MAXIMUM LIKELIHOOD BASED NOISE COVARIANCE MATRIX ESTIMATION FOR MULTI-MICROPHONE SPEECH ENHANCEMENT

Ulrik Kjems and Jesper Jensen

Oticon A/S, Smørum, Denmark

ABSTRACT

Multi-microphone speech enhancement systems can often be decomposed into a concatenation of a beamformer, which provides spatial filtering of the noisy signal, and a single-channel (SC) noise reduction filter, which reduces the noise remaining in the beamformer output. Here, we propose a maximum likelihood based method for estimating the inter-microphone covariance matrix of the noise impinging on the microphone array. The method allows prediction of this covariance matrix for non-stationary noise sources even in signal regions where the target speech signal is present. Although the noise covariance matrix may have several purposes, we use it in this paper for estimating the power spectral density (psd) of the noise entering the SC filter, as this is important for optimal operation of the SC filter. In simulation experiments with a binaural hearing aid setup in a realistic acoustical scenario, the proposed method performs better than existing methods for estimating this noise psd.

Index Terms— Multi-Microphone Speech Enhancement, Noise Covariance Estimation, Noise Power Spectral Density Estimation, Single-Channel Post Filter.

1. INTRODUCTION

Digital speech applications such as mobile phones, voice-controlled devices, hearing aids, etc., must be robust to acoustical background noise and reverberation. For this reason, such devices are often equipped with noise reduction / speech enhancement algorithms. These algorithms can be divided into single- and multi-microphone methods. Although multi-microphone methods generally require more space and are often more demanding in terms of computational and hardware complexity, they are often employed as they can deliver better performance than single-microphone methods. More specifically, multi-microphone methods can be seen as a concatenation of a beamformer algorithm and a single-channel noise reduction algorithm; therefore multi-microphone methods can perform spatial filtering in addition to the spectro-temporal filtering offered by stand-alone single-channel systems.

Recently, the Multi Channel Wiener filter (MWF) for speech enhancement [1] and derivatives thereof, e.g. [2], have received a significant amount of attention. The MWF is the optimal linear estimator in mean-squared error sense of a target signal, given that the microphone signal is the target signal perturbed by uncorrelated, additive noise. It can be shown, e.g. [3], that the MWF can be decomposed into a concatenation of a Minimum Variance Distortionless Response (MVDR) beam former and a single-channel (SC) Wiener filter. While these two systems are theoretically identical, the decomposed system is advantageous in practice over a brute-force implementation of the MWF filter. Specifically, one can exploit that the spatial signal statistics, which need to be estimated to implement the MVDR beamformer, change across time at a different (often slower) rate than the signal statistics that need to be estimated to implement the SC filter.

Most, if not all, SC filters rely on an estimate of the power spectral density (psd) of the noise entering the SC filter. Considering a multi-microphone noise reduction system as a concatenation of a beamforming algorithm and an SC filter, it is obviously possible to estimate the noise psd directly from the output signal of the beamformer, using well-known single-channel noise tracking algorithms, e.g. [4, 5]. However, generally speaking, better performance can be obtained by taking advantage of having multiple microphone signals available when estimating the psd of the noise entering the SC filter.

The idea of using multiple microphone signals for estimating the psd of the noise that enters the SC post filter is not new. In [6], Zelinski used multiple microphone signals to estimate the noise psd observed at the microphones under the assumption that the noise sequences were uncorrelated between microphones, i.e., the inter-microphone noise covariance matrix was diagonal. McCowan [7] and Lefkimmiatis [8] replaced this often unrealistic model with a diffuse (homogenous, isotropic) model of the noise field. More recently, Wolff [9] considered the beamformer in a generalized sidelobe canceller (GSC) structure, and used the output of an (adaptive) blocking matrix, combined with a voice activity detection (VAD) algorithm, to compute an estimate of the psd of the noise entering the SC filter.

In this paper we propose an algorithm which, for each frequency, estimates the time-varying inter-microphone noise covariance matrix. Although this noise covariance matrix es-

The authors wish to thank Oticon A/S for fascilitating this work and Jørn Skovgaard, Oticon A/S, for providing the room impulse responses.

timate may have several other purposes, we use it here for estimating the psd of the noise entering the SC filter. The proposed algorithm shows similarities to the method in [9], but unlike the slightly *ad hoc* scheme presented there, we propose a scheme which is optimal in a maximum likelihood sense.

2. SIGNAL MODEL AND ASSUMPTIONS

Let the noisy signal impinging on the m'th microphone be given by

$$y_m(n) = x_m(n) + v_m(n), \qquad m = 1, \ldots, M,$$

where $y_m(n)$, $x_m(n)$, and $v_m(n)$ denote signal samples of the noisy, clean target, and noise signal, respectively, $M > 1$ is the number of available microphone signals, and where we have ignored analog-to-digital conversion and simply used the discrete-time index n for convenience. We assume, for mathematical convenience, that the observations are realizations of zero-mean Gaussian random processes, and that the noise process is statistical independent of the target process.

Each microphone signal is passed through a discrete Fourier Transform (DFT) filterbank, leading to complex DFT coefficients

$$Y_m(l,k) = \sum_{n=0}^{N-1} y_m(lD - n) w_A(n) e^{-\frac{2\pi jkn}{N}},$$

where l and k denote frame and frequency bin indices, respectively, N is the frame length, D is the filterbank decimation factor, $w_A(n)$ is the analysis window function, and $j = \sqrt{-1}$ is the imaginary unit.

We employ the standard assumption that DFT coefficients are independent across frame and frequency index, which allows us to process each DFT coefficient independently. Thus, without loss of generality, for a given frequency index, we can collect the DFT coefficients of frame l for each microphone in a vector $Y(l) \in \mathbb{C}^M$ as

$$Y(l) \triangleq [Y_1(l) \ldots Y_M(l)]^T.$$

Similar equations describe the target vector $X(l) \in \mathbb{C}^M$ and the noise vector $V(l) \in \mathbb{C}^M$.

We model the target signal as a point source impinging on the array. Let $d(l) = [d_1(l,k) \cdots d_M(l,k)]^T$ denote the (complex-valued) propagation vector whose elements d_m represent the acoustic transfer function from the source to the m'th microphone, evaluated at frequency index k. Then, $X(l)$ may be written as, $X(l) = x(l)d(l)$, where $x(l)$ is the target DFT coefficient with frame index l at the frequency index in question.

Now the correlation matrix $\Phi_{YY}(l) = E\left[Y(l)Y^H(l)\right]$ can be written as

$$\Phi_{YY}(l) = \underbrace{\phi_{xx}(l)d(l)d^H(l)}_{\Phi_{XX}(l)} + \underbrace{E\left[V(l)V^H(l)\right]}_{\Phi_{VV}(l)},$$

where the superscript H denotes Hermitian transposition, and $\phi_{xx}(l) = E[|x(l)|^2]$ is the psd of the target signal.

Finally, let us assume the following model for the development of the noise covariance matrix across time,

$$\Phi_{VV}(l) = c^2(l)\Phi_{VV}(l_0), \qquad l > l_0, \tag{1}$$

where $c(l) \in \mathbb{R}$ is a time-varying scaling factor, and $\Phi_{VV}(l_0)$ is the noise covariance matrix at the most recent frame index l_0 where the target was absent. Thus, Eq. (1) represents the evolution of $\Phi_{VV}(l)$ when speech is present; the noise process does not need to be stationary, but the covariance structure must remain fixed up to a scalar multiplication. Thus, this model can be seen as a relaxation of the methodology known from early single-channel noise reduction systems, where the noise psd estimated in the most recent noise-only region is assumed to remain constant across time when speech is present.

3. MAXIMUM LIKELIHOOD ESTIMATION OF THE NOISE COVARIANCE MATRIX

The goal in this section is to derive an estimate of the noise covariance matrix $\Phi_{VV}(l), l > l_0$, that is, when speech is present. The general idea is to do this based on the output of a set of linearly independent target cancelling beamformers, sometimes refered to as a blocking matrix in GSC terminology [10], see also [11, Chap.5] and the references therein.

Consider any full-rank matrix $B(l) \in \mathbb{C}^{M \times M-1}$ which satisfies

$$B^H(l)d(l) = 0.$$

Obviously, many such matrices exist. Assume that $d(l)$ is known and normalized to unit length, and let $H(l) = I_M - d(l)d^H(l)$, where I_M is the M-dimensional identity matrix. Then, it can be verified that one such matrix $B(l)$ is given by the first $M-1$ columns of matrix, $H(l)$, that is

$$[B(l)\; h(l)] = H(l), \tag{2}$$

where $h(l)$ is simply the M'th column in $H(l)$.

Each column of matrix $B(l)$ can be considered a target-cancelling beamformer, because when applied to the noisy input vector $Y(l)$, the output $Z(l) \in \mathbb{C}^{M-1}$ is only noise related

$$Z(l) = B^H(l)Y(l) = B^H(l)V(l). \tag{3}$$

From Eq. (3) the covariance matrix of $Z(l)$ is given by

$$\Phi_{ZZ}(l) \triangleq E\left[Z(l)Z^H(l)\right] = B^H(l)\Phi_{VV}(l)B(l). \tag{4}$$

Inserting Eq. (1) in Eq. (4) we find

$$\Phi_{ZZ}(l) = c^2(l)B^H(l)\Phi_{VV}(l_0)B(l), \qquad l > l_0. \tag{5}$$

From the Gaussian assumption, it follows that vector $Z(l)$ obeys a zero-mean (complex, circular symmetric) Gaussian distribution, that is,

$$
\begin{aligned}
f_{Z(l)}(Z(l)) =& \frac{1}{\pi^{M-1}\left|\Phi_{ZZ}(l)\right|} \times \\
& \exp\left(-Z^H(l)\Phi_{ZZ}^{-1}(l)Z(l)\right),
\end{aligned}
\qquad (6)
$$

where $|\cdot|$ denotes the matrix determinant. The matrix $\Phi_{ZZ}(l)$ is invertible when $\Phi_{VV}(l_0)$ is invertible (see Eq. (5)), which is usually the case.

Inserting Eq. (5) in Eq. (6), the log-likelihood function $\mathcal{L}$ can be written as

$$
\begin{aligned}
\mathcal{L} =& \log f_{Z(l)}(Z(l)) \\
=& -(M-1)\log\pi - (M-1)\log c^2(l) \\
& - \log\left|B^H(l)\Phi_{VV}(l_0)B(l)\right| \\
& - c^{-2}(l)Z^H(l)\left[B^H(l)\Phi_{VV}(l_0)B(l)\right]^{-1}Z(l).
\end{aligned}
$$

Maximizing $\mathcal{L}$ with respect to the unknown scaling factor $c^2(l)$ leads to the maximum likelihood estimate

$$
c^2_{ML}(l) = \frac{1}{M-1}Z^H(l)\left[B^H(l)\Phi_{VV}(l_0)B(l)\right]^{-1}Z(l). \qquad (7)
$$

Note that $c^2_{ML}(l) \geq 0$ such that the noise covariance estimate

$$
\hat{\Phi}_{VV}(l) = c^2_{ML}(l)\hat{\Phi}_{VV}(l_0), \qquad l > l_0, \qquad (8)
$$

remains positive definite as long as the noise covariance estimate $\hat{\Phi}_{VV}(l_0)$ obtained in the most recent noise-only region is positive definite.

Finally, let $w(l) \in \mathbb{C}^M$ denote the linear beamformer filter such that the beamformer output is given by $\tilde{x}(l) = w^H(l)Y(l)$. Then an estimate of the psd of the noise in the beamformer output is given by

$$
\hat{\phi}_{VV}(l) = w^H(l)\hat{\Phi}_{VV}(l)w(l). \qquad (9)
$$

Fig. 1 shows a block diagram of a beamformer-SC filter noise reduction system including the proposed algorithm for estimating the inter-microphone noise covariance matrix $\Phi_{VV}(l)$ and the psd $\phi_{VV}(l)$ of the noise component in $\tilde{x}(l)$.

4. SIMULATION EXPERIMENTS

We now present results of simulation experiments with synthetic and real audio signals. More specifically, we study the performance of the proposed algorithm in estimating the noise psd $\phi_{VV}(l)$. The proposed scheme (PROP) is compared to existing algorithms, namely the algorithms by Lefkimmiatis *et. al.* (LEF) [8] and Wolff *et. al.* (WOL) [9], which both make use of multiple microphone signals for estimating $\phi_{VV}(l)$. We also compare to the state-of-the-art *single-channel* noise psd tracking method by Hendriks *et. al.* (HEN) [5], applied to the MVDR output, $\tilde{x}(l)$, to demonstrate the advantage of using multiple microphone signals.

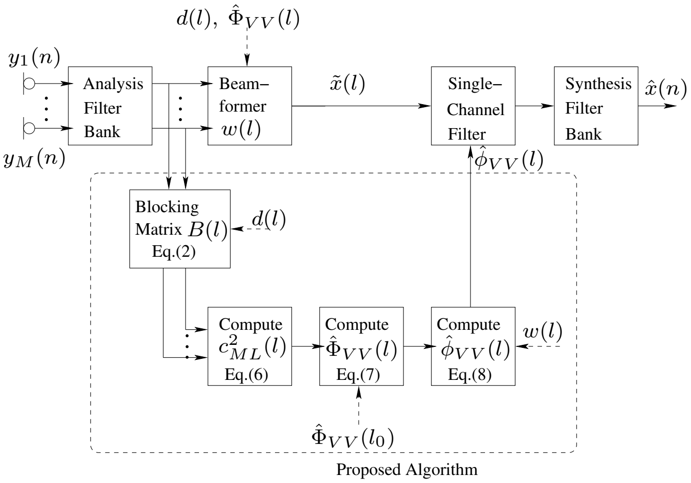

Fig. 1. Block diagram of multi-microphone noise reduction system with proposed algorithm for estimating the psd of the noise signal entering the SC filter.

Speech and noise signals are sampled at a rate of 16 kHz. The analysis filter banks use frames of $N = 512$ samples, and a decimation factor of $D = 256$. The analysis window function is a square-root Hann window. We use an ideal acoustical propagation vector $d(l)$ estimated in an offline calibration procedure, where the target sound source is played in isolation from a location directly in front of the microphone array (in an endfire configuration). We use a fixed MVDR beamformer,

$$
w(l) = \frac{\hat{\Phi}_{VV}^{-1}(l_0)d(l)d^*_{\triangle}(l)}{d^H(l)\hat{\Phi}_{VV}^{-1}(l_0)d(l)},
$$

in all simulations, where $d^*_{\triangle}(l)$ is the complex conjugate of the element in vector $d(l)$ corresponding to the reference microphone (we use $d_{\triangle}(l) = d_1(l)$). The matrix $\hat{\Phi}_{VV}(l_0)$ is estimated from a known noise-only signal region prior to speech activity; in practice, this requires use of a voice activity detection (VAD) algorithm. For the proposed algorithm, matrix $B(l)$ is found from Eq. (2). Our implementation of WOL uses the blocking matrix $B(l) = H(l) \in \mathbb{C}^{M\times M}$.

The noise psd tracking performance is evaluated using the symmetric log-error distortion measure

$$
\text{LogErr} = \frac{1}{KL}\sum_k\sum_l\left|10\log_{10}\left(\frac{\bar{\hat{\phi}}_{VV}(k,l)}{\phi_{VV}(k,l)}\right)\right| \quad \text{[dB]},
$$

where the frequency index k covers frequencies 300-4500 Hz, $K = 134$ is the number of DFT coefficients in this frequency range, the frame index l covers signal frames with speech activity, L is the number of such frames, $\bar{\hat{\phi}}_{VV}(k,l)$ is the noise psd estimates $\hat{\phi}_{VV}(k,l)$ smoothed in a first-order lowpass filter with a time constant of 50 ms, and $\phi_{VV}(k,l)$ denotes the true psd of the noise signal entering the SC filter.

4.1. Anechoic Scenario with Modulated Noise

In our first experiment $M = 4$ microphones are arranged in a uniform linear array with a constant microphone distance of 1 cm. The target speaker is located directly in front of the array at a distance of 3 m, and a diffuse noise field is modeled using 72 equidistant point noise sources arranged in a 3 m diameter circle in the horizontal plane intersecting the array. The noise sources consist of independent speech-shaped Gaussian noise sequences, amplitude modulated by a square wave function with an amplitude of 10 dB and with a time-varying modulation frequency which increases from 0.1 to 2 Hz, see the thick black line in Fig. 2A. The microphone signals are generated under an anechoic (free-field) assumption, in this initial experiment. In this way we try to meet the underlying assumptions of the proposed method and the methods that we compare to. Specifically, the noise is additive (assumed by all the methods), the noise is Gaussian (assumed in HEN [5]), diffuse (assumed in LEF [8]), and the noise covariance matrix is constant across time up to a scalar multiplier (assumed in LEF [8], WOL [9] and the proposed method). For each noisy signal, the initial 2.5 s consists of noise-only, which is used for initialization of the algorithms. Specifically, for the proposed algorithm, this signal region is used to estimate $\Phi_{VV}(l_0)$.

Fig. 2 shows examples of the noise psd tracking performance for each of the algorithms for a subband centered at $f = 1250$ Hz. The SNR of a given microphone signal (measured across the full noisy signal) is approximately 0 dB. For this example, the proposed method (Fig. 2A) is able to track the underlying noise PSD accurately and with rather small fluctuations. The other multi-microphone based methods, LEF, and WOL (Figs. 2B and 2C), perform quite good, while HEN (Fig. 2D), relying only on the single-channel output signal of the MVDR filter, has difficulties in tracking these very abrupt changes in noise level.

We repeated this experiment for 20 different target speakers (15 female and 5 male). Fig. 3 plots the LogErr scores, averaged across target speakers, for input SNRs in the range -5 – 10 dB (to change the SNR, the noise variance was adjusted, while the target signal variance was kept constant). Clearly, the methods using multiple microphones (PROP, WOL, and LEF) are better than the single-channel method HEN. The proposed method is slightly better than LEF and WOL, especially at lower SNRs.

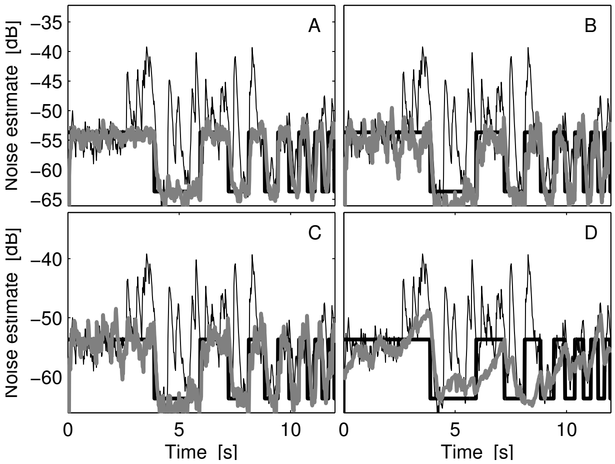

Fig. 2. Noise tracking performance at FFT bin corresponding to $f = 1250$ Hz. Thin black line: noisy psd. Thick black line: true noise psd. Thick gray line: Noise psd estimate. A) Proposed algorithm (PROP). B) WOL [9]. C) LEF [8]. D) HEN [5]

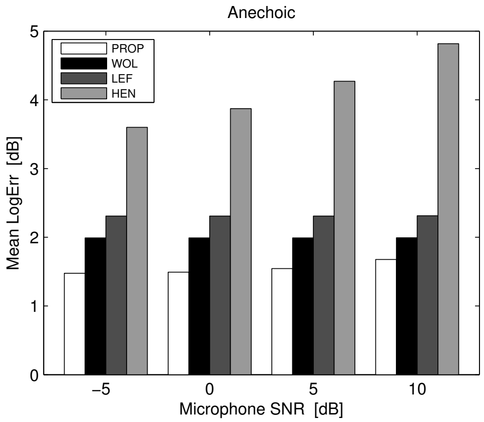

Fig. 3. Average LogErr scores for the proposed algorithm and algorithms WOL [9], LEF [8], and HEN [5].

4.2. Cocktail Party Scenario

We now study the performance in a more realistic acoustic scenario, namely a cocktail party situation. The target sound source is located directly in front as before, and 10 competing speakers are located roughly uniformly in a 10.4 × 12.7 m lecture room, see Fig. 4. As before, the microphone array consists of four microphones, but here we simulate a binaural hearing aid setup, consisting of two pairs of microphones (distance 1 cm) on each ear of a listener (we assume the hearing aids can communicate instantly, and without errors). Noisy microphone signals are generated by convolving target and noise signals with impulse responses measured from the relevant positions in the room to the microphones of dual-microphone behind-the-ear hearing aids placed on a head and torso simulator mannequin. The resulting noisy signals have a duration of 30 s. This acoustical scenario, although synthetically generated, is close to a realistic situation for a hearing aid user. As before, the initial part of each noisy signal realization is a noise-only region used for initialization.

Fig. 5 shows LogErr for the cocktail party scenario, averaged across the 20 different target speakers, for various input SNRs; when computing the input SNR, reflections of the target later than 20 ms of the direct sound are considered noise. For each target speaker, a new random permutation of the competing speakers is used. The proposed method is still

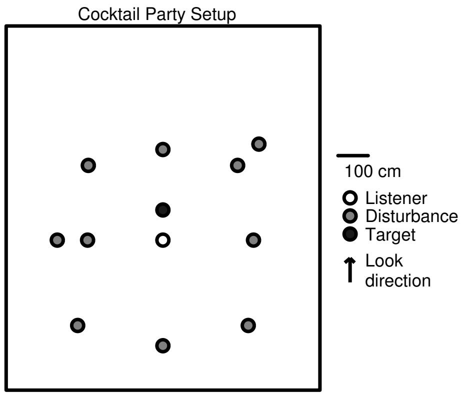

Fig. 4. Configuration of cocktail party scenario in a 10.4 × 12.7 m lecture room. The target source is located directly in front of the listener at a distance of 1 meter.

better than the alternatives. The single-channel method HEN performs better here because babble noise of 10 speakers is more stationary than the difficult modulated noise source in the previous example.

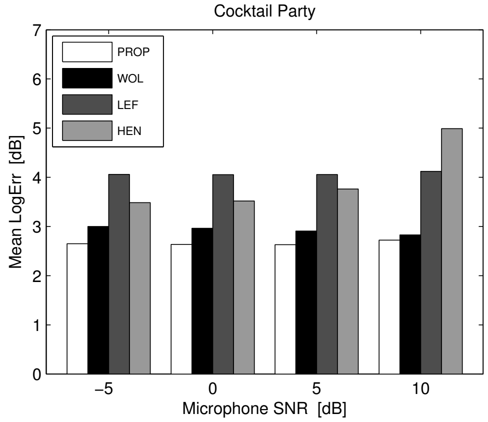

Fig. 5. Average LogErr scores for the proposed algorithm and algorithms WOL [9], LEF [8], and HEN [5].

5. CONCLUSION

Any speech enhancement algorithm, whether multi- or single-microphone, relies on knowledge of the disturbing noise source(s), and often second-order noise signal statistics are estimated based on the available noisy microphone signals. In this context, we have presented a maximum likelihood based method for estimating the inter-microphone covariance matrix of the noise impinging on a microphone array. One possible use of this covariance matrix is for estimating the psd of the noise entering the single-channel post filter in a beamformer-and-post filter speech enhancement system. We demonstrate in simulation experiments that using the multiple microphone signals for estimating this noise psd, as proposed in this paper, is advantageous to estimating the noise psd directly from the output signal of the beamformer.

6. REFERENCES

[1] S. Doclo and M. Moonen, "GSVD-based Optimal Filtering for Single and Multimicrophone Speech Enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, September 2002.

[2] A. Spriet, M. Moonen, and J. Wouters, "Spatially Pre-Processed Speech Distortion Weighted Multi-Channel Wiener Filtering for Noise Reduction," *Signal Processing*, vol. 84, pp. 2367–2387, December 2004.

[3] K. U. Simmer, J. Bitzer, and C. Marro, "Post-Filtering Techniques," in *Microphone Arrays – Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. 2001, Springer Verlag.

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

[5] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise psd tracking with low complexity," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2010, pp. 4266–4269.

[6] R. Zelinski, "A Microphone Array With Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 1988, vol. 5, pp. 2578–2581.

[7] I. A. McCowan and H. Bourlard, "Microphone Array Post-Filter Based on Noise Field Coherence," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[8] S. Lefkimmiatis and P. Maragos, "Optimum post-filter estimation for noise reduction in multichannel speech processing," in *Proc. 14th European Signal Processing Conference*, 2006.

[9] T. Wolff and M. Buck, "Spatial maximum a posteriori post-filtering for arbitrary beamforming," in *Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2008.

[10] L. J. Griffiths and C. W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming," *IEEE Trans. Antennas Propagat.*, vol. 30, no. 1, pp. 27–34, January 1982.

[11] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Third edition, 1996.