A PARTITIONED APPROACH TO SIGNAL SEPARATION WITH MICROPHONE AD HOC ARRAYS

Vincent Mohammad Tavakoli[†]*, Jesper Rindom Jensen*[†]*, Jacob Benesty*[‡]*, and Mads Græsbøll Christensen*[†]

[†]Audio Analysis Lab, AD:MT
Aalborg University, Denmark
{vmt,jrj,mgc}@create.aau.dk

[‡]INRS-EMT
University of Quebec, Montreal, Canada
benesty@emt.inrs.ca

ABSTRACT

In this paper, a blind algorithm is proposed for speech enhancement in multi-speaker scenarios, in which interference rejection is the main objective. Here, the ad hoc array is broken into microphone duples which are used to partition the array into local sub-arrays. The core algorithm takes advantage of differences in signal structure in each duple. A geometric mean filter is then used to merge the output signals obtained with different duples, and to form a global broadband maximum signal-to-interference ratio (SIR) enhancement apparatus. The resulting filter outputs are enhanced acoustic signals in terms of SIR, as shown with experiments.

Index Terms— Speech enhancement, blind signal separation, multichannel, microphone array, ad hoc array.

1. INTRODUCTION

Speech enhancement, through noise reduction and interference suppression, is a necessary part of many applications including, hearing aids, teleconferencing, etc. Microphone arrays have been useful means to speech enhancement for many years. Conventional array signal processing techniques assume known geometries for the array and take advantage of its manifold, often to estimate the direction of arrival (DOA) of acoustic wavefronts, such as in [1], and use it implicitly or explicitly to form beamformers to reject directional noise and interferences. The state-of-the-art minimum variance distortionless response (MVDR) beamformer [2], the controllable linearly constrained minimum variance (LCMV) beamformer [3], the informed parametric spatial filter [4], and cooperative noise reduction [5], follow this approach. For distributed and ad hoc microphone arrays, where the array geometry is unknown or the manifold is complex, enhancement apparatus based on other spatial or spectral fingerprints, such as the acoustic transfer function (ATF), the relative transfer function (RTF), and the pseudo-coherence vector, have been proposed during the past few years. Such attempts include the speech distortion weighted multichannel Wiener filter (SDW-MWF) [6, 7, 8], the nested generalized sidelobe canceler (GSC) [9], distributed GSC in local and global stages [10], geometrically constrained TRINICON [11], and the pseudo-coherence-based MVDR beamformer [12].

Without a known array geometry, speech enhancement algorithms with distributed and ad hoc microphone arrays require estimates of the statistics (usually in the form of covariance matrix) of noise, interference or desired signal. The speech presence probability (SPP) is frequently used to estimate the background noise [13, 14]; however, for restoring the interfered speech in multi-speaker scenarios, it is not sufficient. Templates of covariance matrices for desired and/or interfering signals are required in these scenarios. Unfortunately, these are not available in practical situations, which necessitate a blind speech signal separation prior to beamforming. Comprehensive review of subspace methods and specifically joint diagonalization can be found in [15]; however, algorithms which take into account the characteristics of ad hoc microphone arrays are still missing.

In this paper, such an approach is followed by dividing the ad hoc array into microphone duples, and then through searching for the signal structure inside these microphone duples and compare them using passing and rejecting masks in the short-time Fourier tranform (STFT) domain. The separation process is finalized through a global geometric mean beamformer. The rest of this paper is organized as follows. The problem is formulated in Section 2, where the signal model is defined in Subsection 2.1. The signal separation process using microphone duples, array partitioning, and global geometric mean filter are introduced in Subsections 2.2, 2.3, and 2.4, respectively. Results from experiments on harmonic signals are presented in Section 3, followed by conclusions in Section 4.

2. PROBLEM FORMULATION

2.1. Signal Model

We consider the problem of separation and enhancement of P spatially-constrained acoustic sources in a reverberant envi-

This work was supported in part by the Villum Foundation and the Danish Council for Independent Research, grant ID: DFF 1337-00084.

ronment using an ad hoc microphone array. The ad hoc array consists of $M_{\text{tot}} \geq P$ randomly positioned omni-directional microphones. It is assumed that at least one microphone is placed closer to each acoustic source than any other source; however, no further information is available a priori on the geometry of the problem.

The signal captured at the time t' with the m-th microphone ($m \in \{1, \ldots, M_{\text{tot}}\}$) is

$$y_m(t') = \sum_{p=1}^{P} x_m^p(t') + v_m(t'),$$
$$x_m^p(t') \triangleq g_m^p(t') * s^p(t'), \tag{1}$$

where $s^p(t')$ is the p-th source signal, $v_m(t')$ is the additive noise, $x_m^p(t')$ is the clean (but reverberated) received signal from source p, $g_m^p(t')$ is the acoustic impulse response from this source to the m-th microphone, and $*$ denotes convolution. Acoustic impulse responses are assumed to be time invariant. Furthermore, the signals $x_m^p(t')$ and $v_m(t')$ are assumed to be zero mean, stationary, real, broadband, and uncorrelated.

Assuming a sufficiently long analysis window, (1) can be written in the time-frequency domain as

$$Y_m(k,t) = \sum_{p=1}^{P} X_m^p(k,t) + V_m(k,t), \tag{2}$$
$$X_m^p(k,t) = G_m^p(k) S^p(k,t), \tag{3}$$

where $S^p(k,t)$, $X_m^p(k,t)$, $V_m(k,t)$, and $Y_m(k,t)$ are the STFTs of $s^p(t')$, $x_m^p(t')$, $v_m(t')$, and $y_m(t')$, respectively, at the time frame t and the frequency bin k, and $G_m^p(k)$ is the acoustic transfer function between source p and the m-th microphone of the ad hoc array.

The ad hoc array can be partitioned into a set of N sub-arrays, so that the n-th sub-array ($n \in \{1, \ldots, N\}$) contains $M(n) \geq 1$ microphones, and is bounded by a hull which does not intersect with the bounding hull for any other sub-array. The set of all microphones in the sub-array n, i.e., S_n, provides the STFT-domain signal that can be stacked into the vectors:

$$\mathbf{y}_n(k,t) = \text{Vect}\Big(\big\{Y_{m'}(k,t)\big\}_{m' \in S_n}\Big),$$
$$\mathbf{x}_n^p(k,t) = \text{Vect}\Big(\big\{X_{m'}^p(k,t)\big\}_{m' \in S_n}\Big),$$
$$\mathbf{v}_n(k,t) = \text{Vect}\Big(\big\{V_{m'}(k,t)\big\}_{m' \in S_n}\Big),$$

where $\text{Vect}()$ vectorises its operand in column form.

Then the stacked STFT-domain received and clean signals at the n-th sub-array are

$$\mathbf{y}_n(k,t) = \sum_{p=1}^{P} \mathbf{x}_n^p(k,t) + \mathbf{v}_n(k,t), \tag{4}$$
$$\mathbf{x}_n^p(k,t) = \mathbf{g}_n^p(k) S^p(k,t), \tag{5}$$

where

$$\mathbf{g}_n^p(k) = \text{Vect}\Big(\big\{G_{m'}^p(k)\big\}_{m' \in S_n}\Big), \tag{6}$$

is the stacked ATF for sub-array n.

The signal model based on the ATF in (4) can be modified further to obtain models based on the RTF and the pseudo-coherence (PC) vector. As shown in [12], the later is more informative for beamforming with sub-arrays, giving the chance to use the norm of pseudo-coherence vectors as a selection criterion. Thus, the signal model of interest in this paper is

$$\mathbf{y}_n(k,t) = \sum_{p=1}^{P} \boldsymbol{\rho}_{\mathbf{x}_n^p, X_r^p}(k,t) X_r^p(k,t) + \mathbf{v}_n(k,t), \tag{7}$$

where

$$\boldsymbol{\rho}_{\mathbf{x}_n^p, X_r^p}(k,t) = \frac{E\Big[\mathbf{x}_n^p(k,t) X_r^{p*}(k,t)\Big]}{E\Big[\Big|X_r^p(k,t)\Big|^2\Big]} \tag{8}$$

is the pseudo-coherence vector for sub-array n w.r.t the reference signal $X_r^p(k,t)$, with $E[\cdot]$ and superscript * being mathematical expectation and complex conjugate, respectively. Notice that the index r is different for distinct sources. Our blind signal separation algorithm does not explicitly depend on (7); however, as discussed in Section 4, it is possible to use the proposed algorithm in conjugation with (7) to form extended filters, such as the pseudo-coherence-based MVDR beamformer in [12].

The covariance matrix of $\mathbf{y}_n(k,t)$ can be expressed as

$$\begin{aligned}\boldsymbol{\Phi}_{\mathbf{y}_n}(k,t) &= E\big[\mathbf{y}_n(k,t)\mathbf{y}_n^\dagger(k,t)\big] \\ &= \sum_{p=1}^{P} \boldsymbol{\Phi}_{\mathbf{x}_n^p}(k,t) + \boldsymbol{\Phi}_{\mathbf{v}_n}(k,t),\end{aligned} \tag{9}$$

where the transcript † denotes the Hermitian transpose operator, $\boldsymbol{\Phi}_{\mathbf{x}_n^p}(k,t)$ is the covariance matrix of $\mathbf{x}_n^p(k,t)$, and $\boldsymbol{\Phi}_{\mathbf{v}_n}(k,t) = E\big[\mathbf{v}_n(k,t)\mathbf{v}_n^\dagger(k,t)\big]$ is the covariance matrix of the noise, $\mathbf{v}_n(k,t)$.

It is important to form sub-arrays in a way that a sub-array is more proximal to the p-th acoustic source than any other sub-array, while the rest of the sub-arrays are relatively closer to other acoustic sources. Then, many questions arise. Which microphones from the ad hoc array should be assigned to a sub-array? How the proximal microphones to one of the sources is found and grouped into a sub-array? How is the partitioned array used in signal separation/enhancement? etc. The rest of this section investigates the answers to these questions starting by reduction of the ad hoc microphone array into duples, which are used to differentiate between sources, and then extends the speech separation and enhancement process to all microphones in the ad hoc array.

2.2. Max-SIR Signal Separation with Duples

For every pair of randomly picked microphones from the ad hoc array, designated with the ordered duple $d = (m_1, m_2)$, the STFT-domain signals are also vectorisable. The stacked received signal at the duple d is

$$\mathbf{y}_d(k,t) = \mathbf{x}_d^p(k,t) + \mathbf{i}_d^p(k,t) + \mathbf{v}_d(k,t), \tag{10}$$

where

$$\mathbf{i}_d^p(k,t) = \sum_{\substack{q=1 \\ q \neq p}}^{P} \mathbf{x}_d^q(k,t) \tag{11}$$

is the interference component for the p-th acoustic source received with the duple. Without loss of generality, it can be assumed that microphone m_1 is relatively closer to source p while microphone m_2 is relatively closer to one of its interferences, so that m_1 and m_2 belong to two distinct sub-arrays. Then, the reference signal of interest that we want to separate from the mixture is $X_{m_1}^p(k,t)$. According to the signal model in (7), microphone m_2 weakly receives the reference signal, but captures the weighted summation of interferences (at least for the one which m_2 is proximal to) much more stronger than microphone m_1. This inspires the use of complex weights, $\mathbf{h}_d^p(k,t)$, to recover (separate) $X_{m_1}^p(k,t)$ from $\mathbf{y}_d(k,t)$, i.e.,

$$Z_d^p(k,t) = \mathbf{h}_d^{p\dagger}(k,t)\mathbf{y}_d(k,t). \tag{12}$$

The noise portion of the received signal is usually spatially white, and is also unknown at this point; therefore, the filter in (12) cannot suppress it. However, it is possible to find optimum weights for maximizing the output signal-to-interference-ratio, defined by

$$\mathrm{oSIR}\Big[\mathbf{h}_d^p(k,t)\Big] = \frac{\mathbf{h}_d^{p\dagger}(k,t)\mathbf{\Phi}_{\mathbf{x}_d^p}(k,t)\mathbf{h}_d^p(k,t)}{\mathbf{h}_d^{p\dagger}(k,t)\mathbf{\Phi}_{\mathbf{i}_d^p}(k,t)\mathbf{h}_d^p(k,t)}. \tag{13}$$

The $\mathrm{oSIR}\big[\mathbf{h}_d^p(k,t)\big]$ in (13) is the generalized Rayleigh quotient for the generalized eigenvalue problem of the form:

$$\mathbf{\Phi}_{\mathbf{x}_d^p}(k,t)\mathbf{h}_d^p(k,t) = \lambda^p \mathbf{\Phi}_{\mathbf{i}_d^p}(k,t)\mathbf{h}_d^p(k,t), \tag{14}$$

where $\mathbf{\Phi}_{\mathbf{x}_d^p}(k,t), \mathbf{\Phi}_{\mathbf{i}_d^p}(k,t) \in \mathbb{C}^{2\times 2}$ are Hermitian matrices, and $\mathbf{\Phi}_{\mathbf{i}^p} \in \mathcal{S}_{++}$ (a positive definite matrix), then $\mathbf{\Phi}_{\mathbf{x}^p} - \lambda^p \mathbf{\Phi}_{\mathbf{i}^p}$ is the Hermitian matrix pencil of order 2. As a result of the Courant-Fischer-Weyl min-max principle, $\mathbf{h}_d^p(k,t)$ is equal to the eigenvector regarding the largest eigenvalue of (14).

Equation (14) requires estimates of covariance matrices $\mathbf{\Phi}_{\mathbf{x}_d^p}(k,t)$ and $\mathbf{\Phi}_{\mathbf{i}_d^p}(k,t)$ which are not available a priori; however, it is possible to use the duple d as a differential sensor, i.e., by reversing the order of microphones, and take advantage of relative proximity of m_1 and m_2 to different sources. The reverse ordered duple $\tilde{d} = (m_2, m_1)$ is used together with duple d to form the approximate generalized eigenvalue problem of the form:

$$\mathbf{\Phi}_{\mathbf{y}_d}(k,t)\mathbf{h}_d^p(k,t) = \lambda^p \mathbf{\Phi}_{\mathbf{y}_{\tilde{d}}}(k,t)\mathbf{h}_d^p(k,t), \tag{15}$$

which is specifically correct when desired and interference acoustic signals occupy different frequency bins within a time-frame. Music and voiced speech are signals where it is expected that clean time-frequency bins be obtained from the weights equal to the principal eigenvector of (15).

The degree of freedom in each duple is only sufficient for rejecting the interfering source which is dominantly captured by microphone m_2 at time-frequency bin (k,t); however, it is possible to merge various duples by fixing the first element (m_1) and spanning the second one (m_2) over all microphones in sub-arrays other than the sub-array that m_1 belongs to. The merging process requires partitioning the ad hoc array based on information obtain from the max-SIR speech separation for all possible duples.

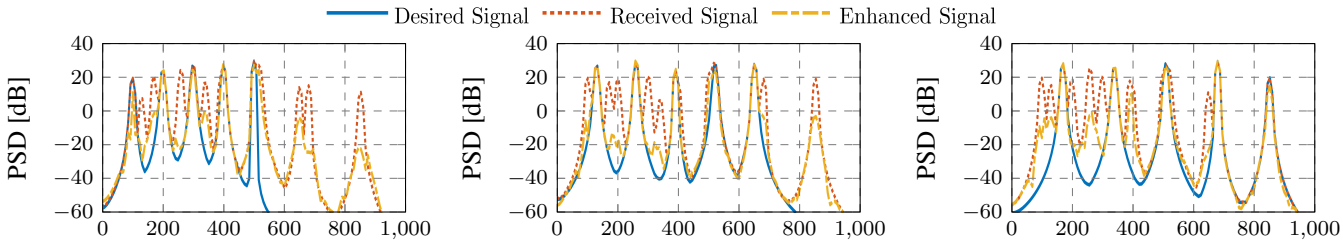
2.3. Partitioning of the Ad Hoc Array

With $m \in \{1, \ldots, M_{\mathrm{tot}}\}$ microphones in the ad hoc array, it is possible to form $\frac{1}{2}M_{\mathrm{tot}}!$ distinct duples. Part of the microphones may share similar geometrical constraints, such as relative closeness to one source, so that they should be grouped into a sub-array. A total number of $N = P$ sub-arrays are required, $\{S_n\}_{n=1}^p$, in order to separate and enhance P acoustic sources with the ad hoc array.

The set of narrowband filter outputs for distinct duples is

$$\Big\{Z_{m_1,m_2}(k,t)\Big\}_{\substack{m_1 \in \{1,\ldots,M_{\mathrm{tot}}\} \\ m_2 \in \{m_1+1,\ldots,M_{\mathrm{tot}}\}}}, \tag{16}$$

where the superscript p is dropped since it is not clear yet which source is extracted using each duple. However, (16) provides sufficient information to find this which leads to partitioning the ad hoc array. The set in (16) can be treated as the lower triangle part of a $M_{\mathrm{tot}} \times M_{\mathrm{tot}}$ matrix. The upper triangle can also be formed; however, it provides redundant information, so it is ignored. To extract the required information, the power spectral density of filter outputs at each time-frequency is compared to the power spectral density of the signal received with the first element of the duple to obtain two STFT-domain masks, namely, the passing and rejecting masks. The elements of the passing mask are equal to one for time-frequency bins that the signal is untouched up to a threshold, and the elements of the rejecting mask are equal to one for time-frequency bins that the signal is attenuated more than a different threshold. These STFT-domain masks, are then used as fingerprints to group microphones into sub-arrays based on the following criteria. The similarity of passing masks for two duples means that their first microphones are geometrically similar, and the similarity of rejecting masks for two duples means that their second microphones are geometrically similar.



(a) Desired source p_1 (b) Desired source p_2 (c) Desired source p_3

Fig. 1. Desired, received, and enhanced signals with the global geometric mean filter for three acoustic sources.

2.4. Geometric Mean Filter

Now that the P sub-arrays are formed in connection to the P acoustic sources, max-SIR filters using duples with the same first element, i, can be merged to remove all interferers of the source closer to microphone i. Suppose that the acoustic source p relatively close to microphone i is the desired source, i.e., $i \in S_p$, then the geometric mean filter:

$$Z_i^p(k,t) = \prod_{\forall j \in \{S_q\}_{q=1}^P \setminus S_p} Z_{d_{i,j}}^p(k,t) \qquad (17)$$

removes all interfering signals from the received signal. This can be regarded as cascading notch filters to remove the interferences and to restore the desired signal.

3. EXPERIMENTS

To show the applicability of the proposed algorithm, an experiment with synthesized signals using room impulse responses is performed here. The image method is used [16] to produce 3 acoustic sources, one of which is designated as the desired speaker, and the other two as interfering sources. The acoustic enclosure is a room of size $5\text{m} \times 5\text{m} \times 3\text{m}$ with reverberation time (T_{60}) equal to 150 ms. The acoustic sources are positioned at $\{2.5, 0.5, 1.5\}$, $\{4.23, 3.5, 1.5\}$, and $\{0.77, 3.5, 1.5\}$, while the microphones are positioned at $\{2.5, 1.54, 1.5\}$, $\{3.36, 3.0, 1.5\}$), and $\{1.63, 3.0, 1.5\}$. Here, the minimum number of microphones is used, i.e., one microphone being relatively closer to each acoustic source; however, no other prior knowledge is available to the enhancement apparatus regarding the problem. The acoustic clean sources are sampled at $f_\text{s} = 8$ kHz. For the STFT, the length of each time frame is set to 128 ms with 75% overlap among neighboring frames which corresponds to 32 ms hop. Recursive averaging with a forgetting factor, $\lambda = 0.2$, is used in estimation of covariance matrices. The desired speaker is iterated for three simulations such that for the first one, p_1 is the desired signal while interferences are p_2 and p_3. All synthetic sources are harmonic signals with pitch and harmonics according to Table 1. As can be seen in this table, some frequencies are close for different sources, so that it is expected to have less enhancement (interference suppression) or more of distortion in the enhanced signal. The results are shown in Fig. 1.

Table 1. Fundamental frequencies (F_0) and their harmonics ($H_i, i \in \{1, \ldots, 4\}$) for three acoustic signals.

	F_0	H_1	H_2	H_3	H_4
p_1	100	200	300	400	500
p_2	130	260	390	520	650
p_3	170	340	510	680	850

4. CONCLUSIONS AND FURTHER REMARKS

According to the results of the experiment in Section 3, merging microphone duples with a global geometric mean filter is an applicable method to separate acoustic sources with microphone ad hoc arrays. The results shows interference suppression of up to 40 dB in a reverberant environment; however, for certain frequencies the interference suppression is as low as 10 dB. In addition to speech enhancement, the proposed algorithm can also be used to provide separated signals suitable for estimating necessary parameters for other enhancement techniques. For instance, for each pair of microphones, m_1' and m_2', grouped into a sub-array according to the criteria in Subsection 2.3, it is possible to estimate the norm, $\aleph_{m_1',m_2'}^p$, and the group-delay, $\tau_{m_1',m_2'}^p$, of the pseudo-coherence vector between the received signals to estimate closeness of the microphone in terms of coherency. The pseudo-coherence vector between the received signals is a valid estimate of the pseudo-coherence vector between clean (but reverberated) signals for time-frequency bins at which the passing mask is one, which means that interferences are less probably exist. The norm and group delay are defined as

$$\aleph_{m_1',m_2'}^p = \left|\left|\boldsymbol{\rho}_{Y_{m_1'},Y_{m_2'}}\right|\right|_2^2, \qquad (18)$$

$$\tau_{m_1',m_2'}^p = \frac{\partial \angle \boldsymbol{\rho}_{Y_{m_1'},Y_{m_2'}}}{\partial \omega}, \qquad (19)$$

where the continuous coherencies are estimated from a sparse set of time-frequency bins using 2-D regression. Following this estimation, the state of the art pseudo-coherence-based MVDR beamformer can be used to enhance the acoustic signals without distortion.

5. REFERENCES

[1] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen, "Fast joint DOA and pitch estimation using a broadband MVDR beamformer," in *Proc. European Signal Processing Conf.*, Sept. 2013, pp. 1–5.

[2] C. Pan, J. Chen, and J. Benesty, "On the noise reduction performance of the MVDR beamformer in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 815–819.

[3] S. Karimian-Azari, J. Benesty, J. R. Jensen, and M. G. Christensen, "A broadband beamformer using controllable constraints and minimum variance," in *Proc. European Signal Processing Conf.*, Sept. 2014, pp. 666–670.

[4] O. Thiergart, M. Taseska, and E.A.P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 12, pp. 2182–2196, Dec. 2014.

[5] A. Hassani, A. Bertrand, and M. Moonen, "Cooperative integrated noise reduction and node-specific direction-of-arrival estimation in a fully connected wireless acoustic sensor network," *Elsevier Signal Process.*, vol. 107, no. 0, pp. 68 – 81, 2015.

[6] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "Variable speech distortion weighted multichannel wiener filter based on soft output voice activity detection for noise reduction in hearing aids," *Proc. Intl. Workshop Acoust. Echo Noise Control*, 2008.

[7] S. Markovich-Golan, S. Gannot, and I. Cohen, "A weighted multichannel wiener filter for multiple sources scenarios," in *Electrical & Electronics Engineers in Israel, IEEE 27th Convention of*, 2012, pp. 1–5.

[8] S. Markovich-Golan, S. Gannot, and I. Cohen, "Performance of the SDW-MWF with randomly located microphones in a reverberant enclosure," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 7, pp. 1513–1523, July 2013.

[9] O. Schwartz, S. Gannot, and E. A. P. Habets, "Nested generalized sidelobe canceller for joint dereverberation and noise reduction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., to appear*, 2015.

[10] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed GSC beamforming using the relative transfer function," in *Proc. European Signal Processing Conf.*, 2012, pp. 1274–1278.

[11] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, and W. Kellermann, "Geometrically constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, 2013, pp. 1–4.

[12] V. M. Tavakoli, J. R. Jensen, M. G. Christensen, and J. Benesty, "Pseudo-coherence-based MVDR beamformer for speech enhancement with ad hoc microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 2659 – 2663.

[13] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, Oct. 2011, pp. 145–148.

[14] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.

[15] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters – A Theoretical Study*, Number VII. Springer, 1st edition, 2011.

[16] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, Eindhoven, 2010.