MULTI-CHANNEL SPEECH ENHANCEMENT USING AN ADAPTIVE POST-FILTER WITH CHANNEL SELECTION AND AUDITORY CONSTRAINTS

*Martin Drews, Martin Streckfuß**

Institute for Telecommunications, Technical University of Berlin, Germany
drews@ft.ee.tu-berlin.de

*Siemens AG Semiconductors, Munich, Germany
martin.streckfuss@hl.siemens.de

ABSTRACT

A multi-channel speech enhancement system with 16 microphones is presented which consists of a conventional delay-and-sum beamformer and an adaptive post-filter. The post-filter adaptation is performed by Wiener filter analysis. The adaptation scheme is improved by a method for frequency-dependent channel selection, a modified method for speech spectrum estimation, and auditory constraints.

The selective processing yields small speech spectrum estimation errors, thus providing a high noise reduction. The application of the improved post-filter to the delay-and-sum beamformer results in a clear improvement of the speech signal quality even if only 4 microphones are used.

1 INTRODUCTION

When using hands-free speech communication systems, the speech signal acquisition is usually corrupted by reverberation and background noise which lead to a significant decrease in communication quality. For this reason, techniques for enhancing the desired speech signal are required which reduce the environmental noise. The objectives of speech enhancement are high quality and intelligibility of the output speech signal. Therefore, a noise reduction system is required which significantly attenuates the environmental noise without affecting the speech signal by additional distortions.

Many techniques which are efficient at enhancing noisy speech make use of more than one microphone for speech input [3, 6, 7]. The method presented here is based on a conventional delay-and-sum beamformer [4] combined with an adaptive post-filter. The structure of the speech enhancement system is shown in fig. 1.

The beamformer estimates the time differences of arrival (TDOAs) between the speech signals received by the microphones, compensates for these TDOAs, and sums the resulting signals. The summation of the delay-compensated input signals leads to an attenuation of the uncorrelated components by a factor of $10 \lg M$ dB (M denotes the number of microphones) while the correlated components are retained [2]. The summation introduces only small distortions into the output speech signal which are caused by TDOA estimation errors. In order to attain a relatively high noise power reduction, delay-and-sum beamformers use a large number of microphones. The microphone array used here consists of $M = 16$ microphones which are arranged as depicted in fig. 2. For the estimation of all M-1 TDOAs, the delay estimator presented in [1] is used which performs a speaker localization method.

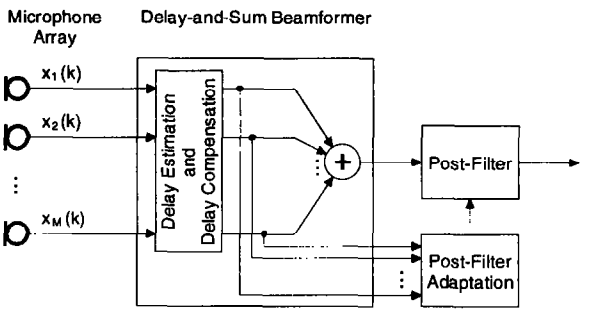

Fig. 1 Structure of the speech enhancement system

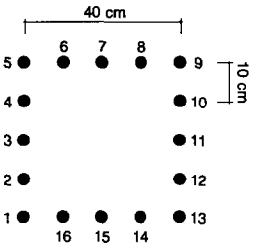

Fig. 2 Construction of the microphone array

Further noise power reduction is obtained by combining the delay-and-sum beamformer with an adaptive post-filter. The main problems in using adaptive post-filters for speech enhancement are associated with the "musical" distortions which usually remain in the speech signal after filtering. The present work focuses on the optimization of the post-filter adaptation scheme.

2 ADAPTATION OF THE POST-FILTER

The post-filter is adapted using Wiener filter analysis in the frequency domain. In this process, the power spectral density (PSD) of the speech signal is estimated via cross-spectral density (CSD) measurement. Due to estimation errors, the adaptive post-filtering introduces additional distortions into the speech signal which degrade its quality and intelligibility. In order to reduce this degradation, the post-filter adaptation scheme is improved by (i) frequency-dependent channel selection for the CSD analysis, (ii) a modified speech PSD estimation method, and (iii) auditory constraints to the post-filter transfer function. The structure is shown in fig. 3.

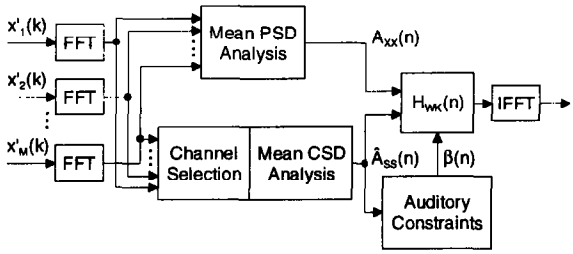

Fig. 3 Structure of the post-filter adaptation scheme

2.1 Estimation of the Speech PSD

The input signals received by the microphones are assumed to consist of highly correlated speech signals and mutually uncorrelated noise signals [2]. Thus, the speech PSD is estimated by averaging the CSDs

$$C_{X_i X_j}(n) = X'_i(n)\, X'^{*}_j(n) \qquad (1)$$

measured over several combinations of different DFT coefficients $X'_i(n)$ and $X'_j(n)$ of the delay-compensated input signals. Since these CSDs contain the PSDs of the mutually correlated speech signals and the uncorrelated CSDs of the noise signals, averaging attenuates the noise CSDs. The mean CSD involving M microphones is calculated from either

$$C_{XX}(n) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} C_{X_i X_j}(n) \qquad (2)$$

or

$$C_{XX}(n) = \frac{1}{M} \sum_{i=1}^{M-1} C_{X_i X_{i+1}}(n) \ . \qquad (3)$$

From the resulting mean CSD the speech PSD is estimated as proposed in [7] according to the following equation

$$\hat{A}_{SS}(n) = \begin{cases} \mathrm{Re}\left[C_{XX}(n)\right] & \text{if} \quad \mathrm{Re}\left[C_{XX}(n)\right] \geq 0 \\ 0 & \text{else} \end{cases} \qquad (4)$$

where $\mathrm{Re}\left[C_{XX}(n)\right]$ denotes the real part of the complex mean CSD.

The cross-correlation of the speech signals and the noise signals is frequency-dependent and further depends on the distance of the microphones: The cross-correlation increases with decreasing frequency and decreasing distance between the microphones. From this dependence a frequency range is defined by a lower frequency limit and an upper frequency limit which both depend on the distance between the microphones; the upper frequency limit further depends on the speaker localization error [2]. Above the lower frequency limit the cross-correlation of the noise signals is sufficiently low for high noise power reduction, whereas below the upper frequency limit the cross-correlation of the speech signals is high. The dependence of this frequency range on the microphone distance and the localization error is shown in fig. 4.

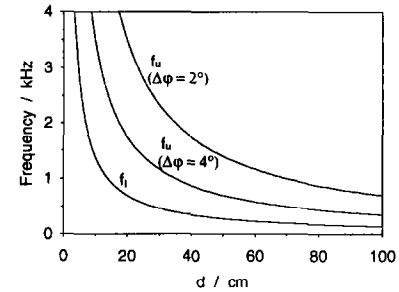

Fig. 4 Lower frequency limit f_l and upper frequency limit f_u as a function of the microphone distance d and the localization error $\Delta\varphi$

Since the array used here consists of 16 microphones with different distances (fig. 2), this dependence is exploited by a frequency-dependent channel selector for the CSD analysis which is designed for minimum speech PSD estimation errors. With increasing frequency, the channel selector selects input signals from microphone pairs with decreasing distances. The frequency ranges optimized for the distances of the microphone array used here are depicted in table 1.

Frequeny range in kHz	0 - 0.58	0.58 - 1.7	1.7 - 4.0
Number of microphones M	4	8	16
Microphone no.	1, 5, 4, 13	1..15 (odd)	1..16
Microphone distance in cm	≥40	20	10
Equation applied	(2)	(3)	(3)

Table 1 Assignment of the microphones to the frequency ranges

For comparative purposes, a speech PSD estimation method is used which calculates the mean CSD without selective processing by applying (2) to all of the 16 microphones. This method is clearly outperformed by

the estimation method which includes the frequency-dependent channel selector. The resulting improvement for the mean square estimation error of the speech PSD is shown in fig. 5. If only speaker localization errors impair the speech PSD estimation, the improvement increases by 2 to 4 dB at frequencies above 1 kHz. In the presence of additional background noise, the selective processing yields similar improvements for the greatest part of the speech spectrum.

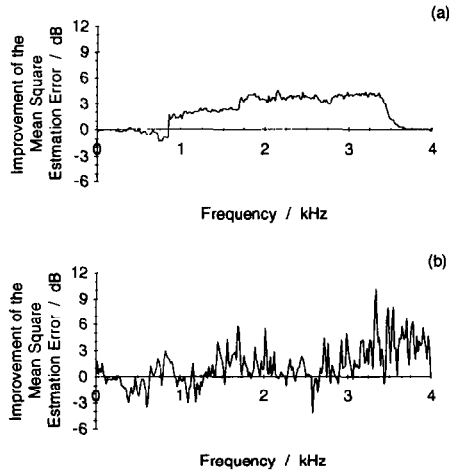

Fig. 5 Improvement for the mean square estimation error of the speech PSD by comparing the speech PSD estimation with frequency-dependent channel selector to the one which calculates the mean CSD according to (2) without selective processing: (a) 4° speaker localization error and no background noise, (b) no localization error and additional background noise at 0 dB input SNR

2.2 Estimation of the Post-Filter Transfer Function

The post-filter is adapted using Wiener filter analysis. In this process, the post-filter is calculated in the frequency domain from

$$H_{\mathrm{WK}}(n) = \frac{\hat{A}_{SS}(n)}{A_{XX}(n)} \tag{5}$$

where $\hat{A}_{SS}(n)$ denotes the speech PSD estimate and $A_{XX}(n)$ denotes the PSD of the output signal of the delay-and-sum beamformer which is estimated by

$$A_{XX}(n) = \left| \frac{1}{M} \sum_{i=1}^{M} X_i'(n) \right|^2 . \tag{6}$$

Another method for estimating $A_{XX}(n)$ is proposed in [7] where the mean PSD of the delay-compensated input signals is employed:

$$A_{XX}(n) = \frac{1}{M} \sum_{i=1}^{M} \left| X_i'(n) \right|^2 . \tag{7}$$

However, this method leads to an over-estimation of the mean noise PSD by a factor of M as shown in [6]. If (4) and (7) are applied to adapt the transfer function, the post-filtering yields a significant reduction of the noise power, but it also introduces clearly audible distortions into the speech signal. If (6) is used instead of (7), less distortions remain in the output speech signal, but on the other hand, the efficiency of the noise reduction decreases.

Various simulations showed that most of the distortions introduced by the post-filter arise from (4) as the speech PSD estimate is set to zero at frequencies where the real part of the complex mean CSD shows negative values. In order to overcome these estimation errors, several authors use smoothing techniques in which the speech PSD estimate is recursively smoothed over some periods. Another method is proposed in [3] where the speech PSD is estimated from the modulus of the mean CSD. By using this method and (6) for the post-filter adaptation, almost no noise power reduction at all is obtained by post-filtering and, as an advantage, no distortions are audible. To provide an improved method for speech PSD estimation, which exploits the advantage of the method proposed in [3] together with the more efficient noise power reduction achieved by (4), the estimation is done according to the following equation

$$\hat{A}_{SS}(n) = \begin{cases} \mathrm{Re}\left[C_{XX}(n)\right] & \text{if} \quad \mathrm{Re}\left[C_{XX}(n)\right] \geq 0 \\ \left|C_{XX}(n)\right| & \text{else} \end{cases} . \tag{8}$$

2.3 Post-Filter with Auditory Constraints

For further reduction of audible distortions in the post-filtered speech signal, the speech PSD estimate is weighted with spectral components of the noisy input speech signals. Accordingly, the transfer function of the post-filter is modified to

$$H_{\mathrm{WK}}(n) = \frac{\beta(n)\,\hat{A}_{SS}(n) + \left(1-\beta(n)\right) A_{XX}(n)}{A_{XX}(n)} \tag{9}$$

where $\beta(n)$ denotes the frequency-dependent weighting factor. Since the mean PSD $A_{XX}(n)$ contains both the PSD of the undistorted speech signal and the PSDs of the noise signals, this modification leads to lower distortions in the output speech signal, but also to a less efficient noise power reduction. In order to obtain the maximum perceptible noise power reduction together with a minimum speech signal distortion, auditory constraints are introduced into the post-filter adaptation scheme. In this process, the transfer function of the post-filter is set to one at frequencies where noise components are masked by speech.

The weighting factor in (9) is adapted by evaluating the noise masking threshold which is calculated from the

speech PSD estimate using critical band analysis [5]; its value spans the range of $0 \leq \beta(n) \leq 1$. In addition, the weighting factor is limited to a frequency-dependent maximum value considering that the estimation error of the speech PSD increases with decreasing frequency. During speech pauses, the weighting factor is set to its maximum value. The short-term speech PSD and its noise masking threshold are shown in fig. 6.

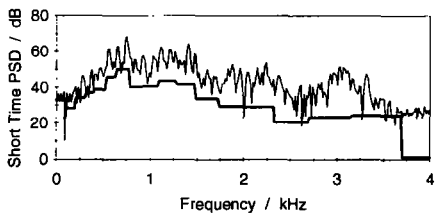

Fig. 6 Short time speech PSD (—) and noise masking threshold of noisy speech at 0 dB SNR (—)

With this adaptation scheme, the post-filter is constrained to reduce only those components of the noise signals which exceed the noise masking threshold. This capability is exploited for further noise reduction by using (7) for the estimation of the post-filter transfer function which provides an over-estimated mean noise PSD.

3 RESULTS AND CONCLUSION

The delay-and-sum beamformer was combined with post-filters which were adapted using different methods as mentioned above. These speech enhancement systems were compared to the pure delay-and-sum beamformer. A series of simulations were carried out using different noisy speech signals at various input SNRs.

As a result of the simulations, the improved post-filter was found in which (7) and (8) are performed to adapt the transfer function (9) *without* using smoothing techniques. Informal listening tests showed that this post-filter yields the most significant improvement with respect to perceptible noise attenuation while retaining a high speech quality. Only perceptible noise components are attenuated and the distortions of the output speech signal are reduced to a minimum. Moreover, the improved post-filter introduces lower distortions into the speech signal than the one whose transfer function results from (5). Especially onsets and offsets of speech signals are more clearly audible. Fig. 7 shows the spectrograms of clean, noisy, and enhanced speech.

It is worth mentioning that the improved post-filter shows comparable performance when added to delay-and-sum beamformers which use only 4 microphones. In this case, the speech PSD estimation is performed by calculating the mean CSD according to (2) without frequency-dependent channel selection.

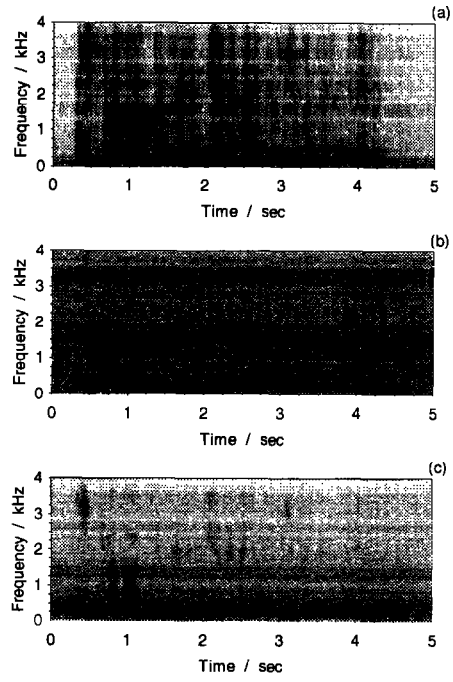

Fig. 7 Spectrograms: (a) clean input speech signal, (b) noisy input speech signals at -3 dB input SNR, (c) output speech signal of the improved post-filter

A extended post-filter adaptation scheme for speech enhancement in reverberant rooms has been presented. The post-filter clearly attenuates the environmental noise without affecting the speech signal when applied to delay-and-sum beamformers using 4 or more microphones.

REFERENCES

[1] Drews, M.: Speaker localization and its application to time delay estimators for multi-microphone speech enhancement systems. Proc. Eusipco 1996, pp. 483-486.

[2] Drews, M.: Construction of microphone arrays for the optimization of multi-channel speech enhancement systems. Frequenz 50 (1996), pp. 223-227 (in German).

[3] Fischer, S.; Simmer, K. U.: Beamforming microphone arrays for speech acquisition in noisy environments. Speech Communication 20 (1996), pp. 215-227.

[4] Haykin, S.: Array signal processing. Prentice Hall, 1985.

[5] Johnston, J. D.: Transform coding of audio signals using perceptual noise criteria. IEEE Journal on Selected Areas of Communications 81 (1988), pp. 314-323.

[6] Simmer, K. U.; Wasiljeff, A.: Adaptive microphone arrays for noise suppression in the frequency domain. Proc. 2nd Cost 229 Workshop on Adaptive Algorithms in Communications 1992, pp. 185-194.

[7] Zelinski, R.: A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. Proc. ICASSP 1988, pp. 2578-2581.