Combination of Two-Channel Spectral Subtraction and Adaptive Wiener Post-Filtering for Noise Reduction and Dereverberation

Matthias Dörbecker Stefan Ernst
Institute of Communication Systems and Data Processing
Aachen University of Technology, Templergraben 55, 52056 Aachen, Germany
Phone: +49 241 806963, Fax: +49 241 8888186
e-mail: matthias@ind.rwth-aachen.de

ABSTRACT

In this contribution a novel structure for the enhancement of speech signals disturbed by acoustic noise is presented which is based on Spectral Subtraction. The Spectral Subtraction technique is combined with a novel estimator for the noise power spectrum which takes advantage of the employment of a second microphone. Due to the extension to a two-microphone system the Spectral Subtraction can be used to reduce realistic, nonstationary noise sources. Additionally, the performance of the system is further improved by the application of a post filter adapted according to Wiener filter techniques. As a result, the proposed speech enhancement system provides a significant suppression of noise in realistic situations as well as a reduction of room reverberation.

1 INTRODUCTION

Spectral Subtraction is a well-known technique to enhance the quality of speech signals disturbed by acoustic background noise. The fundamental components of a noise reduction system based on Spectral Subtraction are an estimator of the power spectrum of the disturbing noise and a noise suppression rule to reconstruct the spectral magnitudes of the speech signal. The principal problem of many conventional noise suppression rules [1, 2], i.e. the presence of musical noise artefacts in the processed signal, can be kept small, if the noise suppression is done according to the *Minimum Mean Square Error Short-Time Spectral Amplitude Estimator* (MMSE log-STSA estimator) proposed by Ephraim and Malah [3, 4]. On the other hand, most of the known techniques to estimate the power spectrum of the disturbing noise fail in realistic situations, since they assume a highly stationary noise.

An alternative approach to enhance noisy speech is the application of several microphones to take into account the spatial distribution of speech- and noise sources. The techniques we are referring to [5, 6, 7, 8] aim at the reduction of diffuse noise sound fields as well as room reverberation.

In this contribution Spectral Subtraction is extended to a two-microphone system, which takes advantage of a novel two-channel estimator of the noise power spectrum. For the derivation of the novel estimator the diffuse character of the noise sound field rather than its stationarity has been the principal assumption. Therefore, the proposed structure can be successfully employed for the suppression of noise in many realistic situations, e.g. babble noise in an office room or in a cafeteria, noise of the engine in a car, etc.

This paper is organized as follows: In Section 2 the novel two-channel estimator for the noise power spectrum is derived. Applying the adaptive post-filter described in Section 3, the performance of the two-channel Spectral Subtraction can be further improved. As an example, we describe the application of the proposed structure to an electronic hearing aid in Section 4. Finally, the results of informal listening tests are presented to evaluate the performance of the noise reduction system.

2 DERIVATION OF THE NOISE POWER SPECTRUM ESTIMATOR

For the derivation of the two-channel noise power spectrum estimator it is assumed that the distance between the speaker and the microphones is such that the microphones pick up a high portion of the direct sound. Therefore, the speech signals received by the two microphones are mutually correlated, i.e. the *magnitude-squared coherence* (MSC) [9] between the two microphone signals is close to one [10].

On the other hand, we suppose that the noise can be characterized as a diffuse sound field. This assumption is suitable for spatial distributed noise sources as well as in reverberant rooms when the distance of the noise source to the microphones is large, so that the reflected sound dominates. It is well-known that a diffuse sound field results in a MSC which is – except for low frequencies – close to zero [6, 10].

Since the coherence of the speech signals received by the microphones is close to one, the acoustic situation can be modeled by two transfer functions $H_1(\Omega)$ and $H_2(\Omega)$ between the speaker and the microphones, where Ω denotes the normalized frequency. Neglecting the high coherence of the diffuse noise sound field at low frequencies, the noise received by the microphones can be represented by two additive, uncorrelated noise sources with the short-time Fourier spectra $N_1(\Omega)$ and $N_2(\Omega)$, respectively, and the cross power spectrum $\Phi_{N_1N_2}(\Omega) \approx 0$. Therefore, the short-time Fourier spectra $X_1(\Omega)$ and $X_2(\Omega)$ of the microphone signals are given by

$$X_1(\Omega) = S(\Omega)\, H_1(\Omega) + N_1(\Omega) \tag{1}$$

$$X_2(\Omega) = S(\Omega)\, H_2(\Omega) + N_2(\Omega) \tag{2}$$

where $S(\Omega)$ denotes the short-time Fourier spectrum of the speech signal.

Furthermore, we suppose that there is almost the same attenuation of the speech signals received by the two microphones, which means that the magnitudes of the transfer functions are similar:

$$|H_1(\Omega)|^2 \approx |H_2(\Omega)|^2 \approx |H(\Omega)|^2 \tag{3}$$

An equivalent assumption holds for the noise power spectra because of the diffuse sound field:

$$\Phi_{N_1N_1}(\Omega) \approx \Phi_{N_2N_2}(\Omega) \approx \Phi_{NN}(\Omega) \tag{4}$$

Due to the described model of the acoustical environment the magnitude-squared cross power spectrum of the microphone signals is equal to

$$\begin{aligned} |\Phi_{X_1X_2}(\Omega)|^2 &= |E\{X_1^*(\Omega)\, X_2(\Omega)\}|^2 \\ &= E\left\{|S(\Omega)|^2\right\}^2 |H_1(\Omega)|^2\, |H_2(\Omega)|^2 \\ &= E\left\{|S(\Omega)|^2\right\}^2 |H(\Omega)|^4 \ . \end{aligned} \tag{5}$$

In the same way, we obtain for the product of the power spectra of the microphone signals (the parameter Ω has been left out for simplicity):

$$\begin{aligned} &\Phi_{X_1X_1}\, \Phi_{X_2X_2} = E\left\{|X_1|^2\right\}\, E\left\{|X_2|^2\right\} \\ &\quad = E\left\{|N_1|^2\right\}\, E\left\{|N_2|^2\right\} + E\left\{|S|^2\right\}^2 |H_1|^2\, |H_2|^2 \\ &\qquad + E\left\{|S|^2\right\} \left(|H_1|^2 E\left\{|N_2|^2\right\} + |H_2|^2 E\left\{|N_1|^2\right\}\right) \\ &\quad = \left(E\left\{|N|^2\right\} + |H|^2\, E\left\{|S|^2\right\}\right)^2 \end{aligned} \tag{6}$$

Taking the square roots and combining eqn. (5) and (6) results in:

$$\begin{aligned} E\left\{|N(\Omega)|^2\right\} &= \sqrt{\Phi_{X_1X_1}(\Omega)\, \Phi_{X_2X_2}(\Omega)} \\ &\quad - |\Phi_{X_1X_2}(\Omega)| \end{aligned} \tag{7}$$

The estimator for the discrete short-time power spectrum of the noise is obtained by replacing the cross- and auto-power spectral densities in eqn. (7) by their discrete short-time estimates

$$\begin{aligned} \hat{\Phi}_{NN}(\nu,\kappa) &= \sqrt{\tilde{\Phi}_{X_1X_1}(\nu,\kappa)\, \tilde{\Phi}_{X_2X_2}(\nu,\kappa)} \\ &\quad - |\tilde{\Phi}_{X_1X_2}(\nu,\kappa)| \end{aligned} \tag{8}$$

where ν and κ denote the discrete frequency and the decimated time index, respectively. The discrete short-time power spectra $\tilde{\Phi}_{X_1X_1}(.)$, $\tilde{\Phi}_{X_2X_2}(.)$, and $\tilde{\Phi}_{X_1X_2}(.)$ can be estimated using first-order IIR-filters, i.e.

$$\begin{aligned} \tilde{\Phi}_{X_1X_2}(\nu,\kappa) &= \beta\, \tilde{\Phi}_{X_1X_2}(\nu,\kappa-1) \\ &\quad + (1-\beta)\, X_1^*(\nu,\kappa)\, X_2(\nu,\kappa) \end{aligned} \tag{9}$$

where $X_1(\nu,\kappa)$ and $X_2(\nu,\kappa)$ denote the discrete short-time Fourier spectra and β is a constant close to one.

As shown in Fig. 1, the estimator of the noise short-time power spectrum is combined with the MMSE log-STSA estimator [4]. Spectral analysis and synthesis are performed by means of a weighted overlap-add FFT filterbank system. In a first step, the block "adaptive post filter" depicted in Fig. 1 should not be considered in this section. The real-valued spectral gain functions $G_1(\nu,\kappa)$ and $G_2(\nu,\kappa)$ obtained by the MMSE log-STSA estimator are used to weight the discrete short-time Fourier spectra $X_1(\nu,\kappa)$ and $X_2(\nu,\kappa)$ of the microphone signals, which leads to a restoration of the spectral amplitudes of the speech signal.

Note, that the structure in Fig. 1 provides a stereophonic output signal, as it is required e.g. in binaural hearing aids. The generation of a single output signal is beyond the scope of this paper but can easily be achieved by an adaptive time delay compensation and a successive addition of both channels.

Informal listening tests showed that the proposed structure provides a significant suppression of incoherent noise sound fields and room reverberation. However, both channels of the processed output signal still contain parts of the disturbing noise as well as low-level musical noise artefacts. Using a stereophonic presentation it can be observed that the listener is not able to designate the direction of incidence of the residual noise. This indicates that the remaining noise components appear mutually uncorrelated in the two output channels. This observation motivates the derivation of an adaptive post filter which benefits from the uncorrelated character of the remaining noise and leads to a further enhancement of the processed speech.

3 ADAPTIVE WIENER POST-FILTER

To derive the adaptation rule for the adaptive post filter, which is realized by means of a real-valued time-varying transfer function $W(\nu,\kappa)$ as depicted in Fig. 1, we take up an approach proposed by Zelinski [7] and Martin [8] for a noise suppression system implemented in the time-domain. In a first step, we regard the zero-phase Wiener filters $W_1(\nu,\kappa)$ and $W_2(\nu,\kappa)$ which aim at the minimization of the differences $\Delta_1(\nu,\kappa)$ and $\Delta_2(\nu,\kappa)$, respectively. The minimization of the difference Δ_1 implies that spectral components of the preprocessed spectrum $\tilde{X}_1$ which are uncorrelated to $\tilde{X}_2$ will be suppressed by the filter W_1. On the other hand, W_2 performs the same task for $\tilde{X}_2$. The short-time transfer functions of the zero-phase Wiener filters are given by

$$W_1(\nu,\kappa) = \frac{|\tilde{\Phi}_{\tilde{X}_1\tilde{X}_2}(\nu,\kappa)|}{\tilde{\Phi}_{\tilde{X}_1\tilde{X}_1}(\nu,\kappa)} \tag{10}$$

$$W_2(\nu,\kappa) = \frac{|\tilde{\Phi}_{\tilde{X}_1\tilde{X}_2}(\nu,\kappa)|}{\tilde{\Phi}_{\tilde{X}_2\tilde{X}_2}(\nu,\kappa)} \ , \tag{11}$$

which are to be considered for any fixed frequency index ν as functions of time κ. The short-time auto- and cross-power spectra $\tilde{\Phi}_{\tilde{X}_1\tilde{X}_1}(.)$, $\tilde{\Phi}_{\tilde{X}_2\tilde{X}_2}(.)$, and $\tilde{\Phi}_{\tilde{X}_1\tilde{X}_2}(.)$ of the preprocessed signals are estimated by first-order IIR filters similar to eqn. (9).

The transfer function W which is used as post filter in both channels is calculated by means of the zero-phase Wiener filters W_1 and W_2. Both authors [7, 8] propose to adapt the post filter according to the mean of both

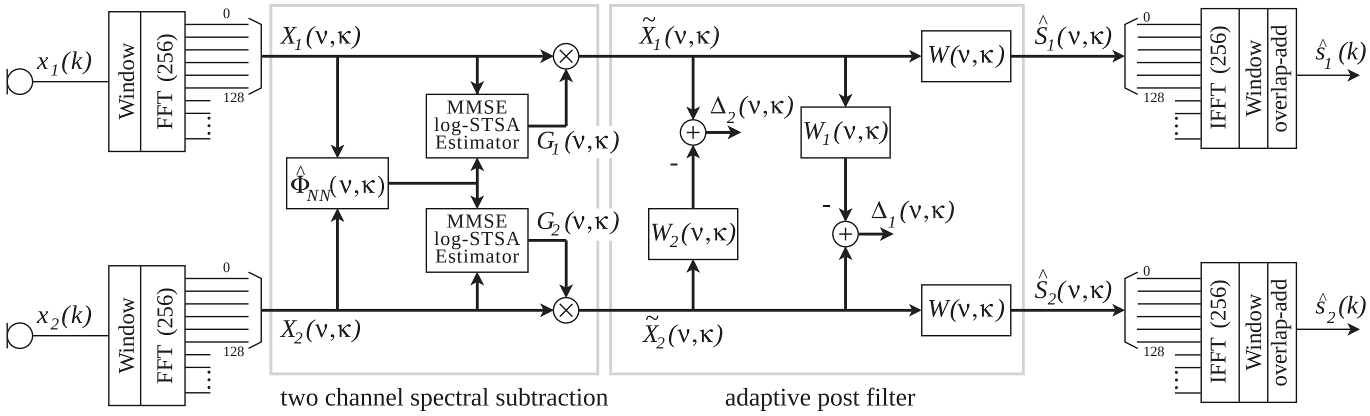

Figure 1: Proposed noise reduction system with two channel Spectral Subtraction and adaptive Wiener post-filtering ν: frequency index, κ: index of decimated time

filters, i.e.

$$W(\nu,\kappa) = \frac{W_1(\nu,\kappa) + W_2(\nu,\kappa)}{2} \ . \tag{12}$$

However, due to the adaptation according to the short-time power spectra determined by a frame-by-frame processing scheme, the adaptation rule (12) is not the optimal choice to attenuate musical noise components appearing mutually uncorrelated within the two channels of the preprocessed signal. Let us consider a time instant with a musical noise spectral component present in $\tilde{X}_1$, while $\tilde{X}_2$ is free of musical noise artefacts at the corresponding frequency. Therefore, only filter W_1 will lead to a suppression of the musical noise spectral component, while W_2 may even tend to an amplification at this frequency. Consequently, the adaptation of the post filter W according to eqn. (12) suffers from the behavior of the filter W_2 and does not lead to a satisfactory attenuation at the frequency disturbed by the musical noise artefact.

One possibility to exclude the negative influence of filter W_2 in this situation is to adapt the post filter according to the minimum of both filters

$$W(\nu,\kappa) = \min\left(W_1(\nu,\kappa), W_2(\nu,\kappa)\right) \ . \tag{13}$$

Alternatively, an even further suppression of mutually uncorrelated spectral components can be obtained using the relation $(\frac{1}{W_1}+\frac{1}{W_2})^{-1} \leq \min(W_1,W_2)$ and a squared adaptation rule:

$$W(\nu,\kappa) = \left(\frac{1}{W_1(\nu,\kappa)} + \frac{1}{W_2(\nu,\kappa)}\right)^{-2} \tag{14}$$

$$= \frac{|\Phi_{\tilde{X}_1\tilde{X}_2}(\nu,\kappa)|^2}{(\Phi_{\tilde{X}_1\tilde{X}_1}(\nu,\kappa) + \Phi_{\tilde{X}_2\tilde{X}_2}(\nu,\kappa))^2} \tag{15}$$

Due to the utilization of the squared magnitude of the cross power spectrum instead of the magnitude itself, this adaptation rule is much easier to realize on a DSP than the adaptation according to eqn. (10), (11), and (13).

Note, that eqn. (15) is exactly the square of the gain function which has been proposed by Allen et. al. [5] to reduce room reverberation.

Informal listening tests showed that the application of the post filter according to eqn. (15) results in a significantly improved performance of the two-channel Spectral Subtraction scheme concerning the residual noise as well as musical noise artefacts.

4 SYSTEM DESCRIPTION AND PERFORMANCE EVALUATION

The system depicted in Fig. 1 has been developed for the application in an electronic hearing aid with two microphones mounted nearby the wearer's ears. The system has been designed for a sampling rate of $f_S = 16\,\text{kHz}$. The filterbank for spectral analysis is realized as a weighted FFT, while the synthesis filterbank is based on inverse FFT and weighted overlap add techniques. The FFT length as well as the lengths of the windows for spectral analysis and synthesis are chosen to $N=256$. Due to the symmetry of the FFT in case of real-valued time domain signals only frequency bands with indices $0 \leq \nu \leq 128$ have to be processed as indicated in Fig. 1.

The overlap between two adjacent frames of the time domain signals is 192 samples, which is equal to a sampling rate reduction of $R = 64$. The windows for spectral analysis and synthesis $w_a(k)$ and $w_s(k)$, respectively, are given by $(0 \leq k < N)$

$$w_a(k) \ = \ w_k(k - \tfrac{N-1}{2}) \ \mathrm{sinc}(\tfrac{1}{N}(k - \tfrac{N-1}{2})) \tag{16}$$

$$w_s(k) \ = \ w_k(k - \tfrac{N-1}{2}) \ \mathrm{sinc}(\tfrac{1}{R}(k - \tfrac{N-1}{2})) \tag{17}$$

where $w_k(.)$ denotes the Kaiser window with its parameter $\alpha_k = 3.0$ and $\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$.

As mentioned previously, the discrete short-time power spectra of the microphone and the preprocessed signals are determinded by first-order IIR filters according to eqn. (9) with filter coefficients $\beta = 0.96$ in case of the noise power spectrum estimator and 0.9 for the adaptive post filter. The MMSE log-STSA estimator has been implemented corresponding to [3, 4] by means of a look-up table of size 31*31 elements. However, listening tests showed that in some situations the MMSE log-STSA estimator embedded in the proposed structure tends to a too strong attenuation of speech components. To avoid the resulting distortions of the speech signal we propose a smoothing of the gain values $G(\nu,\kappa)$ obtained by the MMSE log-STSA estimator towards 1.0, which is realized by the rule

$$G_{\text{smooth}}(\nu,\kappa) = G^{\gamma}(\nu,\kappa) \tag{18}$$

where $\gamma \in [0.7; 1.0]$ is a constant which depends on the listener's subjective impression.

The noise reduction system has been tested with audio signals recorded in a crowded cafeteria. The noise sound field in this situation is highly instationary due to its origin from interferent talkers and the use of dishes. The two microphones are mounted nearby the ears of a dummy head, while the speaker is placed at a distance of 1.8 m in front of the dummy head.

Listening tests proved that the application of the proposed structure leads to a significant reduction of the noise sound field. In the described situation, there are almost no audible distortions and no musical noise artefacts as long as the input-SNR is better than 6 dB. In comparison to conventional multi-microphone noise reduction systems [7, 8] the proposed structure shows a superior performance concerning residual noise, speech distortions, and musical noise artefacts.

Furthermore, the capability of the novel structure to reduce room reverberation has been compared with the system proposed by Allen et. al. [5]. Listening tests using speech signals recorded in a hall with a reverberation time of about 3 seconds confirm that the novel approach provides a significantly improved suppression of room reverberation, concerning the amount of the remaining reverberation, musical noise artefacts, and speech distortions.

To visualize the performance of the system, Fig. 2 shows the short-time power spectra of the undisturbed speech signal, the disturbed signal, and the processed signal in case of the cafeteria-situation. While the influence of the noise hides most of the structures of the speech within the disturbed signal, the application of the noise reduction system provides a restoration of the speech signal.

5 CONCLUSION

In this contribution a novel structure for the reduction of acoustic noise as well as room reverberation has been presented. The proposed concept takes advantage of the combination of a novel two-channel Spectral Subtraction technique and an adaptive Wiener post filter. Listening tests confirm the superiority of the novel noise reduction system in comparison to conventional techniques concerning the amount of noise suppression or the reduction of room reverberation, respectively. Compared to conventional techniques, the proposed structure provides a reduced level of musical noise artefacts and less speech distortions.

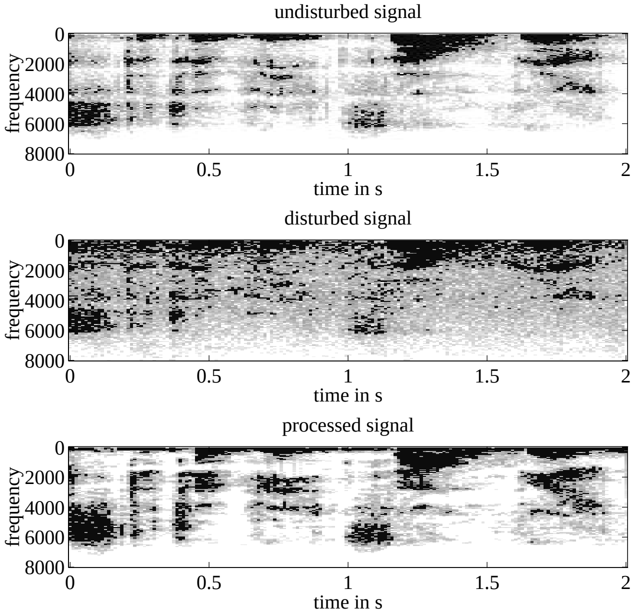

Figure 2: Short-time power spectra of the undisturbed speech signal, disturbed signal, and processed signal

References

[1] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise", in *Proc. IEEE Conf. ASSP*, Apr. 1979, pp. 208–211.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[5] J.B. Allen, D.A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals", *Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, Oct. 1977.

[6] R. Zelinski, "A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms", in *Proc. ICASSP*, 1988, pp. 2578–2581.

[7] R. Zelinski, "Ein Geräuschreduktionssystem mit Mikrofongruppe und LMS-gesteuerter adaptiver Nachfilterung", in *Fortschritte der Akustik – DAGA 1991*, Apr. 1991, pp. 893–896.

[8] R. Martin, *Freisprecheinrichtungen mit mehrkanaliger Echokompensation und Störgeräuschreduktion*, Dissertation, Aachener Beiträge zu Digitalen Nachrichtensystemen, Band 3, Verlag der Augustinus Buchhandlung, Aachen, 1995.

[9] G.C. Carter, "Coherence and Time Delay Estimation", *Proceedings of the IEEE*, vol. 75, no. 2, pp. 236–255, Feb. 1987.

[10] W. Armbrüster, R. Czarnach, and P. Vary, "Adaptive Noise Cancellation with Reference Input – Possible Applications and Theoretical Limits", in *Signal Processing III: Theories and Applications*, I.T. Tong et al., Eds. 1986, pp. 391–394, Elsevier Science Publishers B.V. (North Holland).

This work has been supported by the Deutsche Forschungsgemeinschaft.