ESTIMATION OF THE NOISE CORRELATION MATRIX

Richard C. Hendriks

Signal and Information Processing Lab
Delft University of Technology
2628 CD Delft, The Netherlands
R.C.Hendriks@tudelft.nl

Timo Gerkmann

Sound and Image Processing Lab
KTH Royal Institute of Technology
10044 Stockholm, Sweden
gerkmann@kth.se

ABSTRACT

To harvest the potential of multi-channel noise reduction methods, it is crucial to have an accurate estimate of the noise correlation matrix. Existing algorithms either assume speech absence and exploit a voice activity detector (VAD), or make use of additional assumptions like a diffuse noise field. Therefore, these algorithms are limited with respect to their tracking speed and the type of noise fields for which they can estimate the correlation matrix.

In this paper we present a new method for noise correlation matrix estimation that makes no assumptions about the type of noise field, nor uses a VAD. The presented method exploits the existence of accurate single-channel noise PSD estimators, as well as the availability of one noise reference per microphone pair. For spatially and temporally non-stationary noise fields, the proposed method leads to improved performance compared to widely used state-of-the-art reference methods in terms of both segmental SNR and beamformer response error.

Index Terms— Speech enhancement, noise correlation matrix, multi-microphone

1. INTRODUCTION

In order to reduce degradations due to environmental noise, speech processing applications often exploit single- and multi-channel noise reduction algorithms. The advantage of multi-channel over single-channel noise reduction is that these algorithms are also able to exploit spatial filtering. Most often, single- and multi-channel noise reduction algorithms are implemented in the temporal-spectral domain, e.g., by computing a discrete Fourier transform (DFT). Single-channel noise reduction algorithms then estimate the clean speech DFT coefficients by applying a gain to the noisy DFT coefficients, e.g., [1][2][3], while multi-channel NR algorithms estimate the clean speech DFT coefficients by taking a weighted linear combination of several noisy DFT coefficients from multiple microphones and form a so-called beamformer, e.g., [4][5][6].

A crucial quantity on which all single-channel estimators depend, is the noise power spectral density (PSD). Estimation of this expected value is challenging, in particular for non-stationary noise sources. During recent years, there has been quite some attention for noise PSD estimation of non-stationary noise sources, see e.g., [7][8][9], and references therein.

The equivalence of the noise PSD for multi-channel noise reduction is the noise correlation matrix. Besides the noise PSD per microphone, the noise correlation matrix also caries the cross-correlation between the noise DFT coefficients at the different microphones. Multi-channel noise reduction methods like the minimum variance distortionless response (MVDR) beamformer [10] and the multi-channel Wiener filter [6] exploit the spatial information on the noise field that is contained in the noise correlation matrix. These methods can adaptively steer a beamformer in the direction of interest and reduce the effect of noise sources in other directions.

Correct knowledge of the noise correlation matrix is of high importance for multi-channel noise reduction algorithms in order to have the right suppression in a certain direction. Wrong estimates of the noise correlation matrix can either lead to the situation that disturbances from certain angles are not optimally suppressed, or, even worse, that noise coming from certain angles is amplified.

Similar as for the noise PSD, the noise correlation matrix is an unknown expected value that needs to be estimated from the noisy microphone signals. Although noise PSD estimation received significant interest, estimation of the noise correlation matrix appears to have been less explored. Where estimation of the noise PSD is challenging because noise sources can be non-stationary across time, estimation of the noise correlation matrix can be even more challenging since the noise field can also be spatially non-stationary.

A rather simple approach to estimate the noise correlation matrix is to exploit a voice activity detector (VAD) [11], and estimate the noise correlation matrix when a frequency bin contains no speech. When the noise sources are stationary across time and space, a VAD can be sufficient to estimate the noise correlation matrix. However, in many daily-life situations noise sources are non-stationary across time and space. Since a VAD does not allow to update the correlation matrix estimate during speech presence, it might happen that for temporally and/or spatially non-stationary noise fields the noise correlation matrix is estimated wrongly, resulting in a shape of the formed beam that is not optimally matched to the present noise field.

More recently, in [12], a method was proposed that does not rely on a VAD and can in theory estimate the noise correlation matrix when speech is present. This method relies on the assumption that the noise field is diffuse and estimates only the real part of the noise correlation matrix. However, assuming a diffuse noise field is not always realistic, in particular when noise sources are directional. Then, generally, the noise correlation matrix is complex-valued.

Instead of estimating the noise correlation matrix directly, procedures to estimate the filter coefficients adaptively were for example proposed by Griffiths and Jim [13], also known as the generalized sidelobe canceler (GSC). This adaptive method exploits an unconstrained least-mean-square (LMS) algorithm and allows estimation of the filter coefficients when the noise sources are rather stationary in both space and time. However, its performance degrades when

The research is supported by the Dutch Technology Foundation STW.

T. Gerkmann was with the Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany.

noise sources tend to be more non-stationary in space and time [14].

In this paper we present a method to estimate the noise correlation matrix for spatially and temporally non-stationary noise fields and an M-dimensional microphone array. The presented method makes no assumption on the type of noise field and can estimate the noise correlation matrix also when speech is present at the frequency bin under consideration. Similar as with the GSC we exploit the fact that given the propagation vector of the target source, a noise reference can be obtained. However, different than with the GSC the proposed method does not make use of LMS-based algorithms, but directly computes the elements of the noise correlation matrix.

2. NOTATION AND BASIC ASSUMPTIONS

We consider a multi-channel setup of M microphones. The noisy microphone signals are windowed on a frame-by-frame basis and transformed to the DFT domain, leading to the noisy DFT coefficients $Y_m(k,i)$, where k, i and m denote the frequency-bin, time-frame and microphone number, respectively. Similarly, we define the clean speech and noise DFT coefficients $S_m(k,i)$ and $N_m(k,i)$, respectively. The DFT coefficients are assumed to be random variables, indicated by upper case letters, and their corresponding realizations are indicated by lower case letters. Furthermore, bold faced letters indicate the use of matrices. We assume that the speech and noise DFT coefficients are uncorrelated and additive i.e.,

$$\mathrm{E}\{S_m(k,i)N_n(k,i)\} = 0 \quad \forall \quad k,i,m,n, \tag{1}$$

where $\mathrm{E}\{\cdot\}$ denotes the statistical expectation operator, and

$$Y_m(k,i) = S_m(k,i) + N_m(k,i), \tag{2}$$

respectively. The DFT coefficients are assumed to be independent across time and frequency, which allows us to neglect time- and frequency indices for ease of notation. Let $\mathbf{Y} = [Y_1, ..., Y_M]^T$ be a vector containing the noisy DFT coefficients for each of the M microphones. Similarly, we define $\mathbf{S} \in \mathbb{C}^M$ and $\mathbf{N} \in \mathbb{C}^M$ as vectors containing the M clean and noise microphone DFT coefficients, respectively. Further, let $\mathbf{\Sigma} = \mathrm{E}\{\mathbf{N}\mathbf{N}^H\} \in \mathbb{C}^{M\times M}$ be the noise correlation matrix. We assume the presence of a single target source whose acoustic path to the M microphones is modelled by the frequency dependent propagation vector $\mathbf{d} = [d_1, ..., d_M]^T$. The vector with clean speech DFT coefficients is therefore given by $\mathbf{S} = S\mathbf{d}$, where S is the clean speech DFT at the target speaker location. Altogether this leads to the following compact vector representation

$$\mathbf{Y} = \mathbf{S} + \mathbf{N} = S\mathbf{d} + \mathbf{N}. \tag{3}$$

3. ESTIMATION OF THE NOISE CORRELATION MATRIX

In order to estimate the noise correlation matrix $\mathbf{\Sigma}$, we make a distinction between its diagonal and off-diagonal elements. The diagonal elements are determined by the noise PSD per microphone channel. Using noise PSD estimators that can accurately estimate the noise PSD for non-stationary noise sources, e.g., [9], these diagonal elements can be considered to be known. That reduces the problem of estimating the noise correlation matrix to estimating the off-diagonal elements only. In order to estimate these cross-terms of the noise correlation matrix we make use of three assumptions. At first we assume that the acoustic path, represented by the propagation vector $\mathbf{d}$, is known. Secondly, we assume that the noise PSD per microphone channel is known, and finally, we make use of the assumption already expressed in Eq. (1), that is, the noise and speech DFT coefficients are assumed to be uncorrelated across time, frequency and microphones.

If $\mathbf{d}$ is perfectly known, we can create a noise reference for each microphone pair where the target signal is completely cancelled. Let $d(m)$ denote the mth element of the propagation vector $\mathbf{d}$. Further, let $d_{n,m} = d(n)/d(m)$ denote the complex scaling that needs to be applied to obtain the clean speech DFT coefficient at microphone number n from the clean speech DFT coefficient at microphone number m, that is, $S_n = d_{n,m}S_m$. A noise reference for element (n,m) of the correlation matrix is then obtained by

$$P_{n,m} \quad = \quad Y_n - d_{n,m}Y_m = N_n - d_{n,m}N_m. \tag{4}$$

In order to estimate a cross-term of the correlation matrix, the cross-correlation between $P_{n,m}$ and Y_m^* is computed, where $\cdot^*$ indicates complex conjugation. That is,

$$\begin{aligned} \mathrm{E}\{P_{n,m}Y_m^*\} &= \mathrm{E}\{N_nY_m^*\} - d_{n,m}\mathrm{E}\{N_mY_m^*\} &(5)\\ &= \mathrm{E}\{N_nN_m^*\} - d_{n,m}\mathrm{E}\{|N_m|^2\}, &(6)\end{aligned}$$

where use is made of the assumption that speech and noise DFT coefficients are uncorrelated and that the speech DFT coefficients are perfectly cancelled in $P_{n,m}$ as expressed by Eq. (4). The cross term $\mathrm{E}\{N_nN_m^*\}$ can now be solved from Eq. (6) as

$$\mathrm{E}\{N_nN_m^*\} = \mathrm{E}\{P_{n,m}Y_m^*\} + d_{n,m}\mathrm{E}\{|N_m|^2\}. \tag{7}$$

In practice, the expected values in Eq. (7) are unknown and have to be estimated. To estimate $\mathrm{E}\{|N_m|^2\}$, i.e., the noise PSD, we use the method presented in [9]. The term $\mathrm{E}\{P_{n,m}Y_m^*\}$ can be estimated by means of exponential smoothing, i.e.,

$$\begin{aligned}&\widetilde{\mathrm{E}}\{P_{n,m}(k,i)Y_m^*(k,i)\} = \\ &(1-\alpha)\widetilde{\mathrm{E}}\{P_{n,m}(k,i-1)Y_m^*(k,i-1)\} + \alpha p_{n,m}(k,i)y_m^*(k,i),\end{aligned} \tag{8}$$

where $\widetilde{\mathrm{E}}\{\cdot\}$ denotes an estimate of $\mathrm{E}\{\cdot\}$.

3.1. Reduction of Estimation Errors

The derivation of Eq. (6) from Eq. (5) relies on the assumption that speech and noise are uncorrelated. However, even when truly uncorrelated, estimation of $\mathrm{E}\{P_{n,m}Y_m^*\}$ based on realizations, i.e., $p_{n,m}(k,i)$ and $y_m^*(k,i)$, will give rise to a non-zero contribution due to the speech contained in $y_m^*(k,i)$. Therefore, instead of (6) we obtain

$$\begin{aligned}\widetilde{\mathrm{E}}\{P_{n,m}Y_m^*\} = \widetilde{\mathrm{E}}\{N_nN_m^*\} - d_{n,m}\widetilde{\mathrm{E}}\{|N_m|^2\} \\ + \widetilde{\mathrm{E}}\{N_nS_m^*\} - d_{n,m}\widetilde{\mathrm{E}}\{N_mS_m^*\}.\end{aligned} \tag{9}$$

If we take these estimation errors into account, and substitute Eq. (9) into Eq. (7) we obtain

$$\widetilde{\mathrm{E}_{\mathrm{err}}}\{N_nN_m^*\} = \widetilde{\mathrm{E}}\{N_nN_m^*\} + \widetilde{\mathrm{E}}\{N_nS_m^*\} - d_{n,m}\widetilde{\mathrm{E}}\{N_mS_m^*\}.$$

Similarly, for the cross-correlation term $\widetilde{\mathrm{E}}\{N_mN_n^*\}$, i.e., the complex conjugate of $\widetilde{\mathrm{E}}\{N_nN_m^*\}$, we obtain

$$\widetilde{\mathrm{E}_{\mathrm{err}}}\{N_mN_n^*\} = \widetilde{\mathrm{E}}\{N_mN_n^*\} + \widetilde{\mathrm{E}}\{N_mS_n^*\} - d_{m,n}\widetilde{\mathrm{E}}\{N_nS_n^*\}.$$

With $S_n = d_{n,m}S_m$ and the hermitian property of $\mathbf{\Sigma}$, we can reduce estimation errors by taking the average of $\widetilde{\mathrm{E}_{\mathrm{err}}}\{N_nN_m^*\}$ and

the complex conjugate of $\widetilde{\mathrm{E}_{\mathrm{err}}}\{N_m N_n^*\}$ as

$$\begin{aligned}&\widehat{\mathrm{E}}\{N_n N_m^*\}\\&= (\widetilde{\mathrm{E}_{\mathrm{err}}}\{N_n N_m^*\} + (\widetilde{\mathrm{E}_{\mathrm{err}}}\{N_m N_n^*\})^*)/2\\&= \widetilde{\mathrm{E}}\{N_n N_m^*\} + d_{m,n}^* \Im(\widetilde{\mathrm{E}}\{N_n S_n^*\})j - d_{n,m} \Im(\widetilde{\mathrm{E}}\{N_m S_m^*\})j,\end{aligned} \tag{10}$$

Where $\Im(\cdot)$ denotes the imaginary part and $j = \sqrt{-1}$. That is, estimating $\mathrm{E}\{N_n N_m^*\}$ by means of Eq. (10), removes the real part of the error terms $\widetilde{\mathrm{E}}\{N_n S_n^*\}$ and $\widetilde{\mathrm{E}}\{N_m S_m^*\}$. To further reduce the effect of the imaginary parts of these error terms on $\widehat{\mathrm{E}}\{N_n N_m^*\}$, the smoothing parameter α in Eq. (8) can be increased during periods where the signal-to-noise ratio (SNR) is rather high. In the experimental in Section 4 this smoothing constant is set at $\alpha = 0.9$. Whenever the a posteriori SNR, i.e., $\zeta = |Y_1|^2/\widetilde{\mathrm{E}}\{|N_1|^2\}$, exceeds 7.8 dB, α is increased to $\alpha = 0.99$ for that time-frequency point.

4. EXPERIMENTAL RESULTS

In this section we evaluate the performance of our proposed approach and compare its performance to three reference methods. Similar to a two-microphone hearing aid, we consider an $M = 2$ microphone endfire array with an inter-microphone distance of 1 cm. All signals are sampled at 8 kHz and processed on a frame-by-frame basis with a frame size of 256 samples with 50 % overlap, and are windowed using a square-root-Hann window.

The proposed method is implemented according to the expression in Eq. (10). For comparison, we employ the estimated correlation matrices in an MVDR beamformer, i.e.,

$$\hat{S} = \frac{\mathbf{d}^H \mathbf{\Sigma}^{-1} \mathbf{Y}}{\mathbf{d}^H \mathbf{\Sigma}^{-1} \mathbf{d}}, \tag{11}$$

with $\hat{S}$ the estimated clean speech DFT coefficient. The first reference method is an MVDR combined with the VAD based approach presented in [11] in order to estimate the noise correlation matrix. The VAD proposed in [11] is based on a minimum statistics noise power spectral density estimate [7]. With this approach, speech absence is decided, when the ratio between the recursively smoothed magnitude squared noisy observations and the estimated noise power is smaller than a certain threshold γ for all microphones. In the experimental results we set this threshold to $\gamma = 1.2$. Secondly, we also compare the results to an MVDR beamformer, where the covariance matrix of the noisy signal $\mathrm{E}\{\mathbf{Y}\mathbf{Y}^H\}$ is used, as under ideal conditions this is the same as using the covariance matrix of the noise signal (see e.g. [14]). To estimate $\mathrm{E}\{\mathbf{Y}\mathbf{Y}^H\}$, we recursively smooth $\mathbf{Y}\mathbf{Y}^H$ over time with a smoothing constant $\alpha = 0.9$. Thirdly, we also compare to the GSC summarized in [14, Table 47.2], which is justified by the fact that the analytic expression of the GSC equals the MVDR beamformer under ideal conditions.

As instrumental quality measure we compute the mean squared error between the ideal MVDR beamformer response, where the noise correlation matrix is obtained using the noise only signal, and the beamformer response based on filter coefficients that are estimated using the proposed or reference methods, that is,

$$\mathrm{BR}_{\mathrm{err}} = \frac{1}{|\mathcal{Q}|} \sum_{(\phi,k,i)\in\mathcal{Q}} \|\hat{R}(\phi,k,i) - R(\phi,k,i)\|^2, \tag{12}$$

where $\hat{R}(\phi,k,i)$ and $R(\phi,k,i)$ are the estimated and ideal beamformer responses, ϕ the direction of arrival, and $|\mathcal{Q}|$ the cardinality of the set of all (ϕ,k,i). In addition we use the segmental-SNR

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{|\mathcal{P}|} \sum_{p\in\mathcal{P}} 10\log_{10}\left(\frac{\|\mathbf{s}_p\|^2}{\|\mathbf{s}_p - \hat{\mathbf{s}}_p\|^2}\right), \tag{13}$$

where $\mathbf{s}_p$ and $\hat{\mathbf{s}}_p$ denote a clean and enhanced time-domain signal frame, and $\mathcal{P}$ an index set to denote all clean speech frames with energy within 35 dB of the maximum clean speech frame energy. As the MVDR beamformer is distortionless in the look-direction, we do not include a speech distortion measure in our evaluation.

For the target source we consider 5 female and 5 male speakers from the TIMIT database [15] and two different target source directions. That are, a target source in the endfire direction (0°) that is used to obtain the results in Fig. 1, and a target source at -60° to obtain the results in Fig. 2. We consider the speakers to be in free field and the propagation vector $\mathbf{d}$ to be perfectly known to fulfill Eq. (4). In a reverberant scenario, the estimation of the full acoustic path represented by the propagation vector $\mathbf{d}$ is difficult and (4) is usually not ideally fulfilled. However, we report that first experiments showed promising results also in a reverberant environment.

We disturb the speakers by modulated white Gaussian noise (Fig. 1) and temporally non-stationary train noise (Fig. 2). We use two different types of noise fields. While in subplots (a) of Figs. 1 and 2 the noise sources are spatially stationary, in subplots (b) we have a spatially stationary source at -40° and a second spatially stationary source at 100° that are repeatedly turned on for one second and off for one second, thus creating a spatially highly non-stationary noise field.

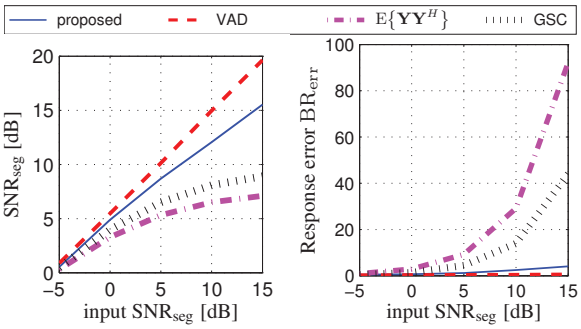
Spatially stationary directional sources can be cancelled by steering a null towards the noise source. For an MVDR beamformer in a dual channel setup and a single directional noise source, this means that the phase of the cross terms $\mathrm{E}\{N_n N_m^*\}$ needs to be estimated. As for a single spatially stationary noise source the phase of $\mathrm{E}\{N_n N_m^*\}$ does not change, a VAD based approach is expected to yield rather robust results even if the noise source is temporally non-stationary. This is confirmed by the results in Figs. 1(a) and 2(a). However, for the spatially highly non-stationary scenarios in Figs. 1(b) and 2(b), the VAD based approach is not capable of tracking the noise field and leads, compared to the proposed approach, to a loss of approximately 7 dB in terms of segmental SNR.

Compared to the GSC and the noisy correlation matrix based MVDR, the proposed approach leads to an improved performance for both noise sources and both the spatially stationary as well as the spatially non-stationary noise field. Similar as compared to the VAD based approach, large improvements are obtained for the situation that the noise is both temporally as well as spatially non-stationary, as visible in Figs. 1(b) and 2(b).

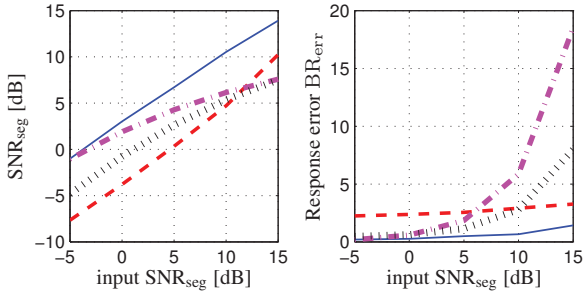
In general, it can be seen from Figs. 1(b) and 2(b), that the proposed approach clearly outperforms the competing approaches, where, dependent on the SNR and the amount of temporal and spatial non-stationarity of the noise source, improvements of 5 to 7 dB in terms of the segmental SNR can be obtained.

5. CONCLUSIONS

In this paper we presented a new method for the estimation of the noise correlation matrix. Existing methods can estimate the noise correlation matrix in stationary or slowly changing noise fields. However, for spatially and temporally non-stationary noise fields, existing methods are not able to accurately estimate the noise correlation matrix. The presented method exploits the availability of

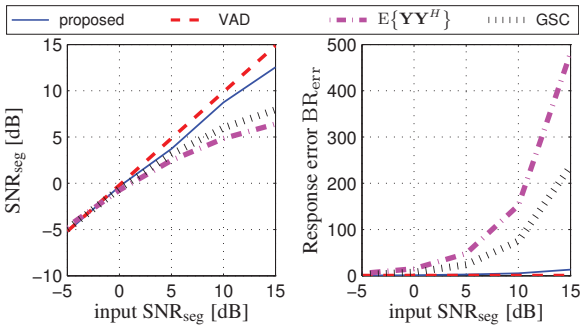


(a) Spatially stationary noise source at $-40°$.

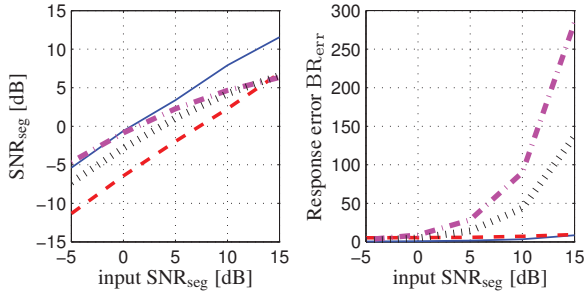


(b) One stationary noise source at $-40°$ and a second source at $100°$ that is repeatedly turned on and off. The present and absent states last one second each.

Fig. 1. Results for a target source at the endfire direction ($0°$) and a modulated white Gaussian noise source.



(a) Spatially stationary noise source at $-40°$.



(b) One stationary noise source at $-40°$ and a second source at $100°$ that is repeatedly turned on and off. The present and absent states last one second each.

Fig. 2. Results for a target source at $-60°$ and a train noise source.

accurate single-channel noise PSD estimators, as well as the availability of one noise reference per microphone pair. For spatially and temporally non-stationary noise fields, the proposed method leads to improved performance compared to exploiting a VAD or the generalized sidelobe canceler in terms of the segmental SNR and the beamformer response error.

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[2] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[3] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio Speech and Language Processing*, vol. 15, no. 6, pp. 1741 – 1752, August 2007.

[4] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 10, Oct. 1987.

[5] M. Brandstein and D. Ward (Eds.), *Microphone arrays*, Springer, 2001.

[6] S. Doclo, *Multi-microphone noise reduction and dereverberation techniques for speech applications*, Ph.D. thesis, Katholieke Universiteit Leuven, 2003.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[8] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 446–475, Sept. 2003.

[9] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *IEEE Int. Conf. Acoust, Speech, Signal Processing*, 2010, pp. 4266–4269.

[10] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *The Journal of the acoustical society of America*, vol. 54, no. 3, pp. 771 – 785, 1973.

[11] X. Zhang and Y. Jia, "A soft decision based noise cross power spectral density estimation for two-microphone speech enhancement systems," in *IEEE Int. Conf. Acoust, Speech, Signal Processing*, 2005, vol. 1, pp. 813–816.

[12] M. Rahmani, A. Akbari B. Ayad, and B. Lithgow, "Noise cross PSD estimation using phase information in diffuse noise field," *Elsevier Signal Processing*, vol. 89, pp. 703–709, 2009.

[13] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. 30, pp. 27–34, Jan. 1982.

[14] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., chapter 47, pp. 945–978. Springer Verlag, Berlin, Heidelberg, 2008.

[15] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology (NIST)*, 1988.