A CROSS-CORRELATION TECHNIQUE FOR ENHANCING SPEECH CORRUPTED WITH CORRELATED NOISE

Yi Hu , Mukul Bhatnagar and Philip C. Loizou

Department of Electrical Engineering
University of Texas at Dallas
Richardson, Texas 75083-0688
{yihuyxy, mukul, loizou}@utdallas.edu

ABSTRACT

Most speech enhancement techniques do not perform very well in correlated or colored noise, as they assume that noise and speech are not correlated. In this paper, we propose a method, based on spectral subtraction, which takes into account possible correlation between noise and speech. Objective measures showed that the proposed method outperformed the power spectral subtraction method resulting in better speech quality and reduced levels of musical noise. Further enhancements in speech quality were obtained by applying a perceptual weighting function (estimated using a psychoacoustics model) that was designed to minimize noise distortion.

1. INTRODUCTION

Reducing noise in corrupted speech remains an important problem and has a broad range of applications, most of which are driven by the explosive growth of mobile communications. Numerous approaches have been proposed for speech enhancement, with the spectral subtraction method being one of the most popular, due to its relatively simple implementation and computational efficiency. It is well suited for real time applications since it works quite well as a single-channel enhancement technique and is capable of handling the nonstationarity of noise to some extent.

Since Boll's original paper [1] many variations of spectral subtraction have been proposed, including the power spectral subtraction with oversubtraction [2], perceptually based enhancement [3] and time-frequency filtering [4]. Most of these techniques make the assumption that noise and speech are uncorrelated. However, in the real world, noise might not always be uncorrelated with speech. For instance, the cafeteria noise (multitalker babble) is highly correlated with speech. Because of the above assumption, spectral subtractive type algorithms do not perform as well with correlated noise. Many other techniques [5][6] have been proposed recently for suppressing correlated noise but they are computationally more expensive compared to spectral subtraction.

In this paper, we propose a speech enhancement method, based on spectral subtraction, which takes into consideration possible correlation between speech and noise. The proposed method subtracts from the corrupted speech not only an estimate of the noise, but also an estimate of the cross-correlation between noise and speech. Enhanced speech produced by the proposed method was used to estimate masking thresholds, which were then used to design a perceptual weighting function. Speech quality improved with reduced levels of residual noise when the perceptual weighting filter was applied.

The paper is organized as follows. In Section 2, we derive the proposed cross-correlation approach, and in Section 3 we derive the perceptual weighting filter based on a psychoacoustical model. Section 4 describes our implementation, and Section 5 presents our results.

2. CROSS-CORRELATION APPROACH

Let the corrupted speech signal $y(n)$ be represented as

$$y(n) = s(n) + d(n) \tag{1}$$

where $s(n)$ is the clean speech signal and $d(n)$ is the noise signal. In the frequency domain, we have

$$Y(k) = S(k) + D(k) \tag{2}$$

The power spectrum of $Y(k)$ can be computed from (2) as follows

$$|Y(k)|^2 = |S(k)|^2 + |D(k)|^2 + S(k)\cdot D^*(k) + S^*(k)\cdot D(k) \tag{3}$$

The terms $|D(k)|^2$, $S(k)\cdot D^*(k)$ and $S^*(k)\cdot D(k)$ cannot be obtained directly and are approximated as $E[|D(k)|^2]$, $E[S^*(k)\cdot D(k)]$ and $E[S(k)\cdot D^*(k)]$, where $E[\cdot]$ denotes the expectation operator. Typically, $E[|D(k)|^2]$ is estimated during the silence periods, and we denote it by $|\hat{D}(k)|^2$. If we assume that $d(n)$ is zero mean and uncorrelated with $s(n)$, then the terms $E[S^*(k)\cdot D(k)]$ and $E[S(k)\cdot D^*(k)]$ reduce to zero. However, if speech and noise are correlated, then we can no longer neglect those cross terms, which represent the cross correlations ($r_{sd}(m)$ and $r_{ds}(m)$) between $d(n)$ and $s(n)$. Unfortunately, we cannot estimate these cross-correlations since we do not have access to $s(n)$. But, since we have access to the

corrupted signal $y(n)$, we can get an estimate of the cross correlation $r_{sd}(m)$ (or r_{ds}) by computing the cross correlation between $y(n)$ and $d(n)$, i.e., $r_{yd}(m)$:

$$r_{yd}(m) = r_{sd}(m) + r_{dd}(m) \tag{4}$$

Note that $r_{yd}(m)$ contains the desired cross correlation $r_{sd}(m)$ as well as the autocorrelation of the noise signal, which in the frequency domain is given by $|D(k)|^2$ and can be lumped with the same term in Eq. (3). By including a short-time (instantaneous) estimate of the cross correlation between $y(n)$ and $d(n)$ we get the proposed cross-correlation spectral subtraction (CCSS) approach:

$$|\hat{S}(k)|^2 = \begin{cases} |Y(k)|^2 - \alpha|\hat{D}(k)|^2 - \delta|Y(k)|\cdot|\hat{D}(k)| & \text{if } |Y(k)|^2 > \alpha|\hat{D}(k)|^2 \\ \beta|\hat{D}(k)|^2 & \text{else} \end{cases} \tag{5}$$

where α is the subtraction factor [2], β is the spectral floor parameter [2], and δ is the cross-correlation coefficient which provides an estimate of the correlation between corrupted speech and noise in the current window frame. The parameter δ is calculated as follows:

$$\delta = \left| \frac{\chi_{y\cdot d} - \mu_y \cdot \mu_d}{\sigma_y \cdot \sigma_d} \right| \tag{6}$$

where

$$\chi_{y\cdot d} = \frac{1}{(N/2)} \sum_k |Y(k)| \cdot |\hat{D}(k)|$$

$$\mu_y = \frac{1}{(N/2)} \sum_k |Y(k)|, \quad \mu_d = \frac{1}{(N/2)} \sum_k |\hat{D}(k)|$$

$$\sigma_y^2 = \frac{1}{(N/2)} \sum_k \{|Y(k)| - \mu_y\}^2, \quad \sigma_d^2 = \frac{1}{(N/2)} \sum_k \{|\hat{D}(k)| - \mu_d\}^2$$

for $0 \le k \le N/2$, N being the FFT size. The value of δ, determines the factor of subtraction, and is proportional to the degree of correlation between speech and noise. Note that Eq. (5) reduces to the original power spectral subtraction method [2] when δ =0.

Informal listening tests showed that the quality of speech processed through the cross-correlation approach was better than the quality produced by the power spectral subtraction approach. The cross-correlation method also reduced the musical noise, commonly found in speech synthesized by the power spectrum subtraction approach. Comparative results as well as spectrograms of enhanced speech are presented later in Section 5.

To further enhance the speech quality, we fed the corrupted speech signal through a perceptual weighting filter, which was estimated using a psychoacoustics model (Fig. 1). The speech output of the cross-correlation method was used to compute the masking thresholds. The cross-correlation method provides a good estimate of the clean signal, which is critical for accurate computation of the masking thresholds. The derivation of the perceptual weighting filter is presented in the next section.

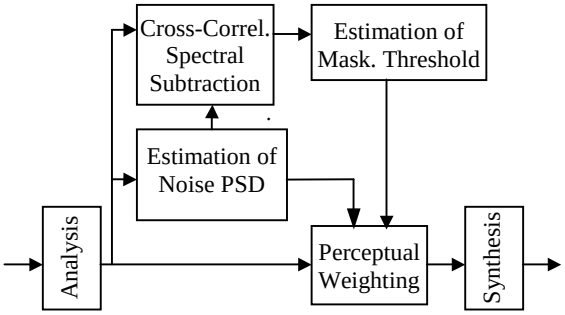

Figure 1 Overview of the proposed enhancement scheme.

3. PERCEPTUAL WEIGHTING

Let $\hat{S}(k) = G(k)\cdot Y(k)$ be the enhanced speech spectrum and *G(k)* the perceptual weighting filter. Forming the error spectrum between the spectrum of the clean signal and the estimated (enhanced) spectrum we get:

$$\begin{aligned} E(k) &= \hat{S}(k) - S(k) = G(k)\cdot Y(k) - S(k) \\ &= [G(k)-1]\cdot S(k) + G(k)\cdot \hat{D}(k) \end{aligned} \tag{7}$$

The power spectrum of the error signal is then given by:

$$\begin{aligned} &|E(k)|^2 = \\ &|G(k)-1|^2 \cdot |S(k)|^2 + |G(k)|^2 \cdot |\hat{D}(k)|^2 \\ &+ (G^*(k)-1)S^*(k)G(k)\hat{D}(k) + S(k)(G(k)-1)G^*(k)\hat{D}^*(k) \end{aligned} \tag{8}$$

The first term in the above equation describes the speech distortion caused by the spectral weighting, the second term describes the noise distortion and the last two terms describe the noise-speech distortion created by the fact that speech and noise are correlated. If we assume that the joint noise-speech distortion is small because it was minimized in the first stage (Fig. 1) by the cross-correlation method, then we are left with the noise and speech distortions. One can then compute a spectral weighting function *G(k)* such that the noise and speech distortions fall below the masking threshold. However, as shown in [7], a complete masking of both distortions cannot be guaranteed and one must settle for the best trade-off between the two distortions. In this study, 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. We therefore chose the weighting function $G(k)$ that satisfied the following criteria:

$$\begin{cases} |G(k)|^2 \cdot \left|\hat{D}(k)\right|^2 \le T(k) \\ 0 \le |G(k)| \le 1 \end{cases}$$

where *T(k)* is the masking threshold. Solving for $G(k)$ we get

$$G(k) = \min\left(\sqrt{\frac{T(k)}{\left|\hat{D}(k)\right|^2}}, 1 \right) \tag{9}$$

as the perceptual weighting filter. The masking threshold *T(k)* was computed as in [8] using the enhanced speech signal estimated by the cross-correlation method.

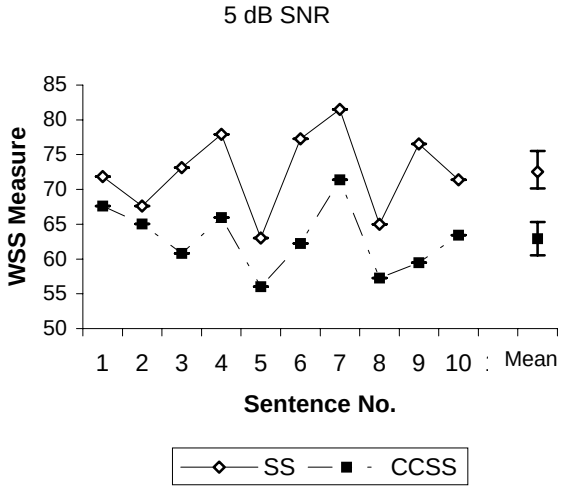

Figure 2 Comparison between the proposed cross-correlation method (CCSS) and the power spectrum subtraction method (SS) using the Weighted Spectral Slope (WSS) measure for 10 sentences from the HINT database at 5 dB SNR.

4. IMPLEMENTATION

A 20-ms Hamming window was used for analysis with 50% overlap. The final enhanced speech was reconstructed by computing the inverse FFT of the estimated spectral amplitude $|\hat{S}(k)|$ combined with the phase of the corrupted speech, and using the standard overlap-and-add method. The subtraction factor α in Eq. (5) was computed as per [2] and was a function of the segmental SNR. The spectral floor parameter β was set to 0.002.

We incorporated a statistical model-based voice activity detection method [9] to detect non-speech frames. This method computed the likelihood ratio of two states, speech absent and speech present, as follows:

$$\frac{1}{N}\sum_{k=0}^{N-1}\left(\frac{|Y(k)|^2}{|\hat{D}(k)|^2}-\log\frac{|Y(k)|^2}{|\hat{D}(k)|^2}-1\right)\begin{array}{c}\textit{speech present}\\ >\\ <\\ \textit{speech absent}\end{array}\eta \qquad (10)$$

where η is a preset threshold. When speech was absent in frame *i*, the noise spectrum was updated according to the following formula:

$$|\hat{D}_i(k)|^2=\lambda_d\cdot|\hat{D}_{i-1}(k)|^2+(1-\lambda_d)\cdot|Y_i(k)|^2 \qquad (11)$$

with $\lambda_d = 0.9$.

5. RESULTS

For evaluation purposes, sentences from the HINT (Hearing in Noise Test) database [10] spoken by a male talker were used. The HINT database is commonly used for speech intelligibility studies and it contains lists of sentences, which were designed to be equally intelligible in noise. Ten sentences from the HINT database were downsampled to 8 kHz and used for testing.

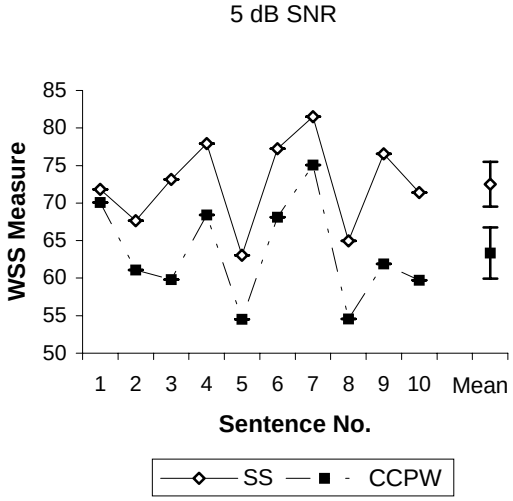

Figure 3 Comparison between the proposed cross-correlation method followed by perceptual weighting (CCPW) and the power spectrum subtraction method (SS) using the WSS measure for 10 sentences from the HINT database at 5 dB SNR.

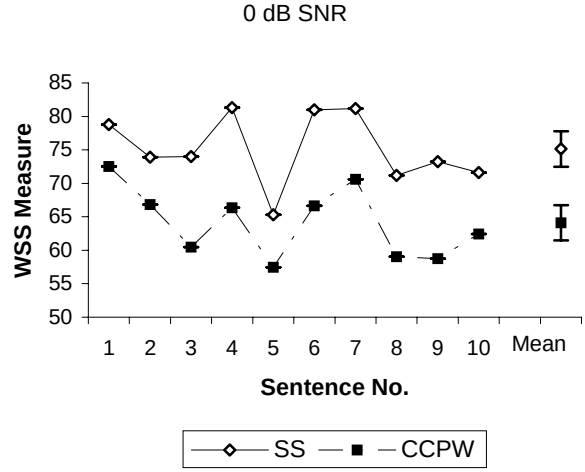

Figure 4 Comparison between the proposed cross-correlation method followed by perceptual weighting (CCPW) and the power spectrum subtraction method (SS) using the WSS measure for 10 sentences from the HINT database at 0 dB SNR.

Speech-shaped noise was added at an SNR of 5 dB and 0 dB. The speech-shaped noise was computed from the long-term spectrum of all the sentences in the database, and matched the spectral characteristics of the male speaker.

For evaluation, the Weighted Spectral Slope (WSS) measure, proposed by Klatt, was used as the objective speech quality measure [11]. We chose the WSS measure, over the SNR and Itakura-Saito measures, because it has a reasonably high correlation (ρ =0.74) with subjective speech quality [11]. Figure 2 shows the comparative results between the proposed cross-correlation approach and the power spectrum subtraction approach for all the 10 sentences tested.

As can be seen, the cross-correlation approach yielded consistently better results (lower WSS values) than the power spectrum subtraction approach. Figures 3 and 4 show the comparison between the psychoacoustically motivated approach (which incorporates the cross-correlation method) and the power spectrum subtraction method for 5 dB SNR and 0 dB SNR respectively. Again, the proposed approach outperformed consistently the power-spectral subtraction method for all sentences.

Informal listening tests confirmed that the proposed speech enhancement method resulted in comparatively much better sound quality and substantially reduced levels of "musical" noise compared to the power spectrum subtraction method (see Fig. 5).

6. CONCLUSIONS

A speech enhancement method was proposed for enhancing speech corrupted with correlated noise. Unlike other speech enhancement techniques which assume that speech and noise are uncorrelated, the proposed method takes into account possible correlation between speech and noise. A perceptual weighting filter that minimized the noise distortion was also developed based on a psychoacoustics model. Results showed that the proposed method outperformed the power-spectrum subtraction method in enhancing speech corrupted with correlated noise.

REFERENCES

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

[2] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 208-211, Apr. 1979.

[3] N. Virag, " Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech and Audio Processing, pp. 126-137,vol. 7, March 1999.

[4] G. Whipple, "Low residual noise speech enhancement utilizing time frequency filtering," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, pp I/5-I/8, vol. 1, 1994.

[5] J.D. Gibson, B. Koo and S.D. Gray, "Filtering of colored noise for speech enhancement and coding," IEEE Trans. Signal Processing, pp. 1732-1742, vol. 39, Aug. 1991.

[6] U. Mittal, N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," IEEE Trans. Speech Audio Proc., pp. 159-167, vol. 8, March 1999.

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

[8] J.D. Johnston, "Transform Coding of Audio Signals using Perceptual Noise Criteria," IEEE Journal on Selected Areas of Communications, Vol. 6, No. 2, pp. 314-323, Feb. 1988.

[9] J. Sohn, N.S. Kim and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 365-368, 1998.

[10] Nilsson, M., Soli, S. and Sullivan, J. "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise, " J. Acoust. Soc. Am., vol. 95, pp. 1085-1099, 1994.

[11] S.R. Quackenbush, T.P. Barnwell,III, M.A. Clements, Objective Measures of Speech Quality. Englewood Cliffs, NJ: Prentice-Hall, 1988.

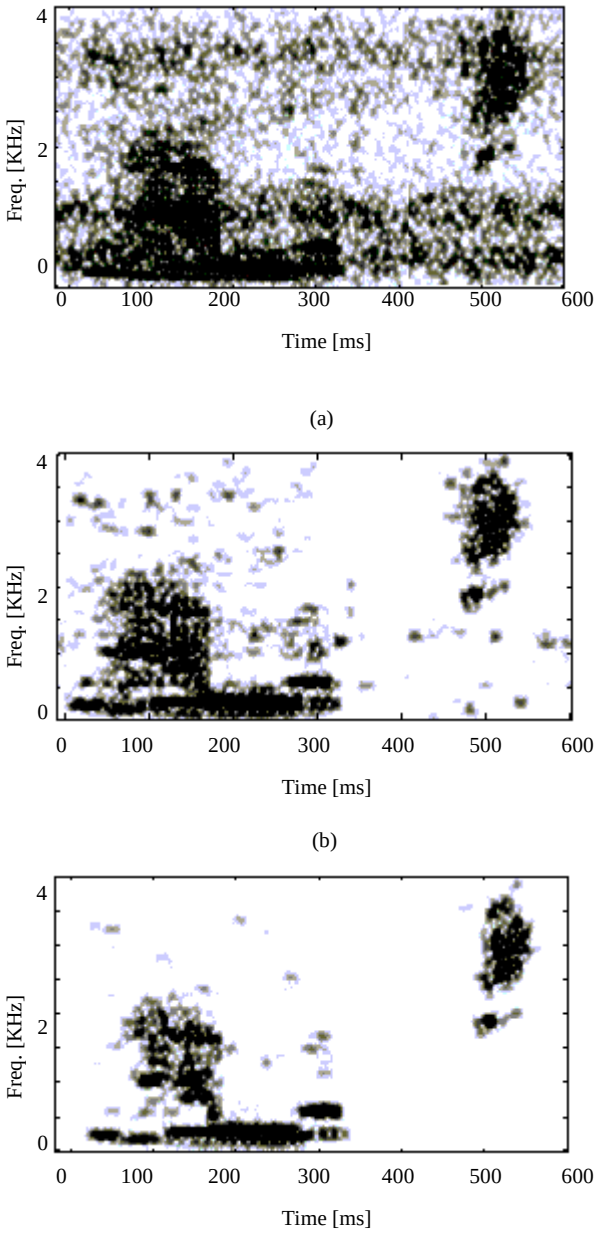

Figure 5. Spectrograms of the sentence "the match ..." (a) Speech corrupted with speech-shaped noise at 5 dB SNR (b) enhanced speech obtained using power spectrum subtraction (c) enhanced speech obtained by the cross-correlation method in conjunction with the perceptual weighting filter.