal has musical noise all the time.

The spectral subtraction method is based on the assumption that the spectrum of the noise is estimated during the nonspeech frames. The problem how to recognize the nonspeech frames is not discussed in this paper, we assumed that we have this information.

In this work we propose a method which can reduce the effect of the musical noise below the value which can obtained if the spectrum of the noise is exactly known.

The musical noise occurs when the difference between the sample spectrum of the noise and the estimate of the power spectral density of the noise process, at discrete frequencies determined by the frame size, is a positive value. Consequently, the musical noise generated within one frame can be considered as a sum of sine waves with known and fixed frequencies and with random amplitudes (included zero) and phases, which remain unchanged for the considered frame. As a result the musical noise considered within a time interval which includes a large number of frames, consists of a multitone signal having a randomly modulated complex envelope. As expected, the spectrum of the musical noise considered over an infinite time interval (a large time interval in practice) is concentrated around known frequencies, namely multiples of the reciprocal of the frame time. Consequently the power of the musical noise can be significantly reduced by applying the "spectrally subtracted" signal (i.e., the enhanced speech signal+musical noise) to a multiple notch filter having notches exactly at frequencies of interest.

A high-order Constrained Notch Filter (CNF) [4,5,6] is proposed here as such a notch filter. The transparency of a such filter is controlled by a so-called *transparency coefficient*. The transparency coefficient has to be chosen as a trade-off of two opposite requirements: on the one hand the CNF must posses high transparency in order not to distort the speech signal; on the other hand the filter must posses low transparency to decrease the transient response and then to reject the musical noise effectively. The first requirement may be sufficiently reduced if the clean signal can be adequately preprocessed prior to its degradation by the noise.

It will be demonstrated that the proposed method, which consists of preprocessing the clean speech, standard spectral subtraction and multiple notch filtering, is superior to the standard spectral subtraction specially for low SNR. A comprehensive listening test indicated that 77% of the listeners strongly preferred the proposed approach over the usual spectral subtraction approach.

2. MUSICAL NOISE PHENOMENON.

As stated the major drawback of the SSS method is that musical residual noise is obtained. It happen when the difference between the sample spectrum of the noise v_t and the estimate of the power spectral density $\left|S_v(\theta)\right|^2$ of the noise process at a normalized frequencies $\{\theta_{k_i}\}_{k_i=1}^{L}$ is a positive value. The subset $\{\theta_{k_i}\}_{k_i=1}^{L}$ is taken at random, from the set of all possible normalized frequencies $\theta_k = \{2\pi k/K\}_{k=0}^{K/2}$. As a result, the whole enhanced signal $\hat{y}_\lambda$ for $-\infty < \lambda < \infty$ is accompanied by a sum of sine waves whose amplitudes and phases are random but whose frequencies are taken from the known a priori set of frequencies. The number *L*, however is random and varies from frame to frame. Consequently the musical noise for $-\infty < \lambda < \infty$ can be represented by the following formula:

$$n_\lambda = \sum_{m=-\infty}^{\infty} \{ \sum_{k_i=1}^{L_m} A_{k_i} \cos[\theta_{k_i}.(\lambda - mT) + \varphi_{k_i}] \} \qquad (1)$$

where L_m is number of tones within the frame *m* and A_{k_i}, φ_{k_i}; $k_i = 1,2,\ldots,L_m$ are their amplitudes and phases respectively.

3. ELIMINATION OF THE MUSICAL NOISE BY USING A HIGH-ORDER CONSTRAINED NOTCH FILTER

The idea of eliminating the musical noise is based on filtering this noise by applying $\hat{y}_\lambda$ to the $K/2$-th order Constrained Notch Filter-CNF which has notches exactly at the normalized frequencies θ_k, $k = i_o,\ldots,i_1$, where i_o denotes the first frequency and i_1 denotes the last frequency to be rejected. The transfer function of the CNF is [4,5,6] :

$$H(z) = \prod_{i=i_0}^{i_1} \left[\frac{1 - 2\cos\left(\frac{2\pi i}{K}\right) z + z^2}{1 - 2\rho\cos\left(\frac{2\pi i}{K}\right) z + \rho^2 z^2} \right] \qquad (2)$$

where the parameter ρ controls the transparency of the filter. For $\rho \to 1$ an ideal transparency can be obtained at the expense of long transient response.

A proper choice of ρ is very important because of two opposing demands :

(1) to choose large ρ (close to 1) for good transparency, i.e., to not distort the speech signal;

(2a) to choose small ρ, for short transient response time of the filter which should be much less than T (or $K = T \cdot f_s$, where f_s is the sampling frequency); or equivalently

(2b) to choose small ρ for effective rejection of the musical noise which is in fact a linearly modulated multitone signal.

It is possible, however, to overcome the distortions introduced by CNF with small ρ if the clean speech signal (if available) is applied to a filter having the transfer function $H^{-1}(z)$ prior to its degradation by the noise. The inverse filter has to compensate for the signal attenuation at the frequencies, at which $H(z)$ has notches. The inverse filter does not exist and consequently the following approximated version of $H^{-1}(z)$, will be considered:

$$G(z) = [1 - \alpha \cdot H(z)]\frac{1}{\alpha} \quad ; \quad 0 < \alpha < 1 \qquad (3)$$

where the parameter α controls the gain of the filter at the notches of $H(z)$. Fig. 2 shows the block scheme of the proposed method, named Musical Noise Rejection (MNR).

4. EXPERIMENTAL RESULTS

An experiment was conducted to evaluate the intelligibility and quality of the MNR and to compare it to the SSS.

4.1 Method

4.1.1 Subjects

Eighteen normal-hearing adults, with ages between 20 and 35 yr., participated in this study. The sentences were heard by means of earphones on both ears.

4.1.2 Test material

Two different kinds of test material were used. The first included 12 lists of everyday sentences. Each list consisted of 10 sentences: 5 statements and 5 questions of varying length (3 to 14 words). The number of total words per list varied from 50 to 70 words/list. In each list each sentence was related to a different topic, such as, family, food, sport, politics, etc. This group of sentences was used for the quality test.

The second kind of test material included 10 lists of single-syllable words. Each list consisted of 10 single-syllable words. This group of single-syllable words was used for the intelligibility test.

The speech signal was sampled at f_s=9.6kHz and processed by a digital computer.

4.1.3 Procedure

Three different processors have been tested and compared: Processor No. 1**:** speech signal with noise (without any processing); Processor No. 2**:** SSS; Processor No. 3**:** MNR.

As a result of large number of preliminary and informal listening tests the following parameters have been chosen as optimal for K=128 (T=13.3msec): $\rho = 0.98$, i_o=4 and i_1=37, i.e., the speech signal is processed by both $G(z)$ and $H(z)$ within the range f_o=300Hz, f_1=2775Hz, where $f_n = i_n \cdot f_s / K$, n=0,1.

Intelligibility test. Every subject listened to different single-syllable words, processed by the three processors for two different values of SSNR: -10dB and -16dB. The subject had to write down the words.

Quality test. Every subject listened to different everyday sentences, processed by the three processors for SSNR=-12dB. Subjects were asked to provide answers to three questions.

(a) what is the *level of concentration* needed to understand the text;

(b) is the quality of the speech *natural* or *metallic*;

(c) which processor she/he *personally preferred*;

4.2 Results

4.2.1 Intelligibility test

The results of the intelligibility test show that the intelligibility of the speech signal processed by the three above-mentioned processors for different SSNR was almost the same, with a small (6%) preference for Processor No.3. The standard deviation of these results is about 8%.

4.2.2 Quality tests

The results related to question (a) are shown in Fig. 3. It is clearly seen that the MNR (Processor No.3) requires the lowest concentration efforts: 11 listeners (out of 18) indicated that lowest concentration efforts are needed when the MNR is used. As a result of question (b), all 18 listeners pointed out that the speech signal processed by the MNR has a metallic quality. The answers to question (c) are shown in Fig. 4. The MNR was preferred by 14 (77%) listeners, the SSS was preferred by 3 (17%) listeners while only one listener (6%) preferred the unprocessed, noisy speech.

4.3 Conclusion

A method to reduce the musical noise which appears in the standard spectral subtraction method is described. The musical noise is reduced by applying the constrained high order notch filter to the output of a standard spectral subtractor. The filtration process distorts the speech signal. It is possible to reduce the level of distortions if the speech signal is preprocessed properly before it is contaminated by the noise. It is assumed, therefore that a clean speech signal is available prior its degradation by the noise.

This method (MNR) has been checked and compared to the unprocessed noisy signal and to the SSS method for low values of SSNR, namely within a range: -16dB to -10dB. A comprehensive listening test shows that the MNR method does not reduce the intelligibility of the processed speech signal as compared to both the SSS method and the unprocessed speech. As a matter of fact the intelligibility obtained was even slightly higher.

The quality test, based on three questions, shows that the MNR method is superior to the SSS and to the unprocessed speech signal.

Two drawbacks, however related to the MNR method are observed:
(1) it assumed that the clean speech signal is available prior the speech is contaminated by noise;
(2) the speech processed by the MNR method has a metallic quality.

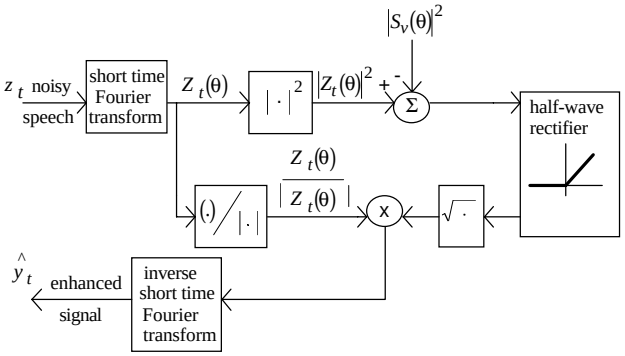

Fig 1. The block scheme of the SSS

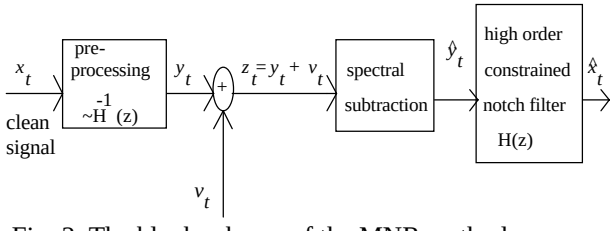

Fig. 2. The block scheme of the MNR method

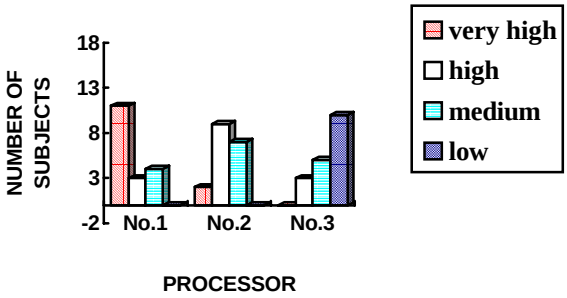

Fig 3. Level of concentration

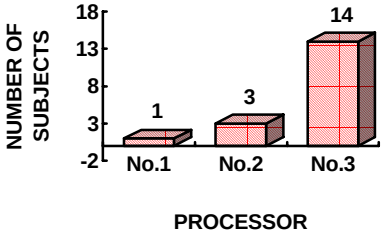

Fig 4. Personal preference

5. REFERENCES

[1] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Tran. Acoust., Speech Signal Processing*, vol.27, pp. 113-120, Apr. 1979 .

[2] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, No.10, October 1992 .

[3]. Y. Ephraim and D. Malah, "Speech enhancement using minimum mean square error log-spectral amplitude estimator," *IEEE Tran. Acoust, Speech Signal Processing*, vol. ASSP-32, No. 6, pp.1109-1121, Dec. 1984.

[4] T. Sang Ng, "Some aspects of an adaptive digital notch filter with constrained poles and zeros" *IEEE Tran. Acoust., Speech Signal Processing*, vol. ASSP-35, No. 2, Feb. 1987.

[5] A. Nehorai, "A minimum parameter adaptive notch filter with constrained poles and zeros," *IEEE Tran. Acoust., Speech Signal Processing*, vol. ASSP-33, pp. 983-996, Aug. 1985.

[6] D. Wulich, E.I. Plotkin and M.N.S. Swamy, "Constrained notch filtering of nonuniformly speech samples for enhancement of an arbitrary signal corrupted by a strong FM interference," *IEEE Tran. Signal Processing*, vol. 39, No. 10, Oct. 1991.