A MULTI-CHANNEL POSTFILTER BASED ON THE DIFFUSE NOISE SOUND FIELD

Lukas Pfeifenberger[1] *and Franz Pernkopf*[1]

[1] Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria
lukas.pfeifenberger@alumni.tugraz.at, pernkopf@tugraz.at

ABSTRACT

In this paper, we present a multi-channel *Directional-to-Diffuse Postfilter* (DD-PF), relying on the assumption of a directional speech signal embedded in diffuse noise. Our postfilter uses the output of a superdirective beamformer like the *Generalized Sidelobe Canceller* (GSC), which is projected back to the microphone inputs to separate the sound field into its directional and diffuse components. From these components the SNR at the output of the beamformer can be derived without needing a *Voice Activity Detector* (VAD). The SNR is used to construct a noise cancelling Wiener filter. In our experiments, the developed algorithm outperforms two recent postfilters based on the Transient Beam to Reference Ratio (TBRR) and the Multi-Channel Speech Presence Probability (MCSSP).

Index Terms— beamforming, multi-channel postfilter, diffuse sound field

1. INTRODUCTION

Speech intelligibility is a paramount issue in modern telecommunication systems. In many applications, background noise is the primary source of speech degradation. While single-channel speech enhancement systems require an inherent trade-off between noise reduction and speech quality, multi-channel speech enhancement systems also exploit the spatial information of the sound field and, thereby achieve a better performance. For this purpose, superdirective beamformers like the *Generalized Sidelobe Canceller* (GSC) [1, 2] in conjunction with multi-channel postfilters have gained the most attraction over the last decade. In this paper we assume a diffuse noise sound field, which can be found in a wide range of applications, such as car interiors, subway stations or roadside emergency telephones [3]. Further, we assume that the speaker is located close to the array, resulting in strong directional components in the *Acoustic Transfer Functions* (ATFs). We therefore model the ATFs as simple time delays, which can be identified by estimating the *Direction of Arrival* (DOA) using one of the algorithms discussed in [4]. The assumption of a diffuse noise sound field has already been used in postfilter concepts like [3] and [5], where a Wiener postfilter is derived from the speech and noise *Power Spectral Densities* (PSDs) at the beamformer output. However, many of these postfilters rely on a VAD and an accurate speech PSD estimate. A comprehensive overview of these methods is given in [6].

Our *Direct-to-Diffuse Postfilter* (DD-PF) algorithm estimates the SNR at the beamformer output by splitting the sound field at the microphones into its directional and diffuse components, using only the assumption of a diffuse noise field. This approach is inspired by the *Signal to Reverberant Ratio* (SRR) [7] and the *multi-channel SNR* in [8]. Other approaches to multi-channel postfilters are, for example: the *Transient Beam to Reference Ratio* (TBRR) [2], which relies on the ratio of transient energies in the beamformer output and in the output of the blocking matrix. These transient energies are determined using noise floor estimates in both the beamformer output and the blocking matrix outputs. The *Multi-Channel Speech Presence Probability* (MC-SPP) [8] algorithm can also be used without a beamformer, as it directly estimates the noise PSD matrix based on an a-priori speech presence probability and recursive averaging. In a similar approach given by [9], the SRR is mapped into a *speech absence probability* (SAP) used for recursive noise PSD estimation. Unlike these approaches, the performance of our postfilter only depends upon target leakage in the blocking matrix, and the diffuse noise field assumption.

This paper is organized as follows: Section 2 verifies the assumptions about the sound fields. Section 3 introduces the signal model and the beamformer. The DD-PF is derived in Section 4. Section 5 presents the experimental setup and performance results, where our postfilter is compared with two other approaches: the TBRR [2] and the MC-SPP [8]. The performance and speech quality is evaluated by using the *Perceptual Evaluation Methods for Audio Source Separation* (PEASS) Toolkit [10, 11]. Section 6 concludes the paper.

2. VERIFICATION OF THE SOUND FIELDS

In our setup, we assume a hands-free telephone situated in a noisy environment. In such a scenario, the speaker is located much closer to the microphone array than the noise source(s). Hence, a mostly directional speaker sound field and a diffuse

noise sound field is expected. To verify these assumptions, we placed $M = 4$ microphones in a linear array with an inter-microphone distance of $d = 5$ cm. The array is located in a 5×8 m wide hall with a RT60 ≈ 550 ms (see Figure 1).

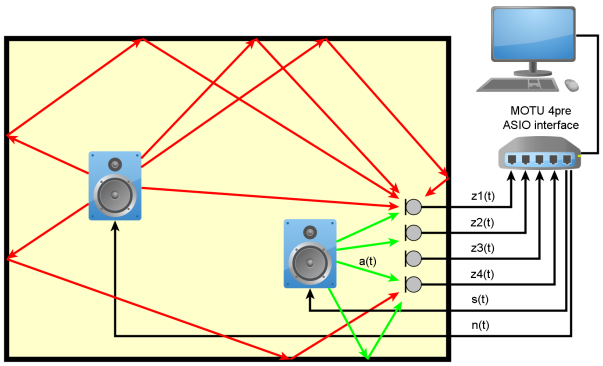

Fig. 1. Setup with a linear array consisting of $M = 4$ microphones at an inter-microphone distance of $d = 5$ cm.

To simulate the speaker, a loudspeaker is placed at a distance of 0.5 m at a DOA of 0° in front of the array. For the diffuse background noise, a second loudspeaker is placed 5 m away from the array. Using the MLS technique for room impulse measurement [12], it can be verified that the speaker sound field has a strong directional component and the noise sound field is mainly diffuse. Figure 2 shows the squared coherence for both scenarios. This result is similar to [3].

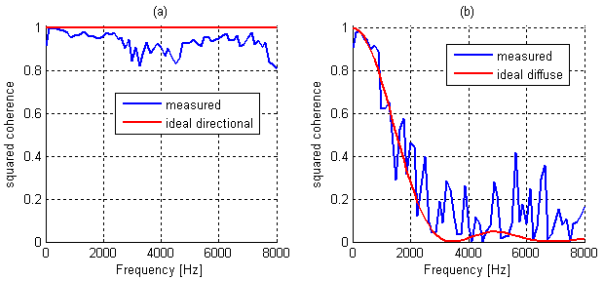

Fig. 2. Measurement of the squared coherence using the first two microphones for the loudspeaker at the position of (a) 0.5 m and (b) 5 m.

3. SIGNAL MODEL

In Figure 1 we denote the ambient noise at the m^{th} microphone as $n_m(t)$, and the ATF from the speaker to the m^{th} microphone as $a_m(t)$. With these definitions, the signal model can be written as $z_m(t) = a_m(t) * s(t) + n_m(t)$ in time-domain. In the fourier-domain $Z_m(j\Omega) = A_m(j\Omega)S(j\Omega) + N_m(j\Omega)$. Covering all M microphones, the signal model can be written in a more compact vector notation as

$$\boldsymbol{Z}(j\Omega) = \boldsymbol{A}(j\Omega)S(j\Omega) + \boldsymbol{N}(j\Omega). \tag{1}$$

While the proposed postfilter can be used in conjunction with any beamformer, we used the GSC for both its robustness and simplicity. It has been implemented as suggested in [2, 13, 14]. Its filter weights are given as $\boldsymbol{W}(j\Omega) = \boldsymbol{F}(j\Omega) - \boldsymbol{H}(j\Omega)\boldsymbol{B}(j\Omega)$, with the *delay and sum* beamformer $\boldsymbol{F}(j\Omega)$, the *blocking matrix* $\boldsymbol{B}(j\Omega)$ and an *adaptive interference canceler* $\boldsymbol{H}(j\Omega)$.

Due to the mainly directional speaker sound field encountered in Section 2, we modeled the ATFs as simple time delays, i.e. $\hat{A}_m(\Omega) = e^{jkd_m \sin\Theta}$, where $k = \frac{\omega}{c}$ is the wave number, d_m is the distance between the m^{th} microphone and an arbitrary reference point [1], and c is the speed of sound. Since the blocking matrix depends on the ATFs, target leakage might occur as a consequence of undermodeling, resulting in a degraded speech signal at the GSC output. However, we found the *signal blocking factor* [15] to be about 16dB in our experiments, which seems quite sufficient.

If the beamformer is steered towards the speech source, e.g. $\hat{\boldsymbol{A}}(j\Omega) \approx \boldsymbol{A}(j\Omega)$, all sounds originating from that direction are allowed to pass, since $\boldsymbol{W}^H(j\Omega)\hat{\boldsymbol{A}}(j\Omega) \approx 1$. This includes the speaker signal, and the portion of the noise impinging from that direction [1]. The beamformer output can therefore be written as

$$\begin{aligned} Y(j\Omega) &= \boldsymbol{W}^H(j\Omega)\boldsymbol{Z}(j\Omega) \\ &= \hat{S}(j\Omega) + \boldsymbol{W}^H(j\Omega)\boldsymbol{N}(j\Omega), \end{aligned} \tag{2}$$

where $\hat{S}(j\Omega)$ is the estimate of the speech source, and $\boldsymbol{W}^H(j\Omega)\boldsymbol{N}(j\Omega)$ is the noise component coming from the direction of the speaker.

4. MULTI-CHANNEL POSTFILTER

Our DD-PF algorithm estimates the SNR at the beamformer output without the need for a speech PSD estimate or a VAD. This is achieved by back-projecting the GSC output $Y(j\Omega)$ to the microphone signals $\boldsymbol{Z}(j\Omega)$ using the ATF model $\hat{\boldsymbol{A}}(j\Omega)$, we obtain

$$\begin{aligned} \hat{\boldsymbol{Z}}' &= \hat{\boldsymbol{A}}Y = \hat{\boldsymbol{A}}\hat{S} + \hat{\boldsymbol{A}}\boldsymbol{W}^H\boldsymbol{N}, \\ \hat{\boldsymbol{Z}}'' &= \boldsymbol{Z} - \hat{\boldsymbol{Z}}' \approx [\boldsymbol{I} - \hat{\boldsymbol{A}}\boldsymbol{W}^H]\boldsymbol{N}, \end{aligned} \tag{3}$$

assuming $\hat{\boldsymbol{A}}\hat{S} = \boldsymbol{A}S$. This assumption holds if the target leakage in the blocking matrix is low. The frequency argument $j\Omega$ has been omitted for brevity. It can be easily seen that $\hat{\boldsymbol{Z}}'$ denotes the directional signal components, and $\hat{\boldsymbol{Z}}''$ the remaining diffuse components. Due to statistical independence of the speech and the noise signal, the spatial PSD matrices of $\hat{\boldsymbol{Z}}'$ and $\hat{\boldsymbol{Z}}''$ can be written as

$$\begin{aligned} \boldsymbol{\Phi}_{\hat{Z}'\hat{Z}'} &= \hat{\boldsymbol{A}}\boldsymbol{\Phi}_{\hat{S}\hat{S}}\hat{\boldsymbol{A}}^H + \hat{\boldsymbol{A}}\boldsymbol{W}^H\boldsymbol{\Phi}_{NN}\boldsymbol{W}\hat{\boldsymbol{A}}^H \\ &= \boldsymbol{\Phi}_{\hat{S}'\hat{S}'} + \boldsymbol{\Phi}_{\hat{N}'\hat{N}'} \quad \text{and} \\ \boldsymbol{\Phi}_{\hat{Z}''\hat{Z}''} &\approx [\boldsymbol{I} - \hat{\boldsymbol{A}}\boldsymbol{W}^H]\boldsymbol{\Phi}_{NN}[\boldsymbol{I} - \boldsymbol{W}\hat{\boldsymbol{A}}^H] \\ &= \boldsymbol{\Phi}_{\hat{N}''\hat{N}''}. \end{aligned} \tag{4}$$

In [8], a multi-channel SNR as generalization from the single-channel case was defined as $\xi = \mathrm{Tr}(\mathbf{\Phi}^{-1}_{\hat{N}'\hat{N}'}\mathbf{\Phi}_{\hat{S}'\hat{S}'})$. Similarly, we evaluate only the power ratio of the main diagonals of these PSD matrices as

$$\xi = \frac{\mathrm{Tr}\left(\mathbf{\Phi}_{\hat{S}'\hat{S}'}\right)}{\mathrm{Tr}\left(\mathbf{\Phi}_{\hat{N}'\hat{N}'}\right)}, \tag{5}$$

because both PSD matrices $\mathbf{\Phi}_{\hat{S}'\hat{S}'}$ and $\mathbf{\Phi}_{\hat{N}'\hat{N}'}$ represent purely directional sound fields. Additionally, we can circumvent the numerically ill-conditioned matrix inversion of $\mathbf{\Phi}_{\hat{N}'\hat{N}'}$, caused by strong spatial correlations for low frequencies. However, we cannot measure $\mathbf{\Phi}_{\hat{S}'\hat{S}'}$ or $\mathbf{\Phi}_{\hat{N}'\hat{N}'}$ directly, but Eqn. (5) can be expressed as

$$\xi = \frac{\mathrm{Tr}\left(\mathbf{\Phi}_{\hat{Z}'\hat{Z}'}\right)}{\mathrm{Tr}\left(\mathbf{\Phi}_{\hat{Z}''\hat{Z}''}\right)} \frac{\mathrm{Tr}\left(\mathbf{\Phi}_{\hat{N}''\hat{N}''}\right)}{\mathrm{Tr}\left(\mathbf{\Phi}_{\hat{N}'\hat{N}'}\right)} - 1. \tag{6}$$

By assuming an ideal spherical diffuse noise sound field at the microphones, the noise PSD matrix $\mathbf{\Phi}_{NN}$ can be written as

$$\mathbf{\Phi}_{NN} = \Phi_{NN}\mathbf{\Gamma}_{NN}, \tag{7}$$

where Φ_{NN} denotes the unknown PSD of the noise source, and the elements of the spatial coherence matrix $\mathbf{\Gamma}_{NN}$ are defined as the coherence function [16] between the i^{th} and the j^{th} microphone, i.e.

$$\Gamma_{N_iN_j}(j\Omega) = \frac{\sin(kd_{ij})}{kd_{ij}}, \tag{8}$$

where d_{ij} is the distance between microphone i and j. Using Eqn. (7), the ratio $\frac{\mathrm{Tr}\left(\mathbf{\Phi}_{\hat{N}''\hat{N}''}\right)}{\mathrm{Tr}\left(\mathbf{\Phi}_{\hat{N}'\hat{N}'}\right)}$ in Eqn. (6) is obtained by

$$\frac{\mathrm{Tr}\left(\mathbf{\Phi}_{\hat{N}''\hat{N}''}\right)}{\mathrm{Tr}\left(\mathbf{\Phi}_{\hat{N}'\hat{N}'}\right)} = \frac{\mathrm{Tr}([\boldsymbol{I} - \hat{\boldsymbol{A}}\boldsymbol{W}^H]\mathbf{\Gamma}_{NN}[\boldsymbol{I} - \boldsymbol{W}\hat{\boldsymbol{A}}^H])}{\mathrm{Tr}(\hat{\boldsymbol{A}}\boldsymbol{W}^H\mathbf{\Gamma}_{NN}\boldsymbol{W}\hat{\boldsymbol{A}}^H)}, \tag{9}$$

using the ATF model $\hat{\boldsymbol{A}}$ and the beamforming filter $\boldsymbol{W}$. The coherence matrix $\mathbf{\Gamma}_{NN}$ is a constant. The directional and diffuse component of the input signal, $\hat{\boldsymbol{Z}}'$ and $\hat{\boldsymbol{Z}}''$, are estimated online using Eqn. (3). Their respective PSDs are found by recursive averaging, e.g. $\mathbf{\Phi}_{\hat{Z}'\hat{Z}',l} = \mathbf{\Phi}_{\hat{Z}'\hat{Z}',l-1}\alpha + (1-\alpha)\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}'^H$, where l is the frame index. The SNR ξ is obtained by using Eqn. (6). This SNR is then used to construct a Wiener filter. We used the *Optimally-Modified Log-Spectral Amplitude Estimator* (OM-LSA) algorithm [17], which is often found in noise cancelling applications.

5. EXPERIMENTS

5.1. Directivity Pattern

The proposed postfilter depends only on the current beamformer state defined by $\hat{\boldsymbol{A}}$ and $\boldsymbol{W}$. Therefore, the postfilter can easily be incorporated into the overall *Directivity Pattern* of the beamformer. The procedure described in [18] is used to simulate the theoretical directivity pattern with a two element array with $d = 5$ cm. The beamformer is fixed to look towards $0°$. In comparison to the beampattern of the GSC without a postfilter [4], figure 3 demonstrates the improved directivity especially for low frequencies.

To measure the real directivity pattern for comparison, we used the room from Figure 1, and the array mounted on a turntable. Figure 4 shows the measured beampattern for two microphones. Especially for low frequencies, it is sharper than the theoretical result. A cause for this effect could be minor gain differences in the microphones, which are not modeled by the simplified ATFs. However, it can be seen that signals impinging from outside $\pm 20°$ are completely suppressed. Increasing the number of microphones up to four did not change the directivity pattern significantly.

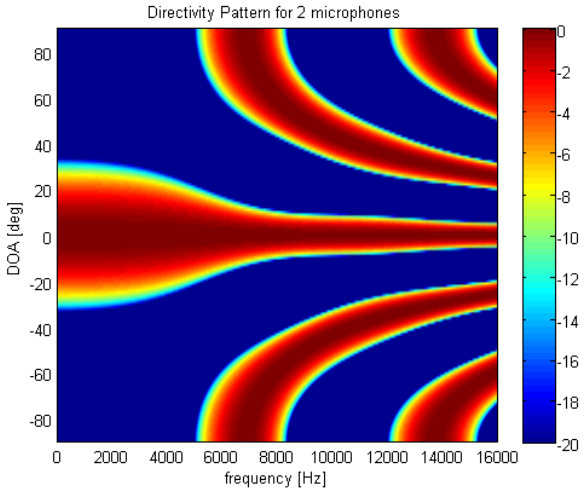

Fig. 3. Simulated directivity pattern for a two microphone beamformer with an aperture of $d = 5$ cm.

5.2. Experimental Setup

To test the speech quality of our MCSE system against a significant amount of speech data, the TIMIT [19], KCORS [20], and (KCOSS) [21] speech corpora have been used. The speech signals have been replayed with the loudspeaker at the 0.5 m position (see Figure 1). For the noise data, recordings from various sources, e.g. traffic noise, industry parks, subway stations and the NOIZEUS database have been replayed with the loudspeaker at the 5 m position. In total, about 60 minutes of test material has been generated. For comparison, we also implemented two other postfilter approaches – the MC-SPP approach and the TBRR. For the GSC beamformer we used a sparse blocking matrix $\boldsymbol{B}(j\Omega)$, which has the same performance as a dense eigenspace blocking matrix [22]. Its main benefit is the linear growth of computational complexity

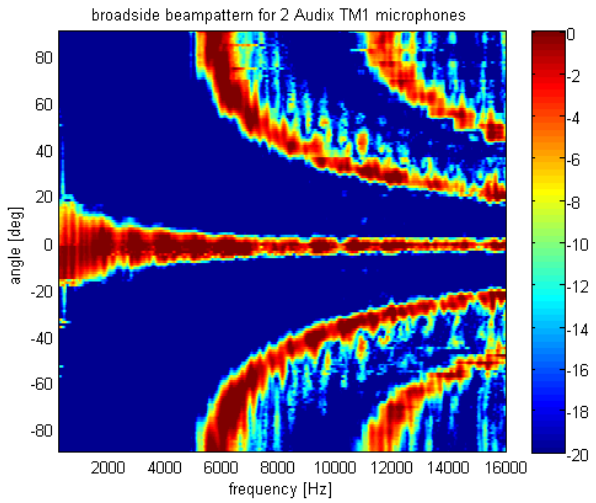

Fig. 4. Measured directivity pattern for 2 Audix-TM1 microphones placed $d = 5$ cm apart.

with the number of microphones. All GSC filters are implemented as non-causal FIR filters, to allow both positive and negative time delays [1]. The sampling frequency is $f_s = 16$ kHz and the SFFT length is 16ms, where we used a hanning window and 50% overlapping frames.

The PEASS Toolkit [10, 11] is used to evaluate the performance of the algorithms in terms of performance and perceptual speech quality. While PEASS might not be intended specifically for speech enhancement tasks, we found it represents the perceived speech quality much better than for example PESQ. PEASS delivers four scores: The *Target Perceptual Score* (TPS) measures the perceptual quality of the desired speech signal contained in the postfilter output. The *Interference Perceptual Score* (IPS) measures the influence of the residual noise components in the beamformer output. The *Artifact Perceptual Score* (APS) measures the influence of artifacts like musical noise generated by the algorithm. And the *Overall Perceptual Score* (OPS) provides a global measure of the perceptual quality of the enhanced output. Each score ranges from 0 to 100 and large values indicate better performance.

5.3. Results

Each algorithm is tested with a signal-to-interference ratio (SIR) ranging from -20 dB to +20 dB in 5 dB steps. Figure 5 shows the performance of the postfilters in terms of the PEASS measures. The OPS score of the TBRR and the MC-SPP postfilters are more or less equal. However, the TBRR performs better than the MC-SPP for the IPS and TPS score, and the APS score indicates that the TBRR introduces the most artifacts. The MC-SPP algorithm has the lowest IPS score, as it relies on the inversion of the spatial noise PSD matrix which is numerically unstable at low frequencies due to high signal correlations. The TBRR algorithm has the lowest APS score, as it relies on recursive noise floor estimation [23,24]. Depending on the instationarity of the noise, this technique is known to introduce musical artifacts. The speech quality of the proposed DD-PF does not depend on spatial speech PSD estimation or a VAD, but only on the estimate of the directional and the diffuse sound components $\mathbf{\Phi}_{\hat{Z}'\hat{Z}'}$ and $\mathbf{\Phi}_{\hat{Z}''\hat{Z}''}$. Their accuracy is determined by the shape of the assumed noise field and the target leakage in the blocking matrix. In our experiments, target leakage was quite low, and the noise sound field was nearly diffuse. Therefore, we achieved both a good speech quality and a good noise suppression at the same time, even for low frequencies. This can be seen by the OPS and TPS score.

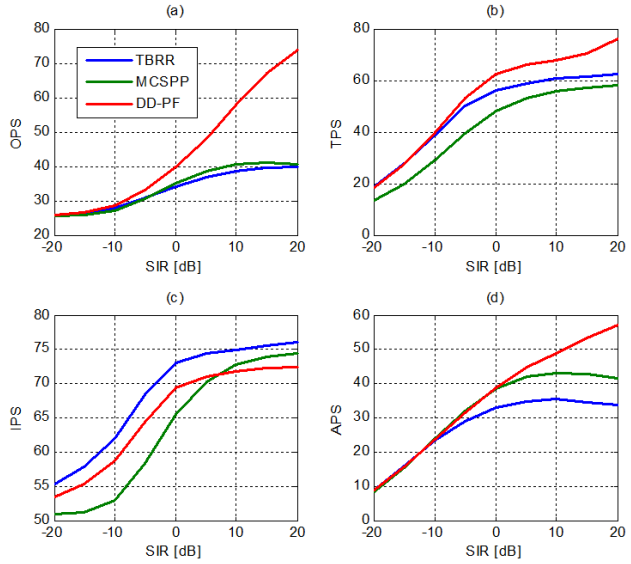

Fig. 5. Comparison of postfilters using PEASS measures; (a) OPS, (b) TPS, (c) IPS and (d) APS.

6. CONCLUSIONS

In this paper, we introduced the *Directional-to-Diffuse Postfilter* (DD-PF), which splits the sound field at the microphones into its directional and diffuse components to derive the SNR at the output of the beamformer, from which a noise reduction Wiener filter is derived. Unlike similar approaches, the algorithm does not depend on spatial speech PSD estimation or a VAD, but only on target leakage in the beamformer and the diffuse noise field assumption. In our experiments, these conditions have been sufficiently met. The achieved directivity pattern is selective even at low frequencies and the speech quality is significantly higher compared to the TBRR and MC-SPP approaches.

REFERENCES

[1] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, Berlin–Heidelberg–New York, 2008.

[2] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function gsc and postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 6, Nov. 2004.

[3] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 11, no. 6, Nov. 2003.

[4] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer, Berlin–Heidelberg–New York, 2008.

[5] N. Ito, N. Ono, E. Vincent, and S. Sagayama, "Designing the wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross spectra," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2818–2821, Nov. 2010.

[6] T. Wolff and M. Buck, "A generalized view on microphone array postfilters," *International Workshop on Acoustic Signal Enhancement*, Sept. 2010.

[7] O. Thiergart, G. Del Galdo, and E. A. P. Habets, "Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 309–312, 2012.

[8] Mehrez Souden, Jingdong Chen, Jacob Benesty, and Sofiene Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, Sept. 2011.

[9] M. Taseska and E. A.P. Habets, "Mmse-based blind source extraction in diffuse noise fields using a complex coherence-based a priori sap estimator," *International Workshop on Acoustic Signal Enhancement*, Sept. 2012.

[10] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," *10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 430–437, 2012.

[11] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, Sept. 2011.

[12] G.B. Stan, J.J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," Tech. Rep., Sound and Image Department, University of Liege, Belgium, 2002.

[13] I. Cohen, "Analysis of two-channel generalized sidelobe canceller (gsc) with post-filtering," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, Nov. 2003.

[14] I.Cohen, "Multichannel post-filtering in nonstationary noise environments," *IEEE Transactions on Signal Processing*, vol. 52, no. 5, May 2004.

[15] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, May 2009.

[16] H. Kuttruff, *Room Acoustics*, Spoon Press, London–New York, 5th edition, 2009.

[17] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, Apr. 2002.

[18] P. Vary and R. Martin, *Digital Speech Transmission*, Wiley, West Sussex, 2006.

[19] "Timit acoustic-phonetic continuous speech corpus," Website, Available online at `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1`; visited on January 16th 2013.

[20] "The kiel corpus of read speech vol. 1," Website, Available online at `http://www.ipds.uni-kiel.de/forschung/kielcorpus.en.html`; visited on January 16th 2013.

[21] "The kiel corpus of spontaneous speech vol. 1-3," Website, Available online at `http://www.ipds.uni-kiel.de/forschung/kielcorpus.en.html`; visited on January 16th 2013.

[22] M. G. Shmulik, S. Gannot, and I. Cohen, "A sparse blocking matrix for multiple constraints gsc beamformer," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2012.

[23] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, Sept. 2003.

[24] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, July 2001.