Multi-Channel Speech Enhancement using a Psychoacoustic Approach for a Post-Filter

Volker Mildner, Stefan Goetze, and Karl-Dirk Kammeyer
University of Bremen, Dept. of Communications Engineering, D-28334 Bremen, Germany
{mildner,goetze}@ant.uni-bremen.de

Abstract

Multi-channel systems containing a Delay&Sum-Beamformer have been previously combined with post-filters to achieve higher speech enhancement in noisy conditions. Problems occuring in the lower frequency regions were compensated by applying a Wiener weighting rule or spectral subtraction. We extend this by referring to the log-STSA rule of Ephraim-Malah.
Furthermore, we introduce a psychoacustically motivated post-filter based on the one presented by Gustafsson for single-channel systems [1]. Relying on Li's subband approach [2] we derive an algorithm which exploits the advantage of multi-channel systems in order to estimate those quantities necessary for computing the psychoacoustic weighting rule.

1 Introduction

We consider a multi-channel system depicted in Figure 1 with M microphones.

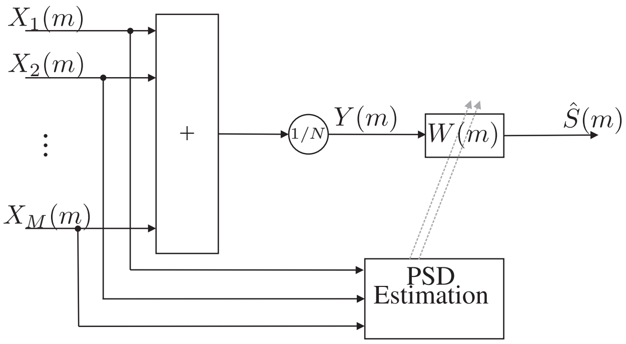

Figure 1: Beamformer with post-filter

In each microphone path a discrete-time signal $x(k)$ is present consisting of the speech signal $s(k)$ and additive noise $n(k)$ such that

$$x(k) = s(k) + n(k) \tag{1}$$

or in the Fourier domain

$$X(m) = S(m) + N(m) \tag{2}$$

with the discrete-time index k and discrete-frequency index m.

The Delay&Sum-Beamformer in Figure 1 is assumed to be pre-steered towards the desired signal $s(k)$ which implies that the microphone signals $x(k)$ are delay-compensated. The output signal is filtered by a postfilter such that $\hat{S}(m) = Y(m) \cdot W(m)$.
In order to realize a Wiener-Filter for spectral weighting

$$W(m) = \frac{\Phi_{SS}(m)}{\Phi_{SS}(m) + \Phi_{NN}(m)} \tag{3}$$

- where $\Phi_{SS}(m)$ and $\Phi_{NN}(m)$ denote the power spectral densities (PSD) of the signal $s(k)$ and $n(k)$ respectively - Zelinski [3] derived a solution under the assumptions that

- the speech signal and the noise signal are uncorrelated $\mathrm{E}\{S^*(m)N_i(m)\} = 0,\ 1 \leq i \leq M$ and that
- noise signals of different microphones are uncorrelated $\mathrm{E}\{N_i^*(m)N_j(m)\} = 0, \forall i \neq j$.

By this one may rewrite the PSDs (while $i \neq j$)

$$\Phi_{X_iX_i}(m) = \Phi_{SS}(m) + \Phi_{NN}(m) \tag{4}$$
$$\Phi_{X_iX_j}(m) = \Phi_{SS}(m) \tag{5}$$

To exploit information from different channels, PSDs are estimated via periodogram averaging

(with $\Re$ as the real part)

$$\hat{\Phi}_{X_iX_i}(m) = \frac{1}{M}\sum_{i=1}^{M} X_i^*(m)X_i(m) \qquad (6)$$

$$\hat{\Phi}_{X_iX_j}(m) = \frac{2}{M(M-1)} \cdot \Re\left\{\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\{X_i^*(m)X_j(m)\}\right\} \qquad (7)$$

For a block oriented implementation a first order recursive smoothing with the discrete block index l is written as

$$\hat{\Phi}_{SS}(m,l) = \alpha\hat{\Phi}_{X_iX_j}(m,l-1) + (1-\alpha)\hat{\Phi}_{X_iX_j}(m,l) \qquad (8)$$

$$\hat{\Phi}_{SS}(m,l) + \hat{\Phi}_{NN}(m,l) = \alpha\hat{\Phi}_{X_iX_i}(m,l-1) + (1-\alpha)\hat{\Phi}_{X_iX_i}(m,l) \qquad (9)$$

This leads to Zelinski's post-filter [3]:

$$W_{Zel}(m,l) = \frac{\hat{\Phi}_{SS}(m,l)}{\hat{\Phi}_{SS}(m,l) + \hat{\Phi}_{NN}(m,l)} \qquad (10)$$

2 Subband Approach

For the derivation of the Zelinski post-filter transfer function the assumption of mutually uncorrelated noise signals in the different microphone paths was made. Unfortunately, in most practical noise environments (diffuse noise fields, e.g.) the noise is highly correlated in the low frequency regions [4]. Thus, a subband approach as previously introduced by Li is considered [2].

For a diffuse noise field the magnitude squared coherence (MSC), which is a measure for the correlation of the noise signals of different microphones, depends only on the intermicrophone-spacing for a linearly spaced array.

$$\Gamma_{X_iX_j}(m) = \text{si}^2(2\pi \cdot m \cdot d_{ij}/c) \qquad (11)$$

Here, d_{ij} is the distance between two microphones and c is the speed of sound. For M linearly spaced microphones those two microphones forming the pair $\{1, M\}$ shall have the greatest possible distance d_{max} of all microphones. This defines the *lowest subband*, for which the noise is correlated in all microphone pairs. The upper bandlimit f_1 of this subband B_1 is defined by examining the first zero of the si-function which results in

$$f_{1,\{1,M\}} = \frac{c}{2d_{max}} \qquad (12)$$

Considering the pairs of microphones $\{i, j\}$ next furthest apart by d_{ij} from each other, it is possible to generally define the upper bandlimit frequency f_t of the subband B_t, $1 \leq t \leq M$, by

$$f_{t,\{i,j\}} = \frac{c}{2d_{ij}} \, . \qquad (13)$$

The microphone array examined exemplarily in this contribution consists of $M = 4$ microphones with an equal spacing of 8 cm, resulting in d_{max} =24 cm and an upper bandlimit for the subband B_1 of $\approx$ 700 Hz The corresponding MSC with the bandlimits for all subbands $B_1 ... B_4$ is depicted in Figure 2.

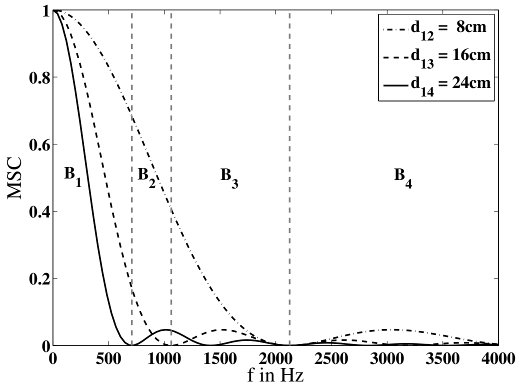

Figure 2: MSC between microphone channels

The consequence of the correlation of the noise signals of different microphones (approximately described by the MSC) is, that the lower the considered subband is the smaller is the number of microphones which may be relied on for the estimate of equation (7). For an arbitrary subband B_t with $2 \leq t \leq M$ the estimate of (7) may be computed via

$$\hat{\Phi}_{X_iX_j,t}(m,l) = \frac{2}{t(t-1)} \cdot \Re\left\{\sum_{i=1}^{t-1}\sum_{j=i+M-t+1}^{M} X_i^*(m,l)X_j(m,l)\right\} \qquad (14)$$

Based on the above equation a PSD estimate of the speech signal for each separate subband $\hat{\Phi}_{SS,t}(m)$ via (8) can be obtained. In the lowest subband B_1 - where there is no pair of microphones at hand to yield an estimate via (14) - Bitzer [4] applies spectral subtraction. Li [2] proposes to use a Wiener-Filter with a decision directed approach [5], which also provides a speech absence probability in order to estimate the noise signal.

As an extension to Li's subband approach, we use Martin's method of minimum statistics [6] to estimate the noise PSD in subband B_1 and apply the log-STSA weigthing rule by Ephraim-Malah [7] to further reduce the problem of musical noise.

3 A Psychoacoustic Approach

Gustafsson introduced a psychoacoustic filter for single channel systems which has the advantage of not suffering from musical noise during speech pauses [1]. The weighting function is defined as

$$W_{IND}(m,l) = \sqrt{\frac{\Phi_{TT}(m,l)}{\Phi_{NN}(m,l)}} + \zeta_n \qquad (15)$$

where ζ_n is a constant value of attenuation and $\Phi_{TT}(m,l)$ the masking threshold caused by the clean speech signal $S(m,l)$.

To obtain a noise estimate $\hat{\Phi}_{NN}(m,l)$ under the condition of a single channel system, Gustaffson used the method of minimum statistics [6]. To have an estimate $\hat{S}(m,l)$ at hand he performed a prefiltering of the noisy signal $X(m,l)$ via the Ephraim-Malah rule [7] with optimized parameters for that purpose. Here, we exploit the knowledge based on the multichannel system.

We compute Zelinski's Filter $W_{Zel}(m,l)$ based on the subband approach of equation (14) for the subbands $B_2 ... B_4$. For subband B_1 we use a Wiener Filter as Li [2] did while yielding our noise estimate in this subband via Martin's method. The noisy signal behind the beamformer $Y(m,l)$ is then prefiltered by this combined filter $W_{Zel,Wien}(m,l)$ to obtain an estimate of the clean speech

$$\hat{S}(m,l) = Y(m,l) \cdot W_{Zel,Wien}(m,l) \qquad (16)$$

The masking threshold shall then be a function of the speech estimate after [1]

$$\hat{\Phi}_{TT}(m,l) = f\left(\hat{S}(m,l)\right) \qquad (17)$$

As previously mentioned the noise PSD in the lowest subband $\hat{\Phi}_{NN,1}(m,l)$ is estimated by minimum statistics. For higher subbands $t \geq 2$ we subtract the estimated PSD of the speech signal (relying on (14)) from the PSD estimate of the noisy signal (9)

$$\hat{\Phi}_{NN,t}(m,l) = \qquad (18)$$
$$\left[\hat{\Phi}_{SS,t}(m,l) + \hat{\Phi}_{NN,t}(m,l)\right] - \hat{\Phi}_{SS,t}(m,l)$$

The obtained estimates of all subbands are then concatenated to form the estimate $\hat{\Phi}_{NN}(m,l)$ of all frequencies.

Finally, Gustafsson's rule is computed as

$$\hat{W}_{IND}(m,l) = \sqrt{\frac{\hat{\Phi}_{TT}(m,l)}{\hat{\Phi}_{NN}(m,l)}} + \zeta_n \qquad (19)$$

and the output signal $\tilde{S}(m,l)$ found by spectral weighting of the beamformer's signal $Y(m,l)$

$$\tilde{S}(m,l) = \hat{W}_{IND}(m,l) \cdot Y(m,l) \qquad (20)$$

4 Simulation Results

The acoustic environment was an office room with a reverberation time of $\tau_{60} = 250$ ms. The microphone signals were obtained by convolution of speech and noise signals with simulated room impulse responses. The noise was shaped like pink noise with additional slight spectral pulses. The time signals are sampled at $f_s = 8$ kHz.

The different methods for noise suppression mentioned in this contribution are compared by instrumental measures: SNR-enhancement (SNRE) (during speech activity only), noise reduction (NR) (during speech pauses only)

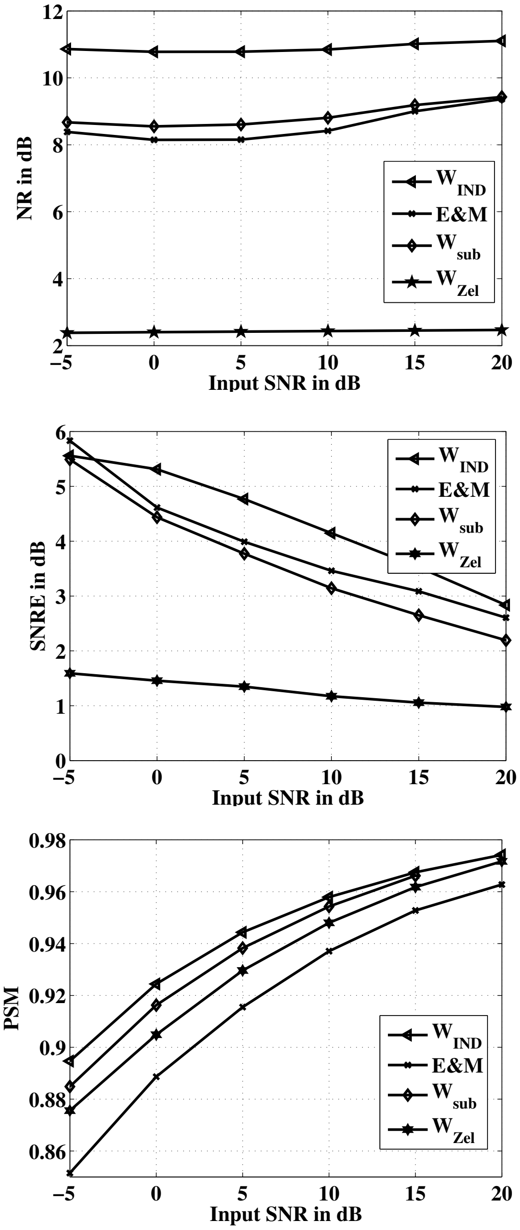

Figure 3: Comparison of the different noise reduction schemes

and the perceptual similarity measure PSM [8] (throughout the whole sample). While the first measure simply compares the increase in speech-to-noise SNR and the second measure only the reduction of the noise power, the third measure is able to tell how similar the processed signal is to the desired (clean) signal [9].

The amount of noise reduction (NR) is highest for the proposed rule W_{IND}, while the rules of Ephraim&Malah (*E&M*) as well as the subband approach W_{sub} also achieve considerable values. The early approach of Zelinski W_{Zel} fails due to the mentioned problems in lower frequency regions. A similar ranking can be seen in terms of speech enhancement SNRE. Attention should be paid to the PSM measure, which unveals that the proposed rule has the best audio quality while at the same time performing high noise reduction.

5 Conclusion

The proposed weighting rule profits from spatial information of a microphone array to enable an algorithm to perform sufficient noise reduction along with providing acceptable audio quality.

References

[1] S. Gustafsson, *Enhancement of Audio Signals by Combined Acoustic Echo Cancellation and Noise Reduction*, Insitut für Nachrichtengeräte und Datenverarbeitung, RWTH Aachen, 1999.

[2] J. Li and M. Akagi, "A Hybrid Microphone Array Post-Filter in a Diffuse Noise Field," in *Proc. Eurospeech 2005*, Lisbon, Portugal, September 2005.

[3] R. Zelinski, *A microphone array with adaptive post-filtering for noise reduction in reverbant rooms*, IEEE ICASSP, New York, 1988.

[4] J. Meyer (Bitzer) and K.U. Simmer, "Multi–Channel Speech Enhancement in a Car Environment Using Wiener Filtering and Spectral Subtraction," in *Proc. IEEE Int. Conference Acoustic, Speech and Signal Processing, ICASSP–97, Munich, Germany*, April 1997.

[5] Y. Ephraim and D. Malah, *Speech Enhancement using a minimum mean-square error short-time spectral amplitude estimator*, IEEE Trans. on Acoustics, Speech and Signal Processing, 1984.

[6] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[7] Y. Ephraim and D. Malah, *Speech Enhancement using a minimum mean-square error log short-time spectral amplitude estimator*, IEEE Trans. on Acoustics, Speech and Signal Processing, 1985.

[8] R. Huber, *Objective Assessment of Audio Quality Using an Auditory Processing Model*, Ph.D. thesis, University of Oldenburg, 2003.

[9] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective Measures for the Evaluation of Noise Reduction Schemes," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2005.