Noise Power Spectral Density Estimation Using MaxNSR Blocking Matrix

Lin Wang, Timo Gerkmann, *Senior Member, IEEE*, and Simon Doclo, *Senior Member, IEEE*

***Abstract*—In this paper, a multi-microphone noise reduction system based on the generalized sidelobe canceller (GSC) structure is investigated. The system consists of a fixed beamformer providing an enhanced speech reference, a blocking matrix providing a noise reference by suppressing the target speech, and a single-channel spectral post-filter. The spectral post-filter requires the power spectral density (PSD) of the residual noise in the speech reference, which can in principle be estimated from the PSD of the noise reference. However, due to speech leakage in the noise reference, the noise PSD is overestimated, leading to target speech distortion. To minimize the influence of the speech leakage, a maximum noise-to-speech ratio (MaxNSR) blocking matrix is proposed, which maximizes the ratio between the noise and the speech leakage in the noise reference. The proposed blocking matrix can be computed from the generalized eigenvalue decomposition of the correlation matrix of the microphone signals and the noise coherence matrix, which is assumed to be time-invariant. Experimental results in both stationary and nonstationary diffuse noise fields show that the proposed algorithm outperforms existing blocking matrices in terms of target speech blocking ability, noise estimation and noise reduction performance.**

***Index Terms*—Blocking matrix, diffuse noise, microphone array, noise power spectral density (PSD) estimation, speech enhancement.**

I. Introduction

HANDS-FREE speech communication is desirable for many applications, such as mobile phones and automatic speech recognition systems, due to the provided convenience and flexibility. However, when the signals are recorded by distant microphones, they may be severely corrupted by interfering speakers or background noise. Single- or multi-microphone noise reduction algorithms are then required to enhance the desired speech signal. In comparison to single-microphone algorithms, which can only use temporal and spectral information, multi-microphone algorithms can additionally exploit spatial information and hence generally achieve a higher noise reduction performance, especially when the speech and the noise sources are spatially separated. This paper addresses a multi-microphone speech enhancement problem with one target speaker talking in a noisy background, which is a scenario that is often encountered in practice. In recent decades a large number of multi-microphone speech enhancement algorithms have been proposed, e.g., fixed and adaptive beamforming [1]–[19], multi-channel Wiener filtering [20]–[22], and blind source separation [23]–[27].

Although spatial filtering techniques work quite well for coherent interferers (point sources), their performance may decrease dramatically in diffuse noise fields, because the correlation of the noise components in the microphone signals is small. Hence spectral post-filtering is typically employed to enhance the output of spatial filtering [28], [29]. Two well-known post-filters have been proposed in [30], [31], which use the auto- and cross-power spectral densities (PSDs) of the multi-microphone signals to compute a Wiener post-filter. However, since the obtained post-filter is essentially constructed from the estimated PSDs of the speech and noise components in the microphone signals, applying it to the output signal after spatial filtering will lead to non-optimal results since the spectrum of the speech and noise components has been modified by the spatial filter. Thus, a better way is to estimate the noise PSD at the spatial filtering output and use it to compute the spectral post-filter.

This paper investigates a speech enhancement system based on the well-known generalized sidelobe canceller (GSC) structure, which consists of a fixed beamformer providing an enhanced speech reference, a blocking matrix providing a noise reference by suppressing the target speech, and a single-channel spectral post-filter. The spectral post-filter requires the PSD of the residual noise in the speech reference, which can, in principle, be estimated from the PSD of the noise reference. The performance of the blocking matrix (more specifically its target speech blocking ability) plays a key role in the speech enhancement performance of the post-filter. Speech leakage at the output of the blocking matrix will lead to overestimation of the noise PSD, giving rise to target speech distortion. It can be shown that the overestimation error is proportional to the ratio between the speech leakage energy and the noise energy at the output of the blocking matrix. Since, in general, it is difficult to cancel the target speech completely with e.g., a null space blocking matrix, a novel maximum noise-to-speech ratio (MaxNSR) blocking matrix is proposed in this paper,

Manuscript received November 06, 2014; revised March 17, 2015; accepted May 18, 2015. Date of publication June 01, 2015; date of current version June 04, 2015. This work was supported by the Postdoctoral Researcher Fellowship by the Alexander von Humboldt Foundation and by the Cluster of Excellence 1077 "Hearing4All," funded by the German Research Foundation (DFG). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Roberto Togneri.

L. Wang is with the Signal Processing Group, Department of Medical Physics and Acoustics, University of Oldenburg, Oldenburg 26111, Germany, and also with the Centre for Intelligent Sensing, Queen Mary University of London, London E1 4NS, U.K. (e-mail: wanglin_2k@yahoo.com; lin.wang@uni-oldenburg.de).

T. Gerkmann and S. Doclo are with the Signal Processing Group, Department of Medical Physics and Acoustics, University of Oldenburg, Oldenburg 26111, Germany (e-mail: timo.gerkmann@uni-oldenburg.de; simon.doclo@uni-oldenburg.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2438542

which aims to suppress the target speech by maximizing the noise-to-speech ratio at its output. Simulation results show that the noise PSD estimate based on the proposed MaxNSR blocking matrix is less affected by speech leakage than the traditional null space blocking matrix. It is further shown that the output signal of the MaxNSR blocking matrix can essentially be represented as a linear combination of the output signals of the null space blocking matrix in the sense of maximum noise-to-speech ratio. While several ways exist to combine the output signals of the null space blocking matrix for noise PSD estimation, the MaxNSR combination shows the best blocking ability.

The paper is organized as follows. Section II reviews related work on beamforming, blocking matrices, and noise PSD estimation. Section III defines the signal model used in the paper. Section IV introduces the GSC-based speech enhancement system along with the blocking matrix based noise PSD estimation. The MaxNSR blocking matrix is proposed in Section V. Extensive experimental results are given in Section VI to compare the performance of different noise PSD estimators (MaxNSR blocking matrix, null space blocking matrix, and single-channel noise PSD estimator) in stationary and non-stationary diffuse noise environments.

II. Review of Related Work

A. Beamforming and Blocking Matrices

Beamforming aims to extract a desired signal impinging on the array from a specific direction while treating all other signals as interfering sources. Beamformers can be designed either as a (data-independent) fixed beamformer, such as a delay-and-sum beamformer or a superdirective beamformer [8], [18], or as a (data-dependent) adaptive beamformer, which combines the spatial focusing of a fixed beamformer with adaptive noise suppression and hence typically reduces noise more efficiently than a fixed beamformer. Several criteria can be applied in the design of an adaptive beamformer [4], [6], [9], [16], e.g., linearly constrained minimum variance (LCMV), minimum variance distortionless response (MVDR), minimum mean-squared error (MMSE), and maximum speech-to-noise ratio (MaxSNR). For these adaptive beamformers, typically an estimate of the noise correlation matrix is required. A widely used approach is to calculate the noise correlation matrix in noise-only periods, with the assistance of a voice activity detector (VAD) [21], [22]. Some other methods that do not rely on a VAD are based on assumptions about the type of the noise field [11], [32], e.g., a spherically or cylindrically isotropic noise field. Another alternative is proposed in [34], where the noise correlation matrix is estimated based on the correlation between the noisy input signal and a noise reference. The noise reference is obtained by steering a null at the target speaker, whose propagation vector or relative transfer function is assumed to be known.

In recent years some perceptually motivated beamformers have been proposed, which allow a tradeoff between noise reduction and dereverberation [15], between the reduction of different noise types [19], or incorporating spatial cue preservation [22]. For instance, in [19], the noise correlation matrix, which is assumed to be known, is decomposed into coherent and incoherent parts. The decomposed noise correlation matrices are then used to derive a beamformer that provides a better control between spatially coherent and incoherent noise reduction. It can be shown that the traditional MVDR, LCMV, and multi-channel Wiener filter are special cases of this tradeoff beamformer.

An alternative (adaptive) implementation of the LCMV or MVDR beamformer is the generalized sidelobe canceler (GSC), consisting of a fixed beamformer, a blocking matrix, and an adaptive noise canceler [2]. The fixed beamformer provides a spatial focus on the speech source, creating a speech reference. The blocking matrix steers nulls in the direction of the speech source, creating one or more noise references. The adaptive noise canceler reduces the residual noise in the speech reference that is correlated with the noise references. Depending on the used blocking matrix, speech may however leak into the noise references of the standard GSC, especially in reverberant environments. A typical way to construct the blocking matrix is to use the projection matrix to the null space of the acoustic transfer function of the target speaker [7], [12], [13], [17]. Depending on the assumed acoustic model for the acoustic transfer function, various types of blocking matrices are obtained. For instance, the delay-and-subtract blocking matrix is derived assuming a pure-delay model under free-field conditions [2]. Since this model only considers the direct path between the target speaker and the microphones, the target speech blocking ability of the delay-and-subtract blocking matrix is quite limited in reverberant environments. As an improvement, a blocking matrix based on the relative transfer functions (RTFs) of the target speaker between the microphones has been proposed. Since this better captures reverberation, an improved target speech blocking ability is obtained [7], [17]. In [7] a least-squares approach is proposed to estimate the RTF by exploiting the nonstationary characteristics of the desired speech signal, while in [17] a generalized eigenvector decomposition approach is proposed to compute the RTF. In addition, the use of blind source separation to construct the blocking matrix has recently been proposed [35], [36]. Another way to limit speech leakage is to use an adaptive blocking matrix, which adapts when the target speaker is dominant [5], [10], [37], [38].

B. Noise PSD Estimation for Post-filter

The accuracy of the noise PSD estimation will directly determine the performance of the post-filter. Different noise PSD estimation methods have been proposed in the literature and use either single-channel or multichannel information.

A commonly used single-channel method is to calculate the noise PSD during speech pauses. This, however, relies on a voice activity detector (VAD) and is not suitable for non-stationary noise. Another well-known method is the minimum statistics algorithm, which estimates the noise PSD by tracking the spectral minima of the smoothed power spectrum of the noisy speech signal [39]. The drawback of the minimum statistics algorithm is that it can only track the noise PSD with a certain delay. Recently, MMSE-based noise PSD estimators, which can track the noise PSD with a smaller delay and thus are even able to handle non-stationary noise to some extent, have been proposed [40], [41].

In general, it is easier to estimate the noise PSD when multichannel information is available. Estimating the noise PSD based on the noise references provided by the blocking matrix in a GSC-structure has been proposed in [35], [36], [42]–[44]. Assuming that the noise references contain little speech leakage, they can be used for computing the PSD of the target noise, e.g., at the microphones or at the output of a spatial filter. It has been shown in [36], [42], [43] that the PSD of the noise reference and the target noise can be related with a frequency-dependent scaling factor, which can be calculated directly from the coefficients of the spatial filter and the noise coherence matrix. If the noise coherence matrix is assumed to be known and time-invariant, the multichannel blocking matrix based noise PSD estimation method can operate during speech-presence periods, which is advantageous especially in fast-changing nonstationary noise fields.

III. Signal Model and MaxSNR Beamformer

In this section we first define the used signal model and then briefly introduce the well-known MaxSNR beamformer, which will be used for the derivation of the proposed algorithm. All derivations throughout the paper will be performed in the frequency domain unless otherwise stated.

A. Signal Model

In this paper we aim to enhance a single target speaker recorded by M microphones in a noisy environment. In the short-time Fourier transform (STFT) domain the noisy microphone signals, $\boldsymbol{x}(k,l) = [x_1(k,l), \cdots, x_M(k,l)]^{\mathrm{T}}$, can be represented as

$$\boldsymbol{x}(k,l) = \boldsymbol{s}(k,l) + \boldsymbol{n}(k,l) = \boldsymbol{h}(k)u(k,l) + \boldsymbol{n}(k,l), \tag{1}$$

where k and l denote the frequency and the block indices, respectively; $u(k,l)$ denotes the clean speech signal; $\boldsymbol{h}(k) = [h_1(k), \cdots, h_M(k)]^{\mathrm{T}}$ denotes the time-invariant acoustic transfer function between the target speaker and the microphones; $\boldsymbol{n}(k,l) = [n_1(k,l), \cdots, n_M(k,l)]^{\mathrm{T}}$ is the noise component received by the microphones and $\boldsymbol{s}(k,l) = [s_1(k,l), \cdots, s_M(k,l)]^{\mathrm{T}}$ is the speech component received by the microphones. The superscript $(\cdot)^{\mathrm{T}}$ represents the transpose.

The speech and noise components in the microphone signals are assumed to be uncorrelated with each other, such that

$$\boldsymbol{\Phi}_{\mathrm{xx}}(k,l) = \boldsymbol{\Phi}_{\mathrm{ss}}(k,l) + \boldsymbol{\Phi}_{\mathrm{nn}}(k,l), \tag{2}$$

where

$$\begin{cases} \boldsymbol{\Phi}_{\mathrm{xx}}(k,l) = \mathrm{E}\{\boldsymbol{x}(k,l)\boldsymbol{x}^{\mathrm{H}}(k,l)\} \\ \boldsymbol{\Phi}_{\mathrm{ss}}(k,l) = \mathrm{E}\{\boldsymbol{s}(k,l)\boldsymbol{s}^{\mathrm{H}}(k,l)\} \\ \boldsymbol{\Phi}_{\mathrm{nn}}(k,l) = \mathrm{E}\{\boldsymbol{n}(k,l)\boldsymbol{n}^{\mathrm{H}}(k,l)\} \end{cases} \tag{3}$$

are the correlation matrices of the noisy speech, speech, and noise components, respectively; $\mathrm{E}\{\cdot\}$ denotes mathematical expectation and the superscript $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose. In addition, based on the single target speaker assumption, $\boldsymbol{\Phi}_{\mathrm{ss}}(k,l)$ is a rank-1 matrix, i.e., $\boldsymbol{\Phi}_{\mathrm{ss}}(k,l) = P_u(k,l)\boldsymbol{h}(k)\boldsymbol{h}^{\mathrm{H}}(k)$ with $P_u(k,l)$ the PSD of the clean speech signal [22].

The noise field is assumed to be spatially homogeneous, but can be either directional, isotropic or uncorrelated, with a time-invariant spatial coherence between the microphones. Under this assumption the noise correlation matrix can be expressed as

$$\boldsymbol{\Phi}_{\mathrm{nn}}(k,l) = P_n(k,l)\boldsymbol{\Gamma}_{\mathrm{nn}}(k), \tag{4}$$

where $P_n(k,l)$ is a time-varying scalar denoting the noise PSD at each time-frequency bin and $\boldsymbol{\Gamma}_{\mathrm{nn}}(k)$ is the time-invariant noise coherence matrix at each frequency. In general, $\boldsymbol{\Gamma}_{\mathrm{nn}}(k)$ can be estimated during noise-only periods as

$$\boldsymbol{\Gamma}_{\mathrm{nn}}(k) = \frac{\bar{\boldsymbol{\Phi}}_{\mathrm{nn}}(k)}{\frac{1}{M}\mathrm{tr}\{\bar{\boldsymbol{\Phi}}_{\mathrm{nn}}(k)\}}, \tag{5}$$

where $\mathrm{tr}\{\cdot\}$ represents the trace of a matrix and $\bar{\boldsymbol{\Phi}}_{\mathrm{nn}}(k)$ represents the noise correlation matrix averaged in noise-only periods, i.e.,

$$\bar{\boldsymbol{\Phi}}_{\mathrm{nn}}(k) = \frac{1}{L_n}\sum_{l\in\mathbb{L}_n}\boldsymbol{n}(k,l)\boldsymbol{n}^{\mathrm{H}}(k,l), \tag{6}$$

where $\mathbb{L}_n$ denotes the set of L_n noise-only frames. Alternatively, $\boldsymbol{\Gamma}_{\mathrm{nn}}$ can also be calculated when prior knowledge about the noise field is available. For instance, for a spherically diffuse noise field the coherence matrix can be expressed as [33]

$$\boldsymbol{\Gamma}_{\mathrm{nn}}^{ij}(k) = \mathrm{sinc}\left(\frac{2\pi k f_k d_{ij}}{c}\right), \tag{7}$$

where the superscript $(\cdot)^{ij}$ denotes the (i,j)-th entry of the matrix, f_k is the frequency at the k-th frequency bin, d_{ij} is the distance between the i-th and the j-th microphone, and c denotes the sound velocity.

Combining (2) and (4), it follows that

$$\boxed{\boldsymbol{\Phi}_{\mathrm{xx}}(k,l) = \boldsymbol{\Phi}_{\mathrm{ss}}(k,l) + P_n(k,l)\boldsymbol{\Gamma}_{\mathrm{nn}}(k)} \tag{8}$$

where $\boldsymbol{\Phi}_{\mathrm{ss}}(k,l)$ and $P_n(k,l)$ are unknown. The goal is to estimate the time-varying noise PSD $P_n(k,l)$ so that a post-filter can be constructed to estimate the speech component $\boldsymbol{s}(k,l)$.

B. MaxSNR Beamformer

The model in (8) is used to derive the MaxSNR beamformer, which aims to maximize the speech-to-noise ratio at the beamformer output [6], [17]. This is expressed as

$$\begin{aligned} \boldsymbol{w}_{\mathrm{MaxSNR}}(k,l) &= \arg\max_{\boldsymbol{w}(k,l)} \frac{\boldsymbol{w}^{\mathrm{H}}(k,l)\boldsymbol{\Phi}_{\mathrm{ss}}(k,l)\boldsymbol{w}(k,l)}{\boldsymbol{w}^{\mathrm{H}}(k,l)\boldsymbol{\Phi}_{\mathrm{nn}}(k,l)\boldsymbol{w}(k,l)} \\ &= \arg\max_{\boldsymbol{w}(k,l)} \frac{\boldsymbol{w}^{\mathrm{H}}(k,l)\boldsymbol{\Phi}_{\mathrm{xx}}(k,l)\boldsymbol{w}(k,l)}{\boldsymbol{w}^{\mathrm{H}}(k,l)\boldsymbol{\Phi}_{\mathrm{nn}}(k,l)\boldsymbol{w}(k,l)}. \end{aligned} \tag{9}$$

When the acoustic transfer function $\boldsymbol{h}(k)$ and the noise coherence matrix $\boldsymbol{\Gamma}_{\mathrm{nn}}(k)$ are both time-independent (as defined in the signal model), the MaxSNR beamformer is also time-independent. After derivation, the time-independent MaxSNR beamformer can be calculated as

$$\boldsymbol{w}_{\mathrm{MaxSNR}}(k) = \arg\max_{\boldsymbol{w}(k)} \frac{\boldsymbol{w}^{\mathrm{H}}(k)\bar{\boldsymbol{\Phi}}_{\mathrm{xx}}(k)\boldsymbol{w}(k)}{\boldsymbol{w}^{\mathrm{H}}(k)\boldsymbol{\Gamma}_{\mathrm{nn}}(k)\boldsymbol{w}(k)}, \tag{10}$$

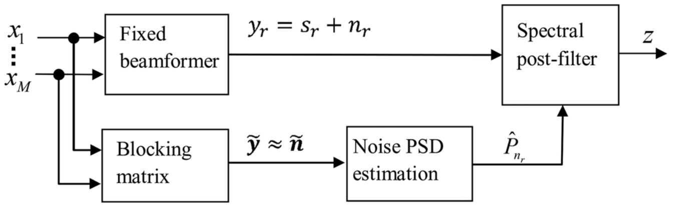


Fig. 1. A GSC-like speech enhancement system using blocking matrix as a noise PSD estimator.

where

$$\bar{\boldsymbol{\Phi}}_{\mathrm{xx}}(k) = \frac{1}{L}\sum_{l \in \mathbb{L}} \boldsymbol{x}(k,l)\boldsymbol{x}^{\mathrm{H}}(k,l), \tag{11}$$

where $\mathbb{L}$ denotes the set of all L time frames in the whole recording period.

For a single target speaker (i.e., $\boldsymbol{\Phi}_{\mathrm{ss}}(k,l)$ is a rank-1 matrix), the MaxSNR beamformer can be calculated as the (principle) generalized eigenvector corresponding to the largest generalized eigenvalue obtained from the generalized eigenvalue decomposition (GEVD) of the matrix pair $(\bar{\boldsymbol{\Phi}}_{\mathrm{xx}}(k), \boldsymbol{\Gamma}_{\mathrm{nn}}(k))$. This is expressed as

$$\bar{\boldsymbol{\Phi}}_{\mathrm{xx}}(k)\boldsymbol{w}(k) = \lambda\boldsymbol{\Gamma}_{\mathrm{nn}}(k)\boldsymbol{w}(k), \tag{12}$$

and

$$\boxed{\boldsymbol{w}_{\mathrm{MaxSNR}}(k) = \eta\boldsymbol{w}_{\max\_\mathrm{eig}}(k)} \tag{13}$$

where η is a complex-valued scalar and $\boldsymbol{w}_{\max\_\mathrm{eig}}$ is the generalized eigenvector corresponding to the largest generalized eigenvalue.

In later sections it will be shown that the concept of the MaxSNR beamformer is useful for deriving the proposed MaxNSR blocking matrix. For conciseness we will omit the indices k and l in the remainder of the paper, except where explicitly required.

IV. Noise PSD Estimation Based on Blocking Matrix Output

The block diagram of the considered speech enhancement system using a blocking matrix as a noise estimator is shown in Fig. 1. This GSC-based structure consists of five blocks: acoustic transfer function estimation (not shown in Fig. 1), fixed beamformer, blocking matrix, noise PSD estimation, and spectral post-filter. First, the acoustic transfer functions between the target speaker and the microphones are estimated to design the fixed beamformer and the blocking matrix. The aim of the fixed beamformer is to spatially focus on the target speaker and thereby reduce background noise not coming from the direction of the target speaker. The output signal of the fixed beamformer, $y_r = s_r + n_r$, also referred to as speech reference, consists of enhanced speech and residual noise, as the noise reduction performance of the fixed beamformer is typically quite limited. The aim of the blocking matrix is to cancel the target speech, yielding one or more noise references, $\tilde{\boldsymbol{y}} = \tilde{\mathbf{s}} + \tilde{\mathbf{n}}$, containing little speech. When no speech leakage occurs, i.e., $\tilde{\boldsymbol{y}} = \tilde{\mathbf{n}}$, the noise reference provides a good reference of the background noise even during speech-presence periods. The aim of the noise PSD estimation block is to estimate P_{n_r}, i.e., the PSD of the residual noise in the speech reference, from the noise reference. Finally, a spectral post-filter is applied to the speech reference y_r, yielding the output signal z.

Compared with the standard GSC system, e.g., [2], [5], [7], [17], the considered system in this paper uses noise PSD estimation and post-filter blocks instead of an adaptive noise canceller, which aims to reduce the residual noise in the speech reference that is correlated with the noise reference. There are two reasons for using the considered system. First, in diffuse noise fields an adaptive noise canceller has limited performance since the correlation between the residual noise in the speech reference and the noise reference is typically small. Second, assuming the noise coherence matrix to be known, an optimal fixed beamformer can be designed that is able to cancel the correlated noise components. The framework of using a blocking matrix as a noise PSD estimator has been investigated intensively in recent years [35], [36], [42]–[44]. The implementation details of the different blocks will be given in the following sections, where we will present several possible ways to combine the noise reference for target noise PSD estimation (cf. Section IV-D).

A. Acoustic Transfer Function Estimation

Since in practice it is very difficult to blindly estimate the acoustic transfer function $\boldsymbol{h}$ between the target speaker and the microphones, a simplified model, e.g., a pure-delay model or an RTF-based model, is generally used instead.

Assuming a free-field scenario and identical omni-directional microphones, a pure-delay model can be used to represent the acoustic path based on the geometrical information of the target speaker and the microphones. The pure-delay model is expressed as

$$\boldsymbol{a}_{\mathrm{delay}}(k) = [1, \mathrm{e}^{-j2\pi f_k \tau_{21}}, \cdots, \mathrm{e}^{-j2\pi f_k \tau_{M1}}]^{\mathrm{T}}, \tag{14}$$

where j represents the imaginary unit with $j^2 = -1$ and τ_{m1}, $m = 2, \cdots, M$, is the relative delay between the m-th microphone and the first microphone. Since the pure-delay model only considers the direct path, degraded modeling performance will be expected for reverberant scenarios.

In reverberant scenarios an RTF model can be used, where the RTF with respect to the first microphone is defined as

$$\boldsymbol{a}_{\mathrm{RTF}} = \frac{\boldsymbol{h}}{h_1} = [1, a_2, \cdots, a_M]^{\mathrm{T}}. \tag{15}$$

Generally, the RTFs can be estimated by exploiting the nonstationarity of speech signals [7] or based on the MaxSNR beamformer [17]. As pointed out in [17], the relationship between the MaxSNR beamformer $\boldsymbol{w}_{\mathrm{MaxSNR}}$ and the acoustic transfer function $\boldsymbol{h}$ can be expressed as

$$\boldsymbol{h} = \zeta\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}_{\mathrm{MaxSNR}}, \tag{16}$$

where ζ is an unknown frequency-dependent complex-valued scalar. Since the first element of $\boldsymbol{a}_{\mathrm{RTF}}$ is 1, the influence of the unknown scalar ζ can be canceled by using

$$\boldsymbol{a}_{\mathrm{RTF}} = \frac{\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}_{\mathrm{MaxSNR}}}{(\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}_{\mathrm{MaxSNR}})_1}, \tag{17}$$

where the subscript $(\cdot)_1$ denotes the first element of a vector.

B. Fixed Beamformer

Given the (acoustic or relative) transfer function $\boldsymbol{a}$ of the target speech and the noise coherence matrix $\boldsymbol{\Gamma}_{\mathrm{nn}}$, a minimum variance distortionless response (MVDR) beamformer can be constructed as [3]

$$\boldsymbol{w}_{\mathrm{BF}} = \frac{(\boldsymbol{\Gamma}_{\mathrm{nn}} + \delta \boldsymbol{I})^{-1}\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{H}}(\boldsymbol{\Gamma}_{\mathrm{nn}} + \delta \boldsymbol{I})^{-1}\boldsymbol{a}}, \tag{18}$$

where δ is the regularization constant and $\boldsymbol{I}$ is the $M \times M$-dimensional identity matrix. Depending on the choice of $\boldsymbol{a}$ and $\boldsymbol{\Gamma}_{\mathrm{nn}}$, different types of beamformers can be obtained. In this paper we will use the MVDR beamformer with the transfer function $\boldsymbol{a} = \boldsymbol{a}_{\mathrm{RTF}}$ estimated using (17) and the noise coherence matrix $\boldsymbol{\Gamma}_{\mathrm{nn}}$ estimated using (5).

In the simplest case, a delay-and-sum beamformer can be obtained by assuming a free-field model $\boldsymbol{a} = \boldsymbol{a}_{\mathrm{delay}}$ and spatially uncorrelated noise with $\boldsymbol{\Gamma}_{\mathrm{nn}} = \boldsymbol{I}$, i.e.,

$$\boldsymbol{w}_{\mathrm{BF}} = \frac{\boldsymbol{a}_{\mathrm{delay}}}{M}. \tag{19}$$

The speech reference y_r is obtained by applying the fixed beamformer to the microphone signals, i.e.,

$$y_r = \boldsymbol{w}_{\mathrm{BF}}^{\mathrm{H}}\boldsymbol{x} = s_r + n_r, \tag{20}$$

where s_r and n_r denote the speech component and the residual noise in the speech reference, respectively.

C. Blocking Matrix

The aim of the blocking matrix is to provide a reference of the background noise by canceling the target speech in the microphone signals. Similarly to the fixed beamformer, the blocking matrix depends on the transfer function $\boldsymbol{a}$ [7], [12], [13], [17]. A typical way to construct the blocking matrix is to use the orthogonal complement matrix of the vector $\boldsymbol{a}$, i.e.,

$$\boxed{\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{a}(\boldsymbol{a}^{\mathrm{H}}\boldsymbol{a})^{-1}\boldsymbol{a}^{\mathrm{H}}} \tag{21}$$

We refer to (21) as the null space blocking matrix. Again, depending on the choice of $\boldsymbol{a}$, different types of blocking matrices can be obtained. In the simplest case, a delay-and-subtract null beamformer is obtained by assuming a free-field model $\boldsymbol{a} = \boldsymbol{a}_{\mathrm{delay}}$. In this paper we will use the null space blocking matrix with $\boldsymbol{a} = \boldsymbol{a}_{\mathrm{RTF}}$.

The blocking matrix in (21) is an $M \times M$-dimensional matrix (with rank $M-1$), i.e., $\boldsymbol{B} = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_M]$, where each column vector $\boldsymbol{b}_m$ satisfies $\boldsymbol{b}_m^{\mathrm{H}}\boldsymbol{a} = 0$. With this blocking matrix, M noise references are generated as

$$\tilde{\boldsymbol{y}} = \boldsymbol{B}^{\mathrm{H}}\boldsymbol{x} = [\tilde{y}_1, \cdots, \tilde{y}_M]^{\mathrm{T}}, \tag{22}$$

where $\tilde{y}_m = \tilde{s}_m + \tilde{n}_m$ is the m-th noise reference, consisting of a noise component $\tilde{n}_m$ and speech leakage $\tilde{s}_m$.

If we assume that the target speech is completely cancelled by the blocking matrix, i.e., no speech leakage occurs (which will typically not be the case), the noise reference only contains noise, i.e.,

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{n}} = [\tilde{n}_1, \cdots, \tilde{n}_M]^{\mathrm{T}}. \tag{23}$$

D. Noise PSD Estimation

The outputs of the blocking matrix can now be used to estimate the PSD of the residual noise in the speech reference. In particular, P_{n_r}, the PSD of the residual noise in the speech reference, and $P_{\tilde{n}_m}$, the PSD of the m-th noise reference, can be related with a frequency-dependent but time-invariant scaling factor r_m [36], [42], [43]. This scaling factor can be calculated directly from the coefficients of th e two spatial filters (the fixed beamformer $\boldsymbol{w}_{\mathrm{BF}}$ and the blocking matrix vectors $\boldsymbol{b}_m$) and the noise coherence matrix $\boldsymbol{\Gamma}_{\mathrm{nn}}$ as

$$\boxed{r_m = \frac{P_{n_r}}{P_{\tilde{n}_m}} = \frac{\boldsymbol{w}_{\mathrm{BF}}^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}_{\mathrm{BF}}}{\boldsymbol{b}_m^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{b}_m}} \tag{24}$$

Since the null space blocking matrix in (21) has M outputs, several possibilities now exist to estimate P_{n_r} using

$$\boxed{\hat{P}_{n_r} = \sum_{m=1}^{I} \alpha_m P_{\tilde{y}_m}} \tag{25}$$

where I is the number of noise references been used and α_m is the combination coefficient defined below.

a) Using one of the M outputs to estimate the target noise PSD as shown in Fig. 2(a), i.e.,

$$I = 1, \alpha_m = r_m. \tag{26}$$

b) Comb1: Averaging the scaled PSDs of all noise references as shown in Fig. 2(b), i.e.,

$$I = M, \alpha_m = \frac{r_m}{M}. \tag{27}$$

c) Comb2: Scaling the average PSD of the noise references as shown in Fig. 2(c), i.e.,

$$I = M, \alpha_m = \frac{\boldsymbol{w}_{\mathrm{BF}}^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}_{\mathrm{BF}}}{\sum\limits_{m=1}^{M}\boldsymbol{b}_m^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{b}_m} = \frac{1}{\sum\limits_{m=1}^{M}\frac{1}{r_m}}. \tag{28}$$

d) Comb3: Scaling a linear combination of M noise references as shown in Fig. 2(d). This possibility will be discussed in more detail in Section V-B.

The combination (d) is a linear combination of noise references (signals), whereas the combinations (b) and (c) are linear combinations of noise PSDs. When no speech leakage occurs, all four combinations lead to the correct noise PSD estimate. In case of speech leakage, these different combinations lead to different noise PSD estimates, hence different performance (cf. simulations in Section VI-C).

E. Spectral Post-filter

Given the noise PSD estimate $\hat{P}_{n_r}$, a spectral post-filter is applied to enhance the speech reference y_r. For instance, in the well-known Wiener filter, the speech component s_r is estimated as

$$z(k,l) = G(k,l)y_r(k,l), \tag{29}$$

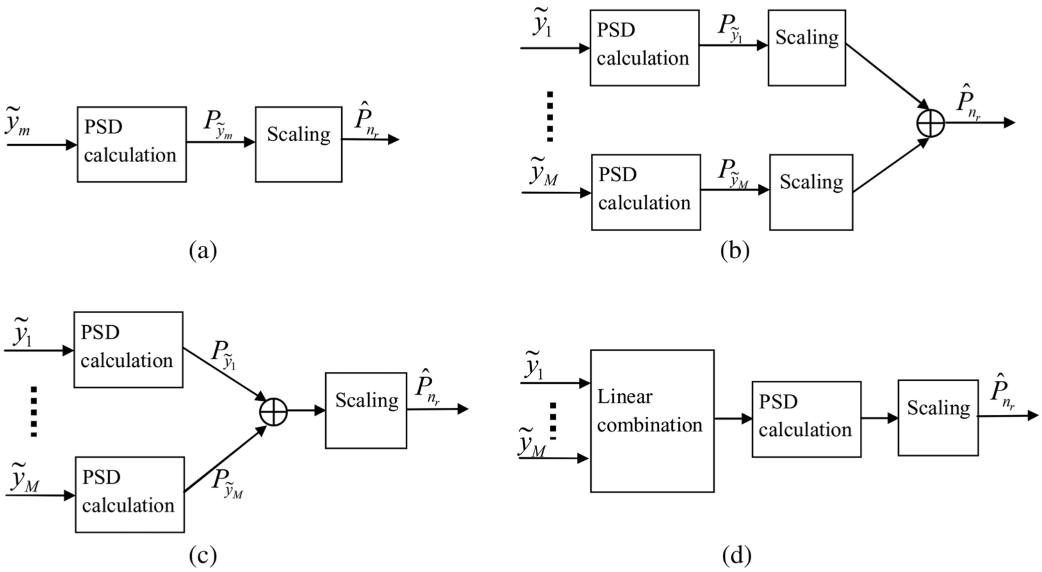


Fig. 2. Several possibilities of using the blocking matrix outputs for noise PSD estimation (a) using one noise reference (b) using M noise references (Comb1) (c) using M noise references (Comb2) (c) using M noise references (Comb3).

where the spectral gain is computed as

$$G(k,l) = \max\left(\frac{\xi_{\mathrm{dd}}(k,l)}{1+\xi_{\mathrm{dd}}(k,l)}, G_{\min}\right), \tag{30}$$

where $G_{\min}$ is the minimum gain to reduce distortions and ξ_{dd} is the estimated *a priori* SNR, e.g., obtained using the decision-directed approach [45] as

$$\xi_{\mathrm{dd}}(k,l) = \alpha_{\mathrm{dd}}\frac{|z(k,l-1)|^2}{\hat{P}_{n_r}(k,l-1)} + (1-\alpha_{\mathrm{dd}})(\frac{|y_r(k,l)|^2}{\hat{P}_{n_r}(k,l)} - 1), \tag{31}$$

where α_{dd} is a smoothing factor.

V. MaxNSR Blocking Matrix

A. Proposed Algorithm

The different noise PSD estimation possibilities outlined in Section IV-D all work perfectly under the assumption that the noise reference does not contain residual speech, i.e., $\tilde{y}_m = \tilde{n}_m$. However, in practice it is typically not possible to completely cancel the target speech, e.g., due to reverberation. In this case a speech leakage component will be present in the noise reference, i.e., $\tilde{y}_m = \tilde{s}_m + \tilde{n}_m$. Using (26) as an example, the noise PSD estimate is then equal to

$$\begin{aligned}\hat{P}_{n_r} &= r_m P_{\tilde{y}_m} = r_m P_{\tilde{s}_m} + r_m P_{\tilde{n}_m}\\ &= P_{n_r}(1+\frac{P_{\tilde{s}_m}}{P_{\tilde{n}_m}}).\end{aligned} \tag{32}$$

This implies that the noise PSD $\hat{P}_{n_r}$ is always overestimated and depends on the ratio of the PSD of the speech leakage $P_{\tilde{s}_m}$ and the PSD of the noise $P_{\tilde{n}_m}$. To minimize the influence of the speech leakage on the noise PSD estimation, both issues should be considered simultaneously, i.e., minimizing $P_{\tilde{s}_m}$ while maximizing $P_{\tilde{n}_m}$. However, existing null space blocking matrix algorithms only address the first issue by suppressing the target speech while not controlling the energy of the remaining noise.

We hence propose a novel blocking matrix, which aims to suppress the target speech while maximizing the noise energy in the output of the blocking matrix. Inspired by the MaxSNR beamformer, we introduce a maximum noise-to-speech ratio (MaxNSR) criterion for designing this blocking matrix. Assuming only one noise reference, i.e.,

$$\tilde{y} = \boldsymbol{w}^{\mathrm{H}}\boldsymbol{x} = \tilde{s} + \tilde{n}, \tag{33}$$

where $\boldsymbol{w}$ denotes the $M\times 1$-dimensional blocking matrix vector, the cost function to be maximized is equal to

$$\tilde{J} = \frac{P_{\tilde{n}}}{P_{\tilde{s}}}. \tag{34}$$

Maximizing $\tilde{J}$ corresponds to minimizing

$$J = 1 + \frac{1}{\tilde{J}} = \frac{P_{\tilde{s}} + P_{\tilde{n}}}{P_{\tilde{n}}} = \frac{P_{\tilde{y}}}{P_{\tilde{n}}}, \tag{35}$$

such that the MaxNSR blocking matrix can be computed as

$$\boldsymbol{w}_{\mathrm{maxNSR}} = \arg\min_{\boldsymbol{w}} \frac{\boldsymbol{w}^{\mathrm{H}}\bar{\boldsymbol{\Phi}}_{\mathrm{xx}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}}, \tag{36}$$

where $\boldsymbol{\Gamma}_{\mathrm{nn}}$ and $\bar{\boldsymbol{\Phi}}_{\mathrm{xx}}$ are defined in (5) and (11), respectively.

This optimization problem can be solved using the generalized eigenvalue decomposition

$$\bar{\boldsymbol{\Phi}}_{\mathrm{xx}}\boldsymbol{w} = \lambda\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}, \tag{37}$$

with λ the generalized eigenvalue. The solution minimizing J is the generalized eigenvector $\boldsymbol{w}_{\min\_\mathrm{eig}}$ corresponding to the smallest generalized eigenvalue, i.e.,

$$\boxed{\boldsymbol{w}_{\mathrm{maxNSR}} = \xi\boldsymbol{w}_{\min\_\mathrm{eig}}} \tag{38}$$

where ξ is a complex-valued scalar. Theoretically, ξ can be any value, but we choose it to be equal to

$$\xi = \sqrt{\frac{1}{\boldsymbol{w}^{\mathrm{H}}_{\min\_\mathrm{eig}}\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}_{\min\_\mathrm{eig}}}}, \tag{39}$$

so that the energy of the input and output noise are equal.

The output of the MaxNSR blocking matrix is expressed as

$$\tilde{y}_{\mathrm{MaxNSR}} = \boldsymbol{w}_{\mathrm{MaxNSR}}^{\mathrm{H}} \boldsymbol{x}. \tag{40}$$

Note that in comparison to the null space blocking matrix with M outputs, the MaxNSR blocking matrix just generate one output. The target noise PSD can be estimated from the MaxNSR blocking matrix output as

$$\boxed{\hat{P}_{n_r} = r_{\mathrm{nsr}} P_{\tilde{y}_{\mathrm{MaxNSR}}}} \tag{41}$$

and

$$\boxed{r_{\mathrm{nsr}} = \frac{\boldsymbol{w}_{\mathrm{BF}}^{\mathrm{H}} \boldsymbol{\Gamma}_{\mathrm{nn}} \boldsymbol{w}_{\mathrm{BF}}}{\boldsymbol{w}_{\mathrm{MaxNSR}}^{\mathrm{H}} \boldsymbol{\Gamma}_{\mathrm{nn}} \boldsymbol{w}_{\mathrm{MaxNSR}}}} \tag{42}$$

Finally, it is worthwhile noting that both the fixed beamformer and the blocking matrix can be obtained simultaneously from the generalized eigenvalue decomposition of $(\bar{\boldsymbol{\Phi}}_{\mathrm{xx}}, \boldsymbol{\Gamma}_{\mathrm{nn}})$: the fixed beamformer relates to the largest generalized eigenvalue (cf. (13)) while the blocking matrix relates to the smallest generalized eigenvalue (cf. (38)).

B. Relationship Between Null Space and MaxNSR Blocking Matrices

From (21) it can be seen that the null space blocking matrix $\boldsymbol{B}$ has rank $M-1$ and the M column vectors $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_M\}$ span the orthogonal complement subspace $\boldsymbol{a}^{\perp}$ of the vector $\boldsymbol{a}$, i.e., any vector that is orthogonal to $\boldsymbol{a}$ can be represented as a linear combination of $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_M\}$.

In Appendix A it is shown that the generalized eigenvector corresponding to the minimum generalized eigenvalue of (37) satisfies

$$\boldsymbol{w}_{\mathrm{min\_eig}}^{\mathrm{H}} \boldsymbol{a} = 0. \tag{43}$$

Since the MaxNSR vector $\boldsymbol{w}_{\mathrm{MaxNSR}}$ is a scaled version of $\boldsymbol{w}_{\mathrm{min\_eig}}$, it is also orthogonal to $\boldsymbol{a}$ and can be represented as a linear combination of the vectors $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_M\}$, i.e.,

$$\boldsymbol{w}_{\mathrm{MaxNSR}} = \boldsymbol{B}\boldsymbol{c}, \tag{44}$$

with $\boldsymbol{c} = [c_1, \cdots, c_M]^{\mathrm{T}}$ the linear combination vector.

Since $\boldsymbol{w}_{\mathrm{MaxNSR}}$ minimizes (36), it follows that the optimal linear combination vector minimizes

$$J = \frac{\boldsymbol{c}^{\mathrm{H}}(\boldsymbol{B}^{\mathrm{H}} \bar{\boldsymbol{\Phi}}_{\mathrm{xx}} \boldsymbol{B})\boldsymbol{c}}{\boldsymbol{c}^{\mathrm{H}}(\boldsymbol{B}^{\mathrm{H}} \boldsymbol{\Gamma}_{\mathrm{nn}} \boldsymbol{B})\boldsymbol{c}}, \tag{45}$$

and hence is the generalized eigenvector corresponding to the smallest generalized eigenvalue of the GEVD

$$(\boldsymbol{B}^{\mathrm{H}} \bar{\boldsymbol{\Phi}}_{\mathrm{xx}} \boldsymbol{B})\boldsymbol{c} = \lambda(\boldsymbol{B}^{\mathrm{H}} \boldsymbol{\Gamma}_{\mathrm{nn}} \boldsymbol{B})\boldsymbol{c}. \tag{46}$$

In summary, with (44)–(46), the MaxNSR vector $\boldsymbol{w}_{\mathrm{MaxNSR}}$ can be represented as a linear combination of the null space vectors $\boldsymbol{b}_m$ in the sense of maximum noise-to-speech ratio. The MaxNSR vector belongs to the null space of the acoustic transfer function vector $\boldsymbol{a}$ and hence has the same target speech blocking ability as the traditional null space vectors. The noise PSD estimation scheme based on a MaxNSR blocking matrix is equivalent to the scheme shown in Fig. 2(d), where the M outputs of a null space blocking matrix are first linearly combined with the weighting vector $\boldsymbol{c}$, before being used to estimate the target noise PSD. Furthermore, as an optimal combination of these null space vectors, the MaxNSR vector maximizes the noisy energy in the output and hence minimizes the influence of speech leakage on the noise PSD estimation, cf. (32). In case of $M = 2$ microphones, the MaxNSR vector is equivalent to a null space vector. However, the advantage of the MaxNSR blocking matrix will become evident for $M > 2$ (cf. experiments in Section VI-C).

C. Complete Algorithm

A complete description of the MaxNSR blocking matrix based noise PSD estimation algorithm is given in Algorithm 1.

Algorithm 1 MaxNSR blocking matrix based noise PSD estimation

```
1) Parameter initialization: $\delta, G_{\mathrm{min}}, \alpha_{\mathrm{dd}}, \alpha_p$
2) VAD
3) Correlation matrix estimation:
       Estimate $\bar{\boldsymbol{\Phi}}_{\mathrm{xx}}$                                  (11)
       Estimate $\bar{\boldsymbol{\Phi}}_{\mathrm{nn}}$                                  (6)
       Estimate $\boldsymbol{\Gamma}_{\mathrm{nn}}$                                  (5)
4) GEVD: $\bar{\boldsymbol{\Phi}}_{\mathrm{xx}}\boldsymbol{w} = \boldsymbol{\lambda}\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}$                         (37)
5) RTF estimation:
     $\boldsymbol{w}_{\mathrm{MaxSNR}} = \eta \boldsymbol{w}_{\mathrm{max\_eig}}$                        (13)
     $\boldsymbol{a}_{\mathrm{RTF}} = \frac{\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}_{\mathrm{MaxSNR}}}{(\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}_{\mathrm{MaxSNR}})_1}$                        (17)
6) Fixed beamformer:
     $\boldsymbol{a} = \boldsymbol{a}_{\mathrm{RTF}}, \boldsymbol{w}_{\mathrm{BF}} = \frac{(\boldsymbol{\Gamma}_{\mathrm{nn}}+\delta\boldsymbol{I})^{-1}\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{H}}(\boldsymbol{\Gamma}_{\mathrm{nn}}+\delta\boldsymbol{I})^{-1}\boldsymbol{a}}$              (18)
     $y_r = \boldsymbol{w}_{\mathrm{BF}}^{\mathrm{H}}\boldsymbol{x}$                                  (20)
7) Blocking matrix:
     $\boldsymbol{w}_{\mathrm{MaxNSR}} = \xi \boldsymbol{w}_{\mathrm{min\_eig}}$                     (38)–(39)
     $\tilde{y}_{\mathrm{MaxNSR}} = \boldsymbol{w}_{\mathrm{MaxNSR}}^{\mathrm{H}}\boldsymbol{x}$                         (40)
8) Noise PSD estimation:
     $r_{\mathrm{nsr}} = \frac{\boldsymbol{w}_{\mathrm{BF}}^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}_{\mathrm{BF}}}{\boldsymbol{w}_{\mathrm{MaxNSR}}^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{nn}}\boldsymbol{w}_{\mathrm{MaxNSR}}}$                   (42)
     $\hat{P}_{n_r} = r_{\mathrm{nsr}} P_{\tilde{y}_{\mathrm{MaxNSR}}}$                          (41)
9) Postfilter:                                     (29)–(31)
```

Algorithm 2 Null space blocking matrix based noise PSD estimation

```
1)-5): same as Algorithm 1
6) Fixed beamformer:
     $\boldsymbol{a} = \begin{cases} \boldsymbol{a}_{\mathrm{RTF}}: & \text{MVDR} \\ \boldsymbol{a}_{\mathrm{delay}}: & \text{delay-and-sum} \end{cases}$
     $\boldsymbol{w}_{\mathrm{BF}} = \frac{(\boldsymbol{\Gamma}_{\mathrm{nn}}+\delta\boldsymbol{I})^{-1}\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{H}}(\boldsymbol{\Gamma}_{\mathrm{nn}}+\delta\boldsymbol{I})^{-1}\boldsymbol{a}}$                    (18)
     $y_r = \boldsymbol{w}_{\mathrm{BF}}^{\mathrm{H}}\boldsymbol{x}$                                  (20)
7) Blocking matrix:
     $\boldsymbol{a} = \begin{cases} \boldsymbol{a}_{\mathrm{RTF}}: & \text{Nullspace} \\ \boldsymbol{a}_{\mathrm{delay}}: & \text{delay-and-substract} \end{cases}$
     $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{a}(\boldsymbol{a}^{\mathrm{H}}\boldsymbol{a})^{-1}\boldsymbol{a}^{\mathrm{H}}$                       (21)
     $\tilde{\boldsymbol{y}} = \boldsymbol{B}^{\mathrm{H}}\boldsymbol{x}$ (22)
8) Noise PSD estimation:
     $\hat{P}_{n_r} = \sum_{m=1}^{I} \alpha_m P_{\tilde{y}_m}$                    (24)–(28)
9) Postfilter:                                     (29)–(31)
```

In Step 3) the noisy speech correlation matrix $\bar{\boldsymbol{\Phi}}_{\mathrm{xx}}$ is estimated using (11) from the microphone signals in the whole

TABLE I
SPECIFIC PARAMETERS FOR THE EXPERIMENT

Parameter	Value
STFT length L_1	1024 (50% overlap)
STFT length L_2	256 (50% overlap)
δ in (18)	10^{-3}
G_{min} in (30)	-14 dB
α_{dd} in (31)	0.98
α_p in (47)	0.8

recording period. The noise coherence matrix $\mathbf{\Gamma}_{\text{nn}}$ can either be computed based on prior knowledge of the noise field (e.g., spherically isotropic noise in (7)) or be estimated using (5)-(6) from the microphone signals in noise-only periods, based on the output of a voice activity detector (VAD). The first strategy does not require a VAD, while the second strategy is suitable when the spatial properties of the noise field are unknown, e.g., semi-incoherent noise. In rare cases where no noise is detected by the VAD and the denominator in (5) equals to zero, the noise coherence matrix $\mathbf{\Gamma}_{nn}$ is set to an identity matrix.

In Step 8) the PSD of the noise reference $P_{\tilde{y}_{\text{MaxNSR}}} = \text{E}(|\tilde{y}_{\text{MaxNSR}}|^2)$ is calculated by recursively smoothing over time, i.e.,

$$P_{\tilde{y}_{\text{MaxNSR}}}(k,l) = \alpha_p P_{\tilde{y}_{\text{MaxNSR}}}(k,l\text{–}1) + (1\text{–}\alpha_p)|\tilde{y}_{\text{MaxNSR}}(k,l)|^2, \tag{47}$$

with α_p the smoothing constant.

For reference, the complete description of the null space blocking matrix based noise PSD estimation algorithm is given in Algorithm 2. Depending on the choice for $\boldsymbol{a}$ the fixed beamformer can be implemented as an MVDR or a delay-and-sum beamformer, while the blocking matrix can be implemented as a null space or a delay-and-subtract blocking matrix. In Step 8), the PSD of the noise reference $P_{\tilde{y}_m}$ is calculated in the same way as (47).

For both algorithms the computationally most demanding part is the GEVD in each frequency bin. Algorithm 2 requires the GEVD to estimate the RTF. Algorithm 1 requires the GEVD to achieve three objectives simultaneously: RTF estimation, MVDR beamformer, and MaxNSR blocking matrix. Since both algorithms require the same GEVD operation, their computational complexity is comparable.

Typical values of the parameters involved in both algorithms will be given in Table I.

VI. Experimental Results

After introducing the experimental setup, which includes the acoustic scenario, the algorithmic implementation details, and objective performance measures, the performance of the proposed algorithm is examined from several aspects. In Section VI-B the noise reduction performance of the fixed beamformer is examined in diffuse noise fields. In Section VI-C, the target speech blocking ability of different blocking matrices is compared. In Section VI-D the noise estimation and speech enhancement performance of the noise PSD estimators using different blocking matrices are evaluated and compared with a single-channel noise PSD estimator, both for stationary as well as for nonstationary noise. Finally, in Section VI-E the performance of the considered algorithms is investigated with diffuse babble noise and with a non-ideal VAD.

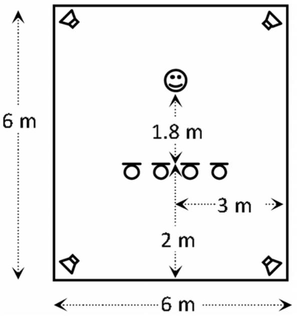


Fig. 3. Illustration of the recording room. The inter-microphone distance is 9 cm.

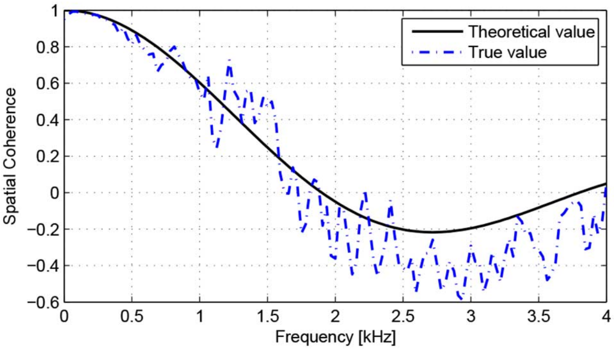


Fig. 4. Spatial coherence of the recorded diffuse noise for the microphone distance of 9 cm.

A. Experimental Setup

1) Acoustic scenario: The experiments have been carried out using recordings from a room with approximate dimensions 6 m × 6 m × 3 m with a reverberation time of approximately 400 ms. Four microphones with an inter-microphone distance of 9 cm are placed in the center of the room, as shown in Fig. 3. For the target speaker a loudspeaker placed around 1.8 m away in the broadside direction of the microphone array is used. All loudspeakers and microphones are 1.3 m high. The target speech component at the microphones is generated by convolving the clean speech signal with the measured impulse responses. The speech signal has a length of 32 s, and the used sampling rate is $f_s = 8$ kHz. In all experiments (except Section VI-E) we have used diffuse-like noise, which has been generated by playing independent stationary white noise from four loudspeakers turned towards the four corners of the room. The spatial coherence of the recorded diffuse-like noise at two microphones of distance 9 cm is plotted in Fig. 4. It can be observed that the measured coherence value (cf. (5)) is consistent with the theoretical value (cf. (7)). Temporally nonstationary noise has been generated by multiplying the recorded stationary diffuse noise with a modulation function in the time domain. The modulation function is defined as

$$f(t) = 1 + 0.8 \cdot \sin(2\pi t f_m / f_s), \tag{48}$$

with modulation depth 0.8 and modulation frequency f_m, which controls how fast the noise amplitude is varying with time [41]. When $f_m = 0$, the modulated noise becomes stationary. The recorded noise is added to the speech components at different input SNRs. Throughout the experiment, all four microphones will be used unless otherwise stated.

2) Algorithmic implementation details: On the one hand, for the fixed beamformer and the blocking matrix a large STFT analysis frame length (e.g., $L_1 = 1024$) is required to enable to capture most of the reverberation. On the other hand, considering the statistical properties of speech, which can only be approximated as a quasi-stationary process within a short time period (20-30 ms), a shorter STFT frame length (e.g., $L_2 = 256$) needs to be used for the single-channel spectral filter. Hence, in order to deal with the inconsistency between the different STFT lengths, a compensation scheme is required. For simplicity, we only consider a single channel, but the scheme can be easily extended to multiple channels. After spatial filtering (with STFT length L_1), we obtain the noise reference $\tilde{y}_m(k_1, l_1)$, the speech reference $y_r(k_1, l_1)$, and the scaling factor $r_m(k_1, l_1)$. After applying the scaling factor to the noise reference $\tilde{y}_m(k_1, l_1)$, the scaled signal

$$\tilde{y}_{m\_\mathrm{sca}}(k_1, l_1) = \sqrt{r_m(k_1, l_1)} \cdot |\tilde{y}_m(k_1, l_1)| \cdot \mathrm{e}^{j\angle y_r(k_1, l_1)} \quad (49)$$

is synthesized to the time domain as $\tilde{y}_{m\_\mathrm{sca}}(t)$. Note that in (49) we use the phase of the speech reference $y_r(k_1, l_1)$ so that after inverse STFT the scaled signal $\tilde{y}_{m\_\mathrm{sca}}(t)$ is closer to the target noise $n_r(t)$. The signal is then reanalyzed using the STFT with frame length L_2 as $\tilde{y}_{m\_\mathrm{sca}}(k_2, l_2)$, from which the target noise PSD is estimated as

$$\hat{P}_{n_r}(k_2, l_2) = P_{\tilde{y}_{m\_\mathrm{sca}}}(k_2, l_2), \quad (50)$$

where $P_{\tilde{y}_{m\_\mathrm{sca}}}(k_2, l_2)$ is calculated similarly to (47).

The noise coherence matrix $\mathbf{\Gamma}_{\mathrm{nn}}$ is estimated from the microphone signals in noise-only periods. A perfect VAD is assumed to be available in all experiments except Section VI-E.

Finally, all algorithmic parameters used in the experiments are listed in Table I.

3) Objective performance measures:

3.1) Fixed beamformer: The noise reduction performance of the fixed beamformer is evaluated using the global SNR improvement. For a time-domain microphone signal $x(t) = s(t) + n(t)$ the global SNR is defined as

$$\mathrm{SNR_{glo}} = 10 \log_{10} \frac{\sum_{t \in \mathbb{T}_s} |s(t)|^2}{\sum_{t \in \mathbb{T}_s} |n(t)|^2}, \quad (51)$$

where $\mathbb{T}_s$ denotes the speech presence periods. The global SNR improvement is calculated by comparing the global SNR before and after the fixed beamformer.

3.2) Blocking matrix: The target speech blocking ability of a blocking matrix is evaluated using two objective measures: global NSR ($\mathrm{NSR_{glo}}$) and scaled NSR ($\mathrm{NSR_{sca}}$).

The global NSR is defined in the time domain in a similar way as the global SNR in (51) by changing the numerator and the denominator. The global NSR improvement is calculated by comparing the global NSR before and after the blocking matrix. The larger the NSR improvement, the better the performance of the blocking matrix.

As indicated by (32), the influence of speech leakage on noise PSD estimation is mainly reflected by the energy ratio of the scaled noise to the scaled speech. Thus, for our noise PSD estimation framework in (25) the scaled NSR of the blocking matrix is defined in the whole time-frequency domain as

$$\mathrm{NSR_{sca}} = 10\log_{10} \frac{\sum\limits_{m=1}^{I} \sum\limits_{k=1}^{K} \sum\limits_{l=1}^{L} \alpha_m(k,l) P_{\tilde{n}_m}(k,l)}{\sum\limits_{m=1}^{I} \sum\limits_{k=1}^{K} \sum\limits_{l=1}^{L} \alpha_m(k,l) P_{\tilde{s}_m}(k,l)}, \quad (52)$$

where, as defined before, L and K are the number of frames and frequency bins respectively, I denotes the number of output channels involved in the noise PSD estimation and $\tilde{y}_m = \tilde{s}_m + \tilde{n}_m$ is the m-th output of the blocking matrix. The larger $\mathrm{NSR_{sca}}$, the better the blocking ability of the blocking matrix.

3.3) Noise PSD estimation: To evaluate the noise PSD estimation performance, two objective measures are employed which compare the target noise PSD $P_{n_r}(k, l)$ and the estimated PSD $\hat{P}_{n_r}(k, l)$ in terms of overestimation and underestimation error [41]. The total estimation error can be represented as

$$\mathrm{Err_{total}} = \mathrm{Err_{ov}} + \mathrm{Err_{un}}, \quad (53)$$

where $\mathrm{Err_{ov}}$ measures the contributions of an overestimation of the true noise PSD as

$$\mathrm{Err_{ov}} = \frac{10}{LK} \sum_{l=1}^{L} \sum_{k=1}^{K} \left| \min\left(0, \log_{10} \frac{P_{n_r}(k,l)}{\hat{P}_{n_r}(k,l)}\right) \right|, \quad (54)$$

while $\mathrm{Err_{un}}$ measures the contributions of an underestimation of the true noise PSD as

$$\mathrm{Err_{un}} = \frac{10}{LK} \sum_{l=1}^{L} \sum_{k=1}^{K} \left| \max\left(0, \log_{10} \frac{P_{n_r}(k,l)}{\hat{P}_{n_r}(k,l)}\right) \right|. \quad (55)$$

An overestimation of the true noise PSD is likely to result in attenuation of the speech signal in a speech enhancement algorithm and thus in speech distortion. On the other hand, an underestimation of the true noise PSD results in reduced noise suppression and is likely to yield an increase of musical noise. To clearly show the influence of the speech leakage on noise PSD estimation, these two objective measures are calculated separately in noise-only periods and speech-presence periods.

3.4) Speech enhancement: For the post-filter with as input the speech reference $y_r(t) = s_r(t) + n_r(t)$ and as output $z(t) = \hat{s}(t) + \hat{n}(t)$ in the time domain, four objective measures are used to evaluate the noise reduction and speech enhancement performance in noise-only periods and speech-presence periods separately.

In speech-presence periods the speech enhancement performance is evaluated using the segmental noise reduction $\mathrm{NRS_{seg}}$, the segmental speech distortion $\mathrm{SD_{seg}}$, and the segmental SNR $\mathrm{SNR_{seg}}$, which are defined over non-overlapping time-domain frames with a length of 15 ms, i.e.,

$$\mathrm{NRS_{seg}} = \frac{1}{Q_s} \sum_{q \in \mathbb{Q}_s} 10\log_{10} \left(\frac{\sum\limits_{t \in \mathrm{frame}_q} |n_r(t)|^2}{\sum\limits_{t \in \mathrm{frame}_q} |\hat{n}(t)|^2} \right), \quad (56)$$

TABLE II
OBJECTIVE EVALUATION MEASURES

Spatial filter	Noise estimation	Speech enhancement
SNR_{glo}	$\text{Err}_{\text{total}}$	NRN_{seg}
NSR_{glo}	Err_{ov}	NRS_{seg}
NSR_{sca}	Err_{un}	SD_{seg}
		SNR_{seg}
		PESQ

$$\text{SD}_{\text{seg}} = \frac{1}{Q_s} \sum_{q \in \mathbb{Q}_s} 10\log_{10} \left(\frac{\sum\limits_{t \in \text{frame}_q} |s_r(t)|^2}{\sum\limits_{t \in \text{frame}_q} |s_r(t) - \hat{s}(t)|^2} \right), \tag{57}$$

$$\text{SNR}_{\text{seg}} = \frac{1}{Q_s} \sum_{q \in \mathbb{Q}_s} 10\log_{10} \left(\frac{\sum\limits_{t \in \text{frame}_q} |s_r(t)|^2}{\sum\limits_{t \in \text{frame}_q} |s_r(t) - z(t)|^2} \right), \tag{58}$$

where $\mathbb{Q}_s$ denotes the set of Q_s frames where speech is present. It should be noted that the measures SD_{seg} and SNR_{seg} may punish the possible dereverberation effect of the post-filter, since the used reference s_r is a reverberant version of the clean speech signal.

In noise-only periods the noise reduction performance is measured by

$$\text{NRN}_{\text{seg}} = \frac{1}{Q_n} \sum_{q \in \mathbb{Q}_n} 10\log_{10} \left(\frac{\sum\limits_{t \in \text{frame}_q} |n_r(t)|^2}{\sum\limits_{t \in \text{frame}_q} |\hat{n}(t)|^2} \right), \tag{59}$$

where $\mathbb{Q}_n$ denotes the set of Q_n noise-only frames.

In addition, PESQ (Perceptual Evaluation of Speech Quality) is used as a global measure to assess the overall speech quality after speech enhancement processing [46].

A short summary of all evaluation measures is given in Table II.

B. Noise Reduction Performance of the Fixed Beamformer

In this experiment the noise reduction performance of two fixed beamformers (delay-and-sum (19) and MVDR (18)) is compared for stationary diffuse noise at a global input SNR of 5 dB. For the delay-and-sum beamformer, $\boldsymbol{a}_{\text{delay}}$ in (14) is calculated assuming the positions of the microphones and the speaker are known. The global SNR improvement achieved by the fixed beamformer for different FFT sizes is shown in Fig. 5. The following observations can be made:

1) The SNR improvement of the delay-and-sum beamformer remains almost constant with respect to the FFT size. It only achieves an SNR improvement of about 1 dB because the used (free-field) transfer function model and (identity) noise coherence matrix are both inaccurate.
2) As expected, the MVDR beamformer performs better than the delay-and-sum beamformer by using a better model for the acoustic transfer function and the noise coherence matrix. Furthermore, its performance improves when increasing the FFT size, because the RTF model becomes more accurate in reverberant environments.

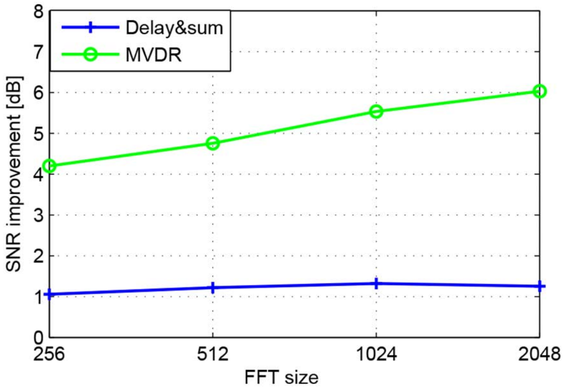


Fig. 5. Global SNR improvement for the two fixed beamformers (using 4 microphones) in stationary diffuse noise at a global input SNR of 5 dB.

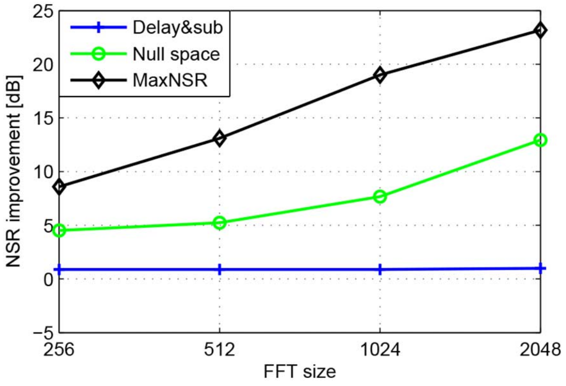


Fig. 6. Global NSR improvement for the three blocking matrices (using 4 microphones) in stationary diffuse noise at a global input SNR of 5 dB.

3) Nonetheless, the SNR improvement of both fixed beamformers is still limited in diffuse noise fields: only a few dB of SNR improvement can be achieved. Hence a post-filter is required to further enhance the target speech.

C. Blocking Ability of the Blocking Matrix

In this experiment the target speech blocking ability of different blocking matrices is compared for stationary diffuse noise at a global input SNR of 5 dB. The MVDR beamformer is used to generate the speech reference, which is used to compute the NSR_{sca} measure in (52).

The global NSR improvement of the delay-and-subtract blocking matrix (first output), the null space blocking (first output), and the MaxNSR blocking matrix is shown in Fig. 6. The following observations can be made:

1) The NSR improvement of the delay-and-subtract beamformer remains almost constant with respect to the FFT size.
2) The MaxNSR and the null space blocking matrix can achieve a larger NSR improvement by using RTFs to model the acoustic transfer function. The MaxNSR blocking matrix clearly yields the best performance.
3) Similarly to the fixed beamformer, the performance of the MaxNSR and the null space blocking matrices increases with FFT size.

For the scaled NSR measure in (52) the following possible blocking matrix outputs for noise PSD estimation are considered:

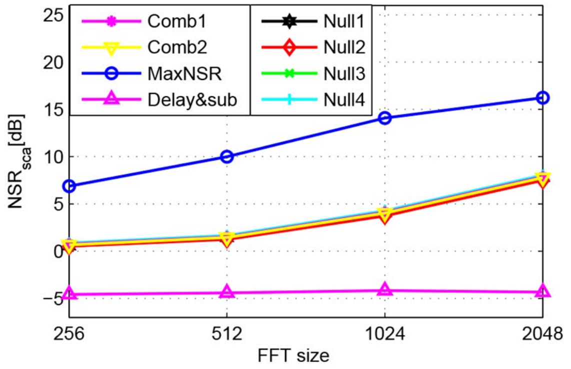


Fig. 7. Blocking ability ($\mathrm{NSR}_{\mathrm{sca}}$) of the null space blocking matrix and the MaxNSR blocking matrix (using 4 microphones) in stationary diffuse noise at a global input SNR of 5 dB. The curves of Comb1, Comb2, and Null1-4 are practically overlapping.

- **Null1-Null4**: The PSD each output of the null space blocking matrix is employed separately for noise PSD estimation using (26), cf. Fig. 2(a).
- **Comb1**: The PSDs of the four outputs of the null space blocking matrix are combined using (27), cf. Fig. 2(b).
- **Comb2**: The PSDs of the four outputs of the null space blocking matrix are combined using (28), cf. Fig. 2(c).
- **MaxNSR**: The MaxNSR blocking matrix which corresponds to a linear combination of the four outputs of the null space blocking matrix in terms of maximum noise-to-speech ratio, cf. Fig. 2(d).
- **Delay&sub**: The first output of the delay-and-subtract blocking matrix is employed for noise PSD estimation using (26), cf. Fig. 2(a).

Fig. 7 depicts the scaled NSR for different FFT sizes for all discussed noise PSD estimation procedures. The following observations can be made:

1) Similarly to the fixed beamformer, the performance of the blocking matrix also increases with FFT size for all noise PSD estimation procedures, except for the delay-and-subtract blocking matrix.
2) All outputs of the null space blocking matrix perform very similarly. In addition, both Comb1 and Comb2, which combine the PSDs of the 4 outputs, perform very similarly to using one output.
3) The MaxNSR blocking matrix clearly outperforms the other possibilities to the outputs.

Furthermore, Fig. 8 compares the scaled NSR of the blocking matrix when using two and four microphones, respectively. As theoretically expected, the null space and MaxNSR blocking matrices perform similarly for two microphones since both null space vectors are equivalent to the MaxNSR vector. When the rank of the null space increases to 3 for four microphones, the outputs of the null space blocking matrix can be combined to obtain a better result, as can be observed. Hence, the proposed MaxNSR blocking matrix is advantageous especially when more than two microphones are used.

D. Noise Estimation and Speech Enhancement

In this experiment the noise PSD estimation performance of different blocking matrix based noise PSD estimation procedures and a single-channel noise PSD estimation procedure is evaluated for stationary and non-stationary diffuse noise. The diffuse noise is added to the microphones at a global input SNR ranging from −5 dB to 15 dB. The following procedures are compared:

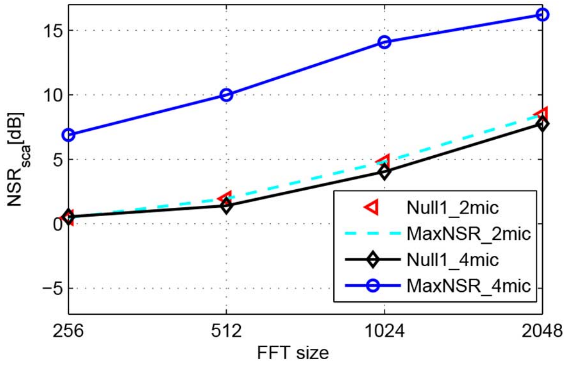


Fig. 8. Blocking ability ($\mathrm{NSR}_{\mathrm{sca}}$) of the null space blocking matrix and the MaxNSR blocking matrix, using 2 and 4 microphones in stationary diffuse noise at a global input SNR of 5 dB. The curves of Null1_2mic and MaxNSR_2mic overlap.

- **Delay&sub**: Multichannel noise PSD estimation using the delay-and-subtract blocking matrix.
- **Null space**: Multichannel noise PSD estimation using the null space blocking matrix (using the first output).
- **MaxNSR**: Multichannel noise PSD estimation using the MaxNSR blocking matrix.
- **UMMSE**: Single-channel unbiased MMSE noise PSD estimation [41], which estimates the noise PSD $\hat{P}_{n_r}$ directly from the speech reference y_r.

D1. Stationary noise: Fig. 9 compares the noise PSD estimation performance (in terms of overestimation error, underestimation error and total error) in stationary diffuse noise for the four considered noise PSD estimators in noise-only and speech-presence periods separately. The following observations can be made:

1) In noise-only periods the noise PSD estimation performance of the four noise PSD estimators is quite similar. As expected, without any speech leakage present, the three multichannel estimators achieve almost the same performance.
2) In speech-presence periods all four estimators tend to overestimate the noise PSD. The overestimation errors of the blocking matrix based noise PSD estimators are caused by the speech leakage (cf. (32)). For the single-channel UMMSE overestimation errors occur especially when the algorithm falsely detects the speech as noise. For all four estimators the overestimation error increases with larger input SNR because the speech leakage increases.
3) In speech-presence periods the MaxNSR noise PSD estimator achieves the lowest overestimation error, especially at high input SNRs, while achieving a similar underestimation error as the other estimators. It hence performs best among all the considered noise PSD estimators.

The speech enhancement performance of the four noise PSD estimators is shown in Fig. 10, in terms of $\mathrm{NRN}_{\mathrm{seg}}$ in noise-only periods, $\mathrm{NRS}_{\mathrm{seg}}$, $\mathrm{SD}_{\mathrm{seg}}$, $\mathrm{SNR}_{\mathrm{seg}}$ in speech-presence periods, and PESQ. As a reference, the $\mathrm{SNR}_{\mathrm{seg}}$ and PESQ of the speech

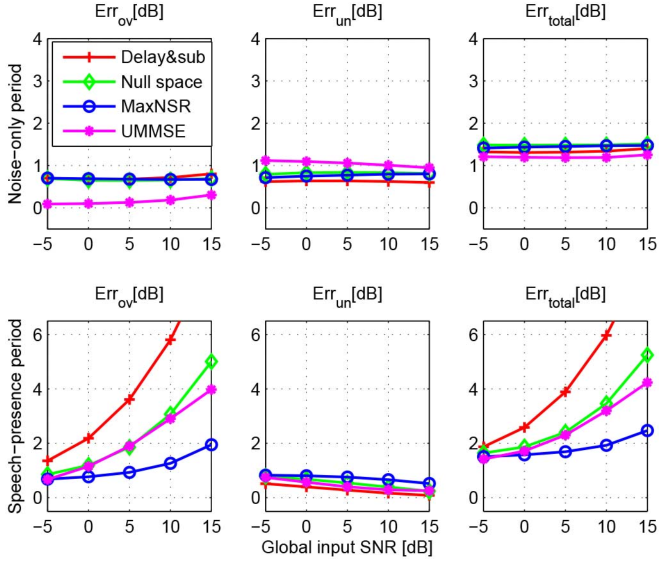


Fig. 9. Noise PSD estimation performance of the four considered noise PSD estimators in stationary diffuse noise at different global input SNRs (for noise-only and speech-presence periods separately).

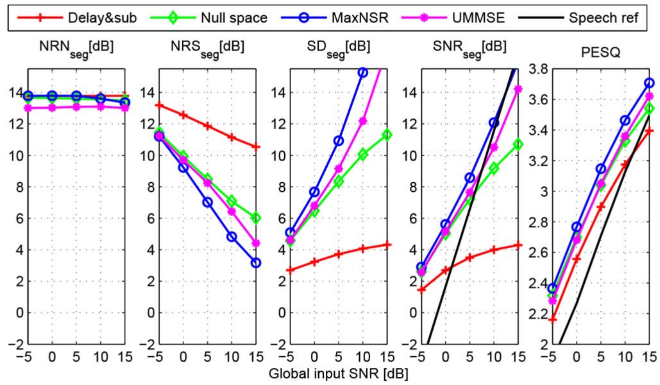


Fig. 10. Speech enhancement performance of the four considered noise PSD estimators in stationary diffuse noise at different global input SNRs.

reference y_r are also given. The following observations can be made:

1) In noise-only periods all four estimators achieve a similar noise reduction performance, with NRN$_{\mathrm{seg}}$ approaching 14 dB corresponding to $G_{\min}$. As expected, without any leakage present, the three multichannel estimators achieve almost the same performance.
2) In speech-presence periods the MaxNSR estimator performs better than the other two multichannel estimators in terms of SNR$_{\mathrm{seg}}$ by introducing less speech distortion (indicated by SD$_{\mathrm{seg}}$), especially at high input SNRs. The delay-and-subtract estimator suffers from a lot of speech leakage, leading to poor performance compared to other blocking matrix based estimators. The MaxNSR estimator also slightly outperforms the single-channel UMMSE estimator, which achieves better performance than the null space estimator.
3) In speech-presence periods all four estimators improve the SNR$_{\mathrm{seg}}$ of the speech reference y_r significantly at low SNRs. The improvement however becomes less evident or even negative at high SNRs where the speech distortion caused by overestimation errors becomes dominant in the SNR$_{\mathrm{seg}}$ measure. The MaxNSR estimator is the only estimator which does not show a decreased SNR$_{\mathrm{seg}}$ at high input SNRs (10 and 15 dB).
4) For all four PSD estimators the PESQ value shows a similar trend as the SNR$_{\mathrm{seg}}$ measure, with the MaxNSR estimator achieving the highest PESQ value for all SNRs.

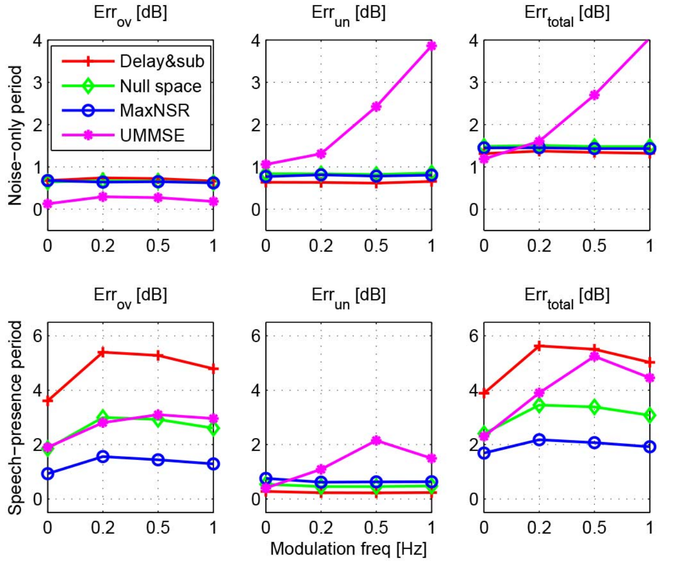


Fig. 11. Noise PSD estimation performance of the four considered noise PSD estimators in non-stationary diffuse noise with different modulation frequencies, at a global input SNR of 5 dB (for noise-only and speech-presence periods separately).

In summary, the speech enhancement results obtained by the four noise PSD estimators are consistent with the results for noise PSD estimation.

D2. Non-stationary noise: In this section the noise PSD estimation and speech enhancement performance of the four considered noise PSD estimators are compared in non-stationary (modulated) diffuse noise. Four modulation frequencies (f_m in (48)) are considered, i.e., 0 Hz, 0.2 Hz, 0.5 Hz, and 1 Hz.

Fig. 11 compares the noise PSD estimation performance (in terms of overestimation error, underestimation error and total error) of the four noise PSD estimators for different f_m. The global input SNR is 5 dB. The following observations can be made:

1) In noise-only periods the noise PSD estimation performance of the three multichannel estimators is quite similar and appears to be independent of f_m. In contrast, the performance of the single-channel UMMSE algorithm degrades significantly with increased f_m since it tends to underestimate the noise especially when the noise is fast-changing.
2) In speech-presence periods all three multichannel estimators tend to overestimate the noise PSD due to the occurring speech leakage. The MaxNSR estimator achieves the lowest overestimation error, while achieving a similar underestimation error as the other estimators. It performs best among all three estimators, followed by the null space estimator, while the delay-and-subtract estimator performs worst.

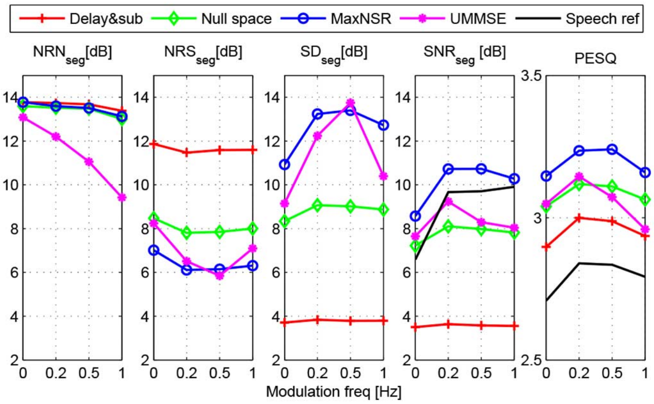


Fig. 12. Speech enhancement performance of the four considered noise PSD estimators in non-stationary diffuse noise with different modulation frequencies, at a global input SNR of 5 dB.

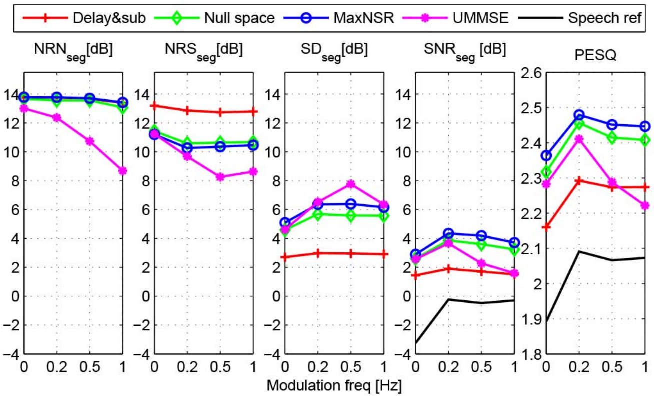


Fig. 13. Speech enhancement performance of the four considered noise PSD estimators in non-stationary diffuse noise with different modulation frequencies, at a global input SNR of -5 dB.

3) In speech-presence periods the estimation error does not vary greatly with f_m for the three multichannel estimators where the MaxNSR shows the smallest variation. The single-channel UMMSE on the other hand tends to both underestimate and overestimate the noise, and the underestimation error in particular grows with increasing f_m.

The speech enhancement performance of the four noise PSD estimators for two different input SNRs is shown in Fig. 12 and 13, in terms of $\mathrm{NRN_{seg}}$ in noise-only periods, $\mathrm{NRS_{seg}}$, $\mathrm{SD_{seg}}$, $\mathrm{SNR_{seg}}$ in speech-presence periods, and PESQ. As a reference, the $\mathrm{SNR_{seg}}$ and PESQ of the speech reference y_r are also given. The following observations can be made:

1) In noise-only periods all three multichannel estimators achieve almost the same noise reduction performance, which is moreover almost independent of f_m. In contrast, the single-channel UMMSE algorithm shows a degraded noise reduction performance with increasing f_m.
2) In speech-presence periods the MaxNSR estimator achieves larger $\mathrm{SNR_{seg}}$ than the null space estimator at both high (5 dB) and low (-5 dB) SNR scenarios. However, the difference between both estimators becomes smaller at low SNR. The delay-and-subtract estimator performs worst among all three multichannel estimators.
3) In speech-presence periods the single-channel UMMSE estimator performs worse than the MaxNSR estimator in all scenarios. Surprisingly, it achieves a larger $\mathrm{SNR_{seg}}$ than the null space estimator at high SNR (5 dB). The reason is that it tends to underestimate the noise PSD in speech-presence periods and hence shows less speech distortion ($\mathrm{SD_{seg}}$) and also less noise reduction ($\mathrm{NRS_{seg}}$). In global, it performs better in terms of $\mathrm{SNR_{seg}}$. However, this is not observed at low SNR (-5 dB), where the null space estimator achieves a larger $\mathrm{SNR_{seg}}$ than the UMMSE estimator.
4) At low SNR (-5 dB) all four estimators are able to significantly improve the $\mathrm{SNR_{seg}}$ relative to the speech reference y_r. However, at high SNR (5 dB), only the MaxNSR estimator is able to improve the $\mathrm{SNR_{seg}}$. The reason is that at high SNR the speech distortion caused by overestimation errors becomes dominant in the $\mathrm{SNR_{seg}}$ measure.
5) For all four estimators the PESQ value shows a similar trend as the $\mathrm{SNR_{seg}}$ measure. Both for high and low SNRs, all four estimators obtain improved PESQ values with respect to the speech reference, with the MaxNSR estimator achieving the highest PESQ value for all cases.

In summary, for nonstationary noise the speech enhancement performance based on the MaxNSR blocking matrix outperforms the other considered blocking matrices, and especially a single-channel algorithm, whose performance is significantly affected.

E. Babble Noise and Non-ideal VAD

In order to investigate the performance of the proposed algorithm for even more realistic acoustic scenarios, in this experiment we have used diffuse babble noise and a non-ideal VAD. The same microphone and target speaker configuration for the previous experiments is used. The diffuse babble noise is simulated using the method presented in [33], assuming a spherically isotropic noise field. The diffuse noise is added to the microphones at a global input SNR ranging from -5 dB to 15 dB. The non-ideal VAD, which is applied to the signal in the first microphone, has been implemented based on the UMMSE algorithm in [41], which also estimates the a *posteriori* speech presence probability at each time-frequency bin. The speech presence probability is averaged across all the frequencies. A time frame is flagged as a speech-presence frame when its averaged speech presence probability is above a threshold (we have used 0.2 as the threshold in the experiment).

Two objective measures are used: $\mathrm{SNR_{seg}}$ and PESQ. Fig. 14 compares the performance of considered algorithms with ideal and non-ideal VADs. As a reference, the $\mathrm{SNR_{seg}}$ and PESQ of the speech reference y_r are also given. The following observations can be made:

1) With an ideal VAD the performance of the four considered noise PSD estimators for diffuse babble noise is similar to the performance for white noise (cf. Fig. 10). The MaxNSR estimator performs best in terms of both $\mathrm{SNR_{seg}}$ and PESQ, followed by the UMMSE and null space estimators. Delay&sub performs the worst among all used algorithms.
2) When a non-ideal VAD is used, the performance of the fixed beamformer degrades especially at low SNRs, as can be observed from the degraded $\mathrm{SNR_{seg}}$ and PESQ values of the speech reference (i.e., fixed beamformer output) in

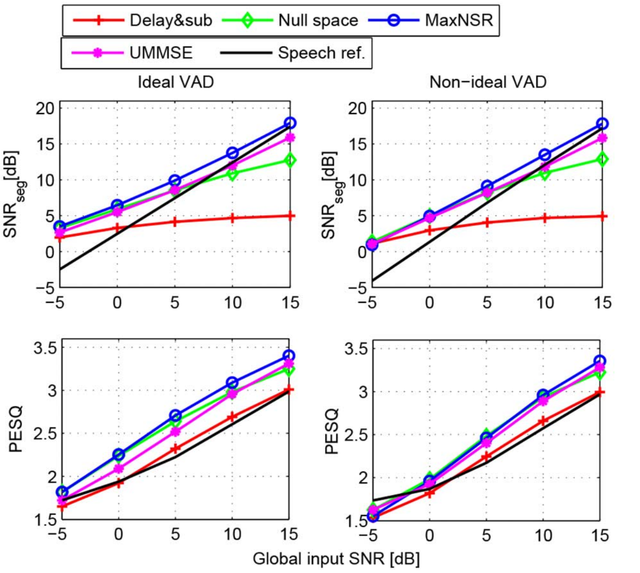


Fig. 14. Performance comparison of the four considered noise PSD estimators with ideal and non-ideal VAD, in diffuse babble noise at different global input SNRs.

comparison to the ideal VAD case. This is because that the noise coherence matrix and the RTF, which are used to construct the MVDR beamformer, are less accurately estimated due to VAD errors, especially at low SNRs.

3) Among the four noise PSD estimators, the MaxNSR and null space estimators are affected by VAD and noise coherence matrix estimation errors, as can be observed from the degraded SNR_{seg} and PESQ values compared to the ideal VAD case, especially at low SNRs. Since the UMMSE and Delay&sub estimators do not depend on the noise coherence matrix, their performance is almost not affected by VAD errors. However, even when using a non-ideal VAD, the MaxNSR estimator still outperforms the other estimators at SNRs larger than 5 dB, whereas the performance is very similar to the null space and UMMSE estimators at lower SNRs.

In summary, among the four noise PSD estimators, the MaxNSR estimator still performs best for diffuse babble noise with an ideal VAD. However, since its performance is sensitive to VAD errors, an accurate estimation of the noise coherence matrix is very important for the MaxNSR estimator.

VII. Conclusions

In this paper a GSC-like speech enhancement system is investigated for enhancing a single target speaker in diffuse noise. The PSD of the residual noise in the speech reference is estimated from the PSD of the noise reference using a scaling factor. To reduce the influence of speech leakage in the noise reference, a MaxNSR blocking matrix is proposed which aims to simultaneously suppress the target speech while maximizing the noise energy in the output of the blocking matrix. The MaxNSR blocking matrix can be computed using the generalized eigenvalue decomposition of the noisy speech correlation matrix and the noise coherence matrix, which is assumed to be known. It is shown that the vector of the MaxNSR blocking matrix is orthogonal to the acoustic transfer function of the target speech and can be represented as a linear combination of the vectors in the traditional null space blocking matrix. The following conclusions can be drawn from the experiments:

1) When more than two microphones are used, the MaxNSR blocking matrix can better suppress the speech leakage and hence performs better in terms of noise PSD estimation and speech enhancement than the null space blocking matrix.
2) In contrast to the single-channel noise PSD estimator, which has problems to track the PSD of nonstationary noise efficiently, the performance of the MaxNSR blocking matrix based noise PSD estimator is almost not affected by the nonstationarity of the noise. It therefore outperforms the single-channel noise PSD estimator especially when the noise is fast-changing.
3) Although the performance of the MaxNSR blocking matrix based noise PSD estimator is sensitive to noise coherence matrix estimation errors (e.g., due to VAD errors), it still outperforms the other considered signal- and multichannel estimators at SNRs larger than 5 dB, whereas its performance is very similar at lower SNRs.

Appendix A
Proof of the Equation (43)

Consider the generalized eigenvalue decomposition (GEVD)

$$\boldsymbol{A}\boldsymbol{w}_k = \lambda_k \boldsymbol{D}\boldsymbol{w}_k \tag{A-1}$$

of the matrices $\boldsymbol{A}$ and $\boldsymbol{D}$, with λ_k and $\boldsymbol{w}_k$, $k = 1, \cdots, M$, the generalized eigenvalues (in descending order) and generalized eigenvectors, respectively. For Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{D}$, it can be easily shown that the generalized eigenvalues are real-valued, i.e., $\lambda_k = \lambda_k^*$.

Given the i-th and j-th generalized eigenvector $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ and assuming that the corresponding generalized eigenvectors λ_i and λ_j are not equal ($\lambda_i \neq \lambda_j$), it can be shown using (A-1) that

$$\begin{cases} \boldsymbol{w}_j^{\text{H}}\boldsymbol{A}\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_j^{\text{H}}\boldsymbol{D}\boldsymbol{w}_i \\ \boldsymbol{w}_j^{\text{H}}\boldsymbol{A}\boldsymbol{w}_i = \lambda_j^* \boldsymbol{w}_j^{\text{H}}\boldsymbol{D}^{\text{H}}\boldsymbol{w}_i = \lambda_j \boldsymbol{w}_j^{\text{H}}\mathbf{D}\boldsymbol{w}_i \end{cases} \tag{A-2}$$

such that for $\lambda_i \neq \lambda_j$ it follows that

$$\boldsymbol{w}_j^{\text{H}}\boldsymbol{D}\boldsymbol{w}_i = 0. \tag{A-3}$$

If the matrix $\boldsymbol{A}$ is a rank-1 matrix, i.e.,

$$\boldsymbol{A} = \boldsymbol{a}\boldsymbol{a}^{\text{H}}, \tag{A-4}$$

the generalized eigenvector $\boldsymbol{w}_1$ corresponding to the largest generalized eigenvalue λ_1 is related to the vector $\boldsymbol{a}$ as

$$\boldsymbol{a} = \zeta \boldsymbol{D}\boldsymbol{w}_1 \tag{A-5}$$

with ζ a complex-valued scalar.

Using (A-3), it follows that all other generalized eigenvectors $\boldsymbol{w}_i$ $(i \neq 1)$ are orthogonal to $\boldsymbol{a}$, i.e.,

$$\boldsymbol{w}_i^{\text{H}}\boldsymbol{a} = 0, \forall i \neq 1. \tag{A-6}$$

References

[1] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[2] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.

[3] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1376, Oct. 1987.

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[5] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[6] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 497–507, Sep. 2000.

[7] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[8] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, pp. 19–38.

[9] H. L. Van Trees, *Optimum Array Processing*. New York, NY, USA: Wiley, 2002.

[10] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing - Applications to Real-World Problems*, J. Benesty and Y. Huang, Eds. Berlin, Germany: Springer-Verlag, 2003, pp. 155–194.

[11] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–14, 2006.

[12] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1420–1434, Jul. 2009.

[13] S. Markovich-Golan, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[14] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds. Hoboken, NJ, USA: Wiley, 2010, pp. 269–302.

[15] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 158–170, Jan. 2010.

[16] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925–4935, Sep. 2010.

[17] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 206–218, Jan. 2011.

[18] M. Crocco and A. Trucco, "Design of robust superdirective arrays with a tunable tradeoff between directivity and frequency-invariance," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2169–2181, May 2011.

[19] E. A. P. Habets and J. Benesty, "Multi-microphone noise reduction based on orthogonal noise signal decompositions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 6, pp. 1123–1133, Jun. 2013.

[20] 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, Sep. 2002.

[21] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Commun.*, vol. 49, no. 7–8, pp. 636–656, Jul. 2007.

[22] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1368–1381, May 2011.

[23] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[24] H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Blind Speech Separation*, S. Makino, T. W. Lee, and H. Sawada, Eds. Berlin, Germany: Springer-Verlag, 2007, pp. 47–78.

[25] L. Wang, H. Ding, and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," *EURASIP J. Audio, Speech, Music Process.*, pp. 1–13, 2010, Article ID 797962.

[26] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 549–557, Mar. 2011.

[27] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1913–1928, Sep. 2013.

[28] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, pp. 39–57.

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

[30] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. Int. Conf. Acoust., Speech, Signal Process*, New York, NY, USA, 1988, pp. 2578–2581.

[31] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2002.

[32] M. Rahmani, A. Akbari, B. Ayad, and B. Lithgow, "Noise cross PSD estimation using phase information in diffuse noise field," *Signal Process.*, vol. 89, no. 5, pp. 703–709, 2009.

[33] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 2911–2917, 2008.

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

[35] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 650–664, May 2009.

[36] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Comput. Speech Lang.*, vol. 27, no. 3, pp. 726–765, May 2013.

[37] D. Van Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," in *Proc. Int. Conf. Acoust., Speech, Signal Process*, Albuquerque, NM, USA, 1990, pp. 833–836.

[38] W. Herbordt and W. Kellermann, "Analysis of blocking matrices for generalized sidelobe cancellers for nonstationary broadband signals," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, USA, 2002, pp. 1–4.

[39] 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, Jul. 2001.

[40] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. Int. Conf. Acoust., Speech, Signal Process*, Dallas, TX, USA, 2010, pp. 4266–4269.

[41] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 20, no. 4, pp. 1383–1393, May 2012.

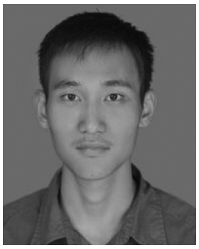
[42] J. Meyer, K. U. Simmer, and K. D. Kammeyer, "Comparison of one-and two-channel noise-estimation techniques," in *Proc. Int. Workshop on Acoustic Echo and Noise Control*, London, U.K., 1997, pp. 137–145.

[43] T. Wolff and M. Buck, "A generalized view on microphone array postfilters," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Tel-Aviv, Israel, 2010, pp. 1–4.

[44] M. Choi and H. Kang, "A two-channel noise estimator for speech enhancement in a highly nonstationary environment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 905–915, May 2011.

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

[46] H. Yi and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.



Lin Wang received the B.S. degree in electronic engineering from Tianjin University, China, in 2003, and the Ph.D. degree in signal processing from Dalian University of Technology, China, in 2010. From 2011 to 2013, he was an Alexander von Humboldt Fellow in the University of Oldenburg, Germany. Since 2014, he has been a Postdoctoral Researcher with Queen Mary University of London, UK. His research interests include video and audio compression, blind source separation, and 3D audio processing.

Timo Gerkmann (S'08–M'10–SM'15) studied electrical engineering at the universities of Bremen and Bochum, Germany. He received his Dipl.-Ing. degree in 2004 and his Dr.-Ing. degree in 2010 both at the Institute of Communication Acoustics (IKA) at the Ruhr-Universität at Bochum, Germany. In 2005, he spent six months with Siemens Corporate Research in Princeton, NJ, USA. During 2010 and 2011, Dr. Gerkmann was a Postdoctoral Researcher at the Sound and Image Processing Lab at the Royal Institute of Technology (KTH), Stockholm, Sweden. Since 2011, he has been a Professor for speech signal processing at the Universität Oldenburg, Oldenburg, Germany. His main research interests are digital speech and audio processing, including speech enhancement, dereverberation, modeling of speech signals, speech recognition, and hearing devices. He is a Senior Member of the IEEE.

Simon Doclo (S'95–M'03–SM'13) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Belgium, in 1997 and 2003. From 2003 to 2007, he was a Postdoctoral Fellow with the Research Foundation - Flanders at the Electrical Engineering Department (Katholieke Universiteit Leuven) and the Adaptive Systems Laboratory (McMaster University, Canada). From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors at the Sound and Acoustics Group in Leuven, Belgium. Since 2009, he has been a Full Professor at the University of Oldenburg, Germany, and scientific advisor for the project group Hearing, Speech, and Audio Technology of the Fraunhofer Institute for Digital Media Technology. His research activities center around signal processing for acoustical and biomedical applications, more specifically microphone array processing, active noise control, acoustic sensor networks and hearing aid processing.

Prof. Doclo received the Master Thesis Award of the Royal Flemish Society of Engineers in 1997 (with Erik De Clippel), the Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control in 2001, the EURASIP Signal Processing Best Paper Award in 2003 (with Marc Moonen) and the IEEE Signal Processing Society 2008 Best Paper Award (with Jingdong Chen, Jacob Benesty, Arden Huang). He was member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing (2008–2013) and Technical Program Chair for the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2013. Prof. Doclo has served as guest editor for several special issues (*IEEE Signal Processing Magazine*, *Elsevier Signal Processing*) and is associate editor for the *EURASIP Journal on Advances in Signal Processing*.