Speech Enhancement With a GSC-Like Structure Employing Eigenvector-Based Transfer Function Ratios Estimation

Alexander Krueger, *Student Member, IEEE*, Ernst Warsitz, *Member, IEEE*, and Reinhold Haeb-Umbach, *Senior Member, IEEE*

***Abstract*—In this paper, we present a novel blocking matrix and fixed beamformer design for a generalized sidelobe canceler for speech enhancement in a reverberant enclosure. They are based on a new method for estimating the acoustical transfer function ratios in the presence of stationary noise. The estimation method relies on solving a generalized eigenvalue problem in each frequency bin. An adaptive eigenvector tracking utilizing the power iteration method is employed and shown to achieve a high convergence speed. Simulation results demonstrate that the proposed beamformer leads to better noise and interference reduction and reduced speech distortions compared to other blocking matrix designs from the literature.**

***Index Terms*—Acoustical transfer function ratios, generalized eigenvalue problem, generalized sidelobe canceler (GSC), microphone array signal processing, speech enhancement.**

I. Introduction

In hands-free speech communication and recognition microphone array signal processing can be used to enhance a desired (speech) signal against background noise and interference. While single-channel methods can only exploit spectral characteristics of the received signals, multichannel methods allow the additional use of spatial information.

In contrast to data independent beamforming, where the design objective is given by the approximation of a desired response for certain directions, this paper will focus on statistically optimum beamforming. The goal is to compute an "optimal" array response based on the incoming data characteristics such that unwanted signals are maximally suppressed while leaving the desired signal unmodified. Typical approaches for statistically optimum beamforming comprise the maximization of the output signal-to-noise ratio (Max-SNR) and the minimization of the output power subject to some linear constraints on the array response, such as the linearly constrained minimum variance (LCMV) method [1]. The most popular example for an LCMV approach is the minimum variance distortionless response (MVDR) beamformer [2]. In [3], Frost proposed to use a constrained least mean squares (LMS) algorithm to obtain an adaptive MVDR solution. Griffiths and Jim introduced the generalized sidelobe canceler (GSC) [4] to transform the constrained minimization into an unconstrained one for ease of realization. The GSC consists of three signal processing blocks. The purpose of the so-called fixed beamformer (FB) is to satisfy the constraints to form a beam in the look direction which leaves the target signal undistorted while suppressing other signal components. The blocking matrix (BM) computes an estimate for the noise subspace of the microphone signals by blocking the speech signal components. The noise references at its output drive a multichannel adaptive noise canceler (ANC) whose coefficients are adapted to suppress the remaining noise in the FB output.

However, the construction of the blocking matrix according to [4] requires perfect knowledge of the direction of arrival (DoA) for the time alignment of the microphone signals. Further, both derivations [3] and [4] assume a free field propagation from the source to the sensors. Both conditions are hardly met in real environments and practical applications. Estimation errors in the DoA and reflections of signals by objects and walls cause leakage of the desired speech signal into the noise references resulting in signal cancellation in the beamformer output.

Several approaches have been proposed to overcome these problems. Hoshuyama *et al.* presented an adaptive blocking matrix (ABM) [5] to produce noise references orthogonal to the output of the FB. They furthermore introduced constraints on the ABM filter coefficients to improve the robustness against estimation errors of the DoA. Herbordt and Kellermann developed an efficient frequency domain realization of the GSC with ABM [6] and showed in [7] that for an optimal performance the adaptation may be carried out only in periods of noise absence. This, however, is a constraint which might be difficult to adhere to in practice. If adaptation of the ABM is carried out in the presence of noise, only suboptimal performance is attainable.

Gannot *et al.* introduced a blocking matrix containing transfer function ratios (TFRs) which can be estimated even in the presence of stationary noise [8]. They showed that knowledge of the ratios of transfer functions from the source to the individual sensors is sufficient to block the desired signal. Knowledge of the transfer functions themselves is not necessary. They proposed to estimate the TFRs using a least squares approach exploiting the nonstationarity of speech signals as opposed to the

Manuscript received June 03, 2008; revised February 22, 2010; accepted March 06, 2010. Date of publication April 01, 2010; date of current version October 01, 2010. This work was supported by the German Science Foundation (DFG) Research Training Group GK-693 of the Paderborn Institute for Scientific Computation (PaSCo) and by the DFG under Contract HA3455/4-1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jingdong Chen.

The authors are with the Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany (e-mail: krueger@nt.uni-paderborn.de; warsitz@nt.uni-paderborn.de; haeb@nt.uni-paderborn.de).

Digital Object Identifier 10.1109/TASL.2010.2047324

stationary character of the noise. Cohen extended the method using a weighted least squares approach which incorporates an indicator function for the presence of the desired signal [9].

Doclo and Moonen suggest filtering the microphone signals such that the mean squared error (MSE) between the speech components of the microphone signals and the filtered signals is minimized in the time domain [10]. They attempt to achieve this by computing a generalized singular value decomposition (GSVD) of data matrices consisting of noise only and noisy speech samples. Although they propose to use recursive algorithms for the GSVD [11] the computational effort remains large. In [12], a spatio-temporal prediction approach for noise reduction in reverberant environments is presented that constructs an optimal filter using a set of inter-sensor transformations. This principle is reformulated in the frequency domain in [13] in order to reduce the computational complexity at the cost of neglecting temporal correlations of the microphone signals.

In [14], a further frequency domain approach was proposed where for each frequency bin the SNR was maximized independently by the solution of a generalized eigenvalue problem (GEVP). While this led to a drastically reduced computational effort, using a narrowband Max-SNR criterion for a broadband beamforming problem introduces an uncontrollable amount of distortion of the desired signal. The reason for this is that the resulting beamformer frequency responses in each frequency bin are unique only up to a complex constant. In order to fix this problem and obtain a distortionless response, different single-channel postfilters have been suggested [14], [15].

To enhance a desired speech signal in the presence of stationary noise and additional nonstationary interference, Reuven *et al.* introduced the dual-source transfer function generalized sidelobe canceler (DS-TF-GSC) [16]. They use a matched beamformer (MB) to suppress the nonstationary interferer while the goal of the noise cancelling branch is to remove the stationary noise from the microphone signals. In [17], they analyzed the performance of the DS-TF-GSC for different acoustical environments and noise fields. They showed that the amount of suppression of the interference and the introduced speech distortion depend highly on the estimation accuracy of the TFRs. The estimation of the dual-source TFRs requires the identification of periods where the desired signal or the nonstationary interferer alone are present. This important and certainly nontrivial issue has not been addressed in [16]. However, several solutions have been proposed in the literature. For instance, if *a priori* information about the desired speaker is available, a model-based speaker segmentation and identification approach using Hidden Markov modeling and Viterbi decoding [18] can be applied. Alternatively, it is possible to perform an online clustering of the time-differences of arrival (TDoA) between the signals of two different sensors [15], [19]. An overview of approaches to obtain TDoA estimates in reverberant environments, comprising the well-known generalized correlation method [20], is given in [21].

In this paper, we present an extension of the approach in [22] for the construction of a blocking matrix and fixed beamformer for speech enhancement with a GSC-like structure in a reverberant enclosure. Our proposed approach requires the estimation of the speaker-sensor TFRs, if the scenario with only stationary noise is considered. If, additionally, a nonstationary interferer is included, the interferer-sensor TFRs have to be estimated as well. We derive the procedure for obtaining the TFRs from a generalized eigenvector decomposition and develop an algorithm to compute the TFRs from the multichannel input signal. Simulation results confirm that the performance of the GSC is noticeably improved by using the eigenvector-based TFR estimation for the computation of the blocking matrix (BM) and fixed beamformer (FB) compared to using previously proposed methods for the computation of the BM and FB from the literature.

This paper is organized as follows. In Section II, we first introduce the considered hands-free communication scenario as well as the notation used in the paper. We then recapitulate the GSC structure and explain the purpose of each GSC component. Subsequently, we show how the solution of the narrowband Max-SNR criterion can be used for the estimation of TFRs from the speaker to the microphones. These TFRs are then employed for the construction of a blocking matrix to be used in a GSC for the suppression of stationary noise. The scenario is then extended by assuming the presence of a nonstationary interferer (e.g., another speaker), in addition to the stationary noise. The GSC is modified in a manner similar to that used in [16]; however, here using the TFRs obtained from the generalized eigenvector decomposition in the construction of the blocking matrix and the fixed beamformer. In Section III, we present a method for the adaptive tracking of the principal eigenvector of a GEVP which is needed for the TFR estimation. Finally, in Section IV we demonstrate the tracking ability of the eigenvector estimation and then present simulation results where we compare the proposed algorithms with selected algorithms from the literature.

II. Speech Enhancement

A. Problem Formulation and Notation

We consider a typical hands-free speech communication scenario (see Fig. 1) where an array of M microphones is located in a reverberant enclosure. Each discrete-time microphone signal $x_i(l), (i = 1, \ldots, M)$, where l denotes the time index and i indicates the microphone, is assumed to consist of three components: a desired signal component $s_{i,\mathrm{D}}(l)$, an interfering signal component $s_{i,\mathrm{I}}(l)$ and a stationary noise term $n_i(l)$

$$\begin{aligned} x_i(l) &= s_{i,\mathrm{D}}(l) + s_{i,\mathrm{I}}(l) + n_i(l) \\ &= \sum_{l'=0}^{\infty} h_{i,l,\mathrm{D}}(l') s_{0,\mathrm{D}}(l - l') \\ &\quad + \sum_{l'=0}^{\infty} h_{i,l,\mathrm{I}}(l') s_{0,\mathrm{I}}(l - l') + n_i(l). \end{aligned} \tag{1}$$

The desired signal component $s_{i,\mathrm{D}}(l)$ is assumed to result from the convolution of the desired (speech) source signal $s_{0,\mathrm{D}}(l)$ with the time-variant room impulse response $h_{i,l,\mathrm{D}}(l')$ from the desired source position to the ith microphone. Here, $h_{i,l,\mathrm{D}}(l')$ denotes the room response to an impulse occurring at the desired source position at the time-index l. Accordingly, the

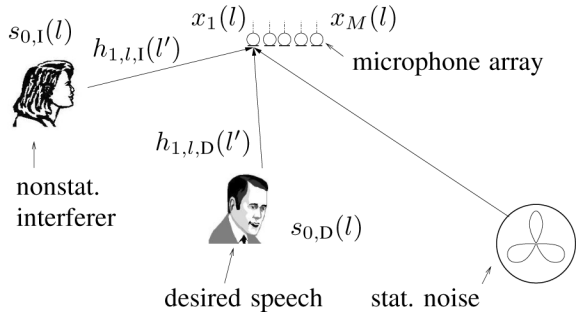

Fig. 1. Typical hands-free speech communication scenario.

interfering signal component $s_{i,\mathrm{I}}(l)$ is assumed to result from the convolution of the interfering (speech) source signal $s_{0,\mathrm{I}}(l)$ with the time-variant room impulse response $h_{i,l',\mathrm{I}}(l)$ from the interfering source position to the ith microphone. In the considered scenario the noise component $n_i(l)$ is assumed to comprise two components: a directional component originating from a point source, e.g., a fan or a personal computer, as well as a noncoherent component occurring in practice due to imperfect sensors. Collecting all microphone signals in a vector

$$\mathbf{x}(l) := [x_1(l), \ldots, x_M(l)]^T \tag{2}$$

and doing so in an analogous way with the signal components and the room impulse responses

$$\mathbf{s}_{\mathrm{D}}(l) := [s_{1,\mathrm{D}}(l), \ldots, s_{M,\mathrm{D}}(l)]^T \tag{3}$$
$$\mathbf{s}_{\mathrm{I}}(l) := [s_{1,\mathrm{I}}(l), \ldots, s_{M,\mathrm{I}}(l)]^T \tag{4}$$
$$\mathbf{n}(l) := [n_1(l), \ldots, n_M(l)]^T \tag{5}$$
$$\mathbf{h}_{l',\mathrm{D}}(l) := [h_{1,l',\mathrm{D}}(l), \ldots, h_{M,l',\mathrm{D}}(l)] \tag{6}$$
$$\mathbf{h}_{l',\mathrm{I}}(l) := [h_{1,l',\mathrm{I}}(l), \ldots, h_{M,l',\mathrm{I}}(l)] \tag{7}$$

where $(\cdot)^T$ denotes transposition, the signal model (1) can be reformulated in a compact vector notation

$$\begin{aligned}\mathbf{x}(l) &= \mathbf{s}_{\mathrm{D}}(l) + \mathbf{s}_{\mathrm{I}}(l) + \mathbf{n}(l) \\ &= \sum_{l'=0}^{\infty} \mathbf{h}_{l,\mathrm{D}}(l') s_{0,\mathrm{D}}(l-l') \\ &\quad + \sum_{l'=0}^{\infty} \mathbf{h}_{l,\mathrm{I}}(l') s_{0,\mathrm{I}}(l-l') + \mathbf{n}(l).\end{aligned} \tag{8}$$

The goal of multichannel speech enhancement is to recover the desired speech signal $s_{0,\mathrm{D}}(l)$ from the microphone signals $x_i(l)$, while suppressing all other interfering and noise components. For this purpose, the microphone signals are typically processed in blocks. This can be expressed by a windowing operation

$$x_i(l,m) := x_i(l) \cdot w(l - mB) \tag{9}$$

where $w(l)$ denotes the discrete-time window function of length L. Here B denotes the advance between successive blocks and m is the block index. The overlap between successive data blocks is then given by $L - B$.

Further, the room impulse responses are considered to be slowly changing in time. To be specific, they are assumed to be time-invariant for the duration B of a frame advance. The windowed room impulse responses corresponding to the mth block are defined by

$$h_{i,\mathrm{D}}(l,m) := h_{i,mB,\mathrm{D}}(l) \cdot w(l) \tag{10}$$
$$h_{i,\mathrm{I}}(l,m) := h_{i,mB,\mathrm{I}}(l) \cdot w(l). \tag{11}$$

Note that here the second argument m of $h_{i,\mathrm{D}}(l,m)$ and $h_{i,\mathrm{I}}(l,m)$ indicates the time variance of the impulse response, while m in $x_i(l,m)$ of (9) denotes the block index. However, we preferred to live with this double meaning to avoid a clumsy notation. The short-time discrete Fourier transform (STDFT) of the signal $x_i(l)$, at time indices mB, is denoted by $X_i(k,m)$, i.e.,

$$X_i(k,m) := \left(\sum_{l=mB}^{mB+L-1} x_i(l,m) e^{-j\omega T(l-mB)} \right) \Bigg|_{\omega=\frac{2\pi k}{LT}}. \tag{12}$$

Accordingly, we define the STDFT of the source signals $s_{0,\mathrm{D}}(l), s_{0,\mathrm{I}}(l)$ by $S_{0,\mathrm{D}}(k,m), S_{0,\mathrm{I}}(k,m)$, the STDFTs of the signal components $s_{i,\mathrm{D}}(l)$ and $s_{i,\mathrm{I}}(l)$ at the microphones by $S_{i,\mathrm{D}}(k,m,)$ $S_{i,\mathrm{I}}$ (k,m), the STDFTs of the widowed room impulse responses $h_{i,\mathrm{D}}(l,m)$ and $h_{i,\mathrm{I}}(l,m)$ by $H_{i,\mathrm{D}}(k,m), H_{i,\mathrm{I}}(k,m)$ and finally, the STDFT of the noise component $n_i(l)$ by $N_i(k,m)$. We assume a frequency resolution of L such that the frequency index $k, (k = 0, \ldots, L-1)$, corresponds to the angular frequency $2\pi k/(TL)$, where $1/T$ is the sampling frequency. For clarity and simplification of the following presentation, we use vector notation

$$\mathbf{X}(k,m) := [X_1(k,m), \ldots, X_M(k,m)]^T. \tag{13}$$

Accordingly, we define the STDFT vectors $\mathbf{S}_{\mathrm{D}}(k,m), \mathbf{S}_{\mathrm{I}}(k,m), \mathbf{H}_{\mathrm{D}}(k,m), \mathbf{H}_{\mathrm{I}}(k,m)$, and $\mathbf{N}(k,m)$.

Using the multiplicative transfer function approximation (MTFA) [23], an assumption commonly made in frequency domain acoustic signal processing and which is justified when the window length L is large compared to the duration of the room impulse response, (8) can be approximated in the STDFT domain by

$$\begin{aligned}\mathbf{X}(k,m) &= S_{0,\mathrm{D}}(k,m)\mathbf{H}_{\mathrm{D}}(k,m) \\ &\quad + S_{0,\mathrm{I}}(k,m)\mathbf{H}_{\mathrm{I}}(k,m) + \mathbf{N}(k,m) \qquad (14) \\ &= S_{1,\mathrm{D}}(k,m)\bar{\mathbf{H}}_{\mathrm{D}}(k,m) \\ &\quad + S_{1,\mathrm{I}}(k,m)\bar{\mathbf{H}}_{\mathrm{I}}(k,m) + \mathbf{N}(k,m) \qquad (15)\end{aligned}$$

where

$$\bar{\mathbf{H}}_{\mathrm{D}}(k,m) := \left[1, \frac{H_{2,\mathrm{D}}(k,m)}{H_{1,\mathrm{D}}(k,m)} \ldots \frac{H_{M,\mathrm{D}}(k,m)}{H_{1,\mathrm{D}}(k,m)}\right]^T \tag{16}$$

$$\bar{\mathbf{H}}_{\mathrm{I}}(k,m) := \left[1, \frac{H_{2,\mathrm{I}}(k,m)}{H_{1,\mathrm{I}}(k,m)} \ldots \frac{H_{M,\mathrm{I}}(k,m)}{H_{1,\mathrm{I}}(k,m)}\right]^T \tag{17}$$

denote the transfer function ratios (TFRs) with respect to the first transfer function and where we assumed that the STDFTs of

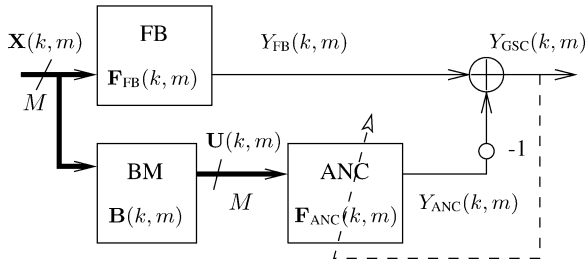

Fig. 2. GSC structure.

the signal components $s_{i,\mathrm{D}}(l)$ and $s_{i,\mathrm{I}}(l)$ can be approximated by

$$S_{1,\mathrm{D}}(k,m) = S_{0,\mathrm{D}}(k,m)H_{1,\mathrm{D}}(k,m) \tag{18}$$

$$S_{1,\mathrm{I}}(k,m) = S_{0,\mathrm{I}}(k,m)H_{1,\mathrm{I}}(k,m). \tag{19}$$

Moreover, we have assumed in (15) that $H_{1,\mathrm{D}}(k,m) \neq 0$ and $H_{1,\mathrm{I}}(k,m) \neq 0$.

B. Generalized Sidelobe Canceler

In order to recover the desired signal $s_{0,\mathrm{D}}(l)$ from the noisy microphone signals $\mathbf{x}(l)$, we employ the structure of a generalized sidelobe canceler (GSC) which is depicted in Fig. 2. The structure consisting of three signal processing units is obtained if the MVDR approach is reformulated with the objective being to minimize the output signal variance under the constraint that the desired signal is left undistorted.

The purpose of the fixed beamformer (FB) is to satisfy the constraint to form a beam in the look direction which leaves the target signal undistorted while suppressing other signal components. With the fixed beamformer transfer functions $F^*_{i,\mathrm{FB}}(k,m)$, which are defined in accordance with the acoustical transfer functions, the fixed beamformer output $Y_{\mathrm{FB}}(k,m)$ in the STDFT domain is given by

$$\begin{aligned} Y_{\mathrm{FB}}(k,m) &= \sum_{i=1}^{M} F^*_{i,\mathrm{FB}}(k,m)X_i(k,m) \\ &= \mathbf{F}^H_{\mathrm{FB}}(k,m)\mathbf{X}(k,m) \end{aligned} \tag{20}$$

where $(\cdot)^*$ denotes the complex conjugation and $(\cdot)^H$ denotes the complex conjugated transposition.

The blocking matrix (BM) computes an estimate for the noise subspace of the microphone signals by blocking the speech signal components. It relates the STDFTs of the microphone signals $\mathbf{X}(k,m)$ at its input to the STDFTs of the output noise references $\mathbf{U}(k,m)$ by means of

$$\mathbf{U}(k,m) = \mathbf{B}^H(k,m)\mathbf{X}(k,m). \tag{21}$$

The noise references at its output drive a multichannel adaptive noise canceler (ANC) whose coefficients are adapted to suppress the remaining noise in the FB output. The adaptation of the ANC transfer functions $\mathbf{F}_{\mathrm{ANC}}(k,m)$ can be carried out in periods of time when only stationary noise is present in the microphone signals. The ANC output is formulated by

$$Y_{\mathrm{ANC}}(k,m) = \mathbf{F}^H_{\mathrm{ANC}}(k,m)\mathbf{U}(k,m). \tag{22}$$

Finally, the GSC output is obtained by

$$Y_{\mathrm{GSC}}(k,m) = Y_{\mathrm{FB}}(k,m) - Y_{\mathrm{ANC}}(k,m). \tag{23}$$

The main concern of the paper will be to propose structures for the blocking matrix and the fixed beamformer for two different scenarios. The first scenario considers the case where only the desired speaker and the stationary noise are present. In the second scenario, the nonstationary interferer is included as well. For the first case, the construction of the proposed BM and the FB requires the knowledge of the desired TFRs $\bar{\mathbf{H}}_{\mathrm{D}}(k,m)$ only. In the second case, both TFRs, $\bar{\mathbf{H}}_{\mathrm{D}}(k,m)$ and $\bar{\mathbf{H}}_{\mathrm{I}}(k,m)$, are employed.

Generally, the acoustical environment where the beamformer is used, is changing over time because of speaker movements, temperature changes, etc. For that reason the components of a beamformer, or in our case the GSC, have to be time-varying in order to be able to cope with the changes. However, for the derivation of the transfer functions of the GSC components, it is convenient to assume a steady state. That is why in the following we consider a short period of time, which covers the duration of several blocks, where we assume the TFRs to be time-invariant, i.e., $\bar{\mathbf{H}}_{\mathrm{D}}(k,m) = \bar{\mathbf{H}}_{\mathrm{D}}(k)$ and $\bar{\mathbf{H}}_{\mathrm{I}}(k,m) = \bar{\mathbf{H}}_{\mathrm{I}}(k)$, such that (15) can be reformulated as follows:

$$\mathbf{X}(k,m) = S_{1,\mathrm{D}}(k,m)\bar{\mathbf{H}}_{\mathrm{D}}(k) + S_{1,\mathrm{I}}(k,m)\bar{\mathbf{H}}_{\mathrm{I}}(k) + \mathbf{N}(k,m). \tag{24}$$

Note that the TFRs are much less influenced by changes in the acoustical environment than the transfer functions themselves, such that the commonly made assumption of temporally time-invariant TFRs is reasonable and not too restrictive. In practice, hence, the TFRs are changing over time and we will address this issue in Section III.

In the next section, we will address the TFR estimation before we will present the BM and FB design.

C. Eigenvector Based TFR Estimation

For the TFR estimation we consider periods of time with desired speech signal and stationary noise components only, and periods of time with interferer signal and stationary noise components only. The identification of such periods is an important research issue in its own right. A sufficiently comprehensive treatment of this topic is beyond the scope of this paper and we refer to our own and other work, e.g., [15], and [18], where possible solutions are proposed. As in [16] we assume here that such periods are available and can be detected properly. Without loss of generality, we assume for the further derivation a period of time where only the desired speech signal and stationary noise is present, i.e.,

$$\begin{aligned} \mathbf{X}(k,m) &= \mathbf{S}_{\mathrm{D}}(k,m) + \mathbf{N}(k,m) \\ &= S_{1,\mathrm{D}}(k,m)\bar{\mathbf{H}}_{\mathrm{D}}(k) + \mathbf{N}(k,m). \end{aligned} \tag{25}$$

We will derive the desired TFRs from the transfer functions $\mathbf{F}(k,m)$ of a frequency-domain Max-SNR beamformer whose output in STDFT domain is given by

$$Y(k,m) = \mathbf{F}^H(k,m)\mathbf{X}(k,m). \tag{26}$$

Note that this is only an auxiliary construction and this output will actually not be computed in practice.

The objective function of the frequency-domain Max-SNR beamformer is to determine the beamformer coefficients $\mathbf{F}(k,m)$ such that the signal-to-noise ratio (SNR) is maximized

$$\mathbf{F}_{\mathrm{SNR,D}} := \underset{\mathbf{F}(k,m)}{\operatorname{argmax}} \mathrm{SNR}_{\mathrm{D}}(k,m) \tag{27}$$

where the SNR with respect to the desired signal $\mathbf{S}_{\mathrm{D}}(k,m)$ of the kth frequency bin and mth signal block is defined by

$$\mathrm{SNR}_{\mathrm{D}}(k,m) := \frac{\mathbf{F}^H(k,m)\mathbf{\Phi}_{\mathbf{S}_{\mathrm{D}}\mathbf{S}_{\mathrm{D}}}(k,m)\mathbf{F}(k,m)}{\mathbf{F}^H(k,m)\mathbf{\Phi}_{\mathbf{NN}}(k,m)\mathbf{F}(k,m)}. \tag{28}$$

The short-time power spectral densities (PSDs) of speech and noise are defined by

$$\mathbf{\Phi}_{\mathbf{S}_{\mathrm{D}}\mathbf{S}_{\mathrm{D}}}(k,m) := E\left[\mathbf{S}_{\mathrm{D}}(k,m)\mathbf{S}_{\mathrm{D}}^H(k,m)\right] \tag{29}$$

$$\mathbf{\Phi}_{\mathbf{NN}}(k) := E[\mathbf{N}(k,m)\mathbf{N}^H(k,m)]. \tag{30}$$

The expectation is conducted over all realizations of the signals in the mth block. The dependence of $\mathbf{\Phi}_{\mathbf{S}_{\mathrm{D}}\mathbf{S}_{\mathrm{D}}}(k,m)$ on the block index m is obvious, as speech signals are nonstationary. On the contrary, the noise has been assumed to be stationary resulting in a PSD being independent of the block index.

We further assume that speech and noise are uncorrelated and that each of the signals has zero mean. This allows us to split the PSD of the noisy microphone signals

$$\mathbf{\Phi}_{\mathbf{X}_{\mathrm{D}}\mathbf{X}_{\mathrm{D}}}(k,m) := E[\mathbf{X}(k,m)\mathbf{X}^H(k,m)]|_{\mathbf{X}=\mathbf{S}_{\mathrm{D}}+\mathbf{N}} \tag{31}$$

into two parts

$$\mathbf{\Phi}_{\mathbf{X}_{\mathrm{D}}\mathbf{X}_{\mathrm{D}}}(k,m) = \mathbf{\Phi}_{\mathbf{S}_{\mathrm{D}}\mathbf{S}_{\mathrm{D}}}(k,m) + \mathbf{\Phi}_{\mathbf{NN}}(k). \tag{32}$$

Now the solution of (27) is the eigenvector belonging to the largest eigenvalue $\lambda_{\mathrm{MAX,D}}(k,m)$ of the generalized eigenvalue problem (GEVP) [14]

$$\mathbf{\Phi}_{\mathbf{X}_{\mathrm{D}}\mathbf{X}_{\mathrm{D}}}(k,m)\mathbf{F}_{\mathrm{SNR,D}}(k,m) = \lambda_{\mathrm{MAX,D}}(k,m)\mathbf{\Phi}_{\mathbf{NN}}(k)\mathbf{F}_{\mathrm{SNR,D}}(k,m). \tag{33}$$

The GEVP can be transformed to the special eigenvalue problem (SEVP), if $\mathbf{\Phi}_{\mathbf{NN}}(k)$ is not singular

$$\mathbf{\Phi}_{\mathbf{NN}}^{-1}(k)\mathbf{\Phi}_{\mathbf{X}_{\mathrm{D}}\mathbf{X}_{\mathrm{D}}}(k,m)\mathbf{F}_{\mathrm{SNR,D}}(k,m) = \lambda_{\mathrm{MAX,D}}(k,m)\mathbf{F}_{\mathrm{SNR,D}}(k,m). \tag{34}$$

With the assumption of time-invariant TFRs (24) the PSD of the microphone signals can be rewritten as

$$\mathbf{\Phi}_{\mathbf{X}_{\mathrm{D}}\mathbf{X}_{\mathrm{D}}}(k,m) = \Phi_{S_{1,\mathrm{D}}S_{1,\mathrm{D}}}(k,m)\bar{\mathbf{H}}_{\mathrm{D}}(k)\bar{\mathbf{H}}_{\mathrm{D}}^H(k) + \mathbf{\Phi}_{\mathbf{NN}}(k) \tag{35}$$

where

$$\Phi_{S_{1,\mathrm{D}}S_{1,\mathrm{D}}}(k,m) = E[|S_{1,\mathrm{D}}(k,m)|^2]. \tag{36}$$

Using (35) in (34), the SEVP (34) can be reformulated as follows:

$$\mathbf{\Phi}_{\mathbf{NN}}^{-1}(k)\bar{\mathbf{H}}_{\mathrm{D}}(k)\bar{\mathbf{H}}_{\mathrm{D}}^H(k)\mathbf{F}_{\mathrm{SNR,D}}(k,m) = \frac{\lambda_{\mathrm{MAX,D}}(k,m)-1}{\Phi_{S_{1,\mathrm{D}}S_{1,\mathrm{D}}}(k,m)}\mathbf{F}_{\mathrm{SNR,D}}(k,m) \tag{37}$$

where we assume that $\Phi_{S_{1,\mathrm{D}}S_{1,\mathrm{D}}}(k,m) \neq 0$.

As the rank of the positive semidefinite matrix $\mathbf{\Phi}_{\mathbf{NN}}^{-1}(k)\bar{\mathbf{H}}_{\mathrm{D}}(k)\bar{\mathbf{H}}_{\mathrm{D}}^H(k)$ is one there is obviously only one eigenvector belonging to an eigenvalue greater than zero, which is given by

$$\mathbf{F}_{\mathrm{SNR,D}}(k) = \zeta(k)\mathbf{\Phi}_{\mathbf{NN}}^{-1}(k)\bar{\mathbf{H}}_{\mathrm{D}}(k) \tag{38}$$

where $\zeta(k)$ is an arbitrary complex constant. The result for the eigenvector can be easily verified by substituting (38) into (37). This eigenvector which maximizes (28) is, under the stated assumptions, obviously independent of the block number m, since only the scalar $(\lambda_{\mathrm{MAX,D}}(k,m)-1)/(\Phi_{S_{1,\mathrm{D}}S_{1,\mathrm{D}}}(k,m))$ and thus the eigenvalue corresponding to $\mathbf{F}_{\mathrm{SNR,D}}(k)$, but not $\mathbf{F}_{\mathrm{SNR,D}}(k)$ itself, depends on m.

The drawback of the Max-SNR approach (27) is that the resulting eigenvectors are ambiguous. That means that the Max-SNR criterion determines only the relationship between the components of the eigenvectors, but not their absolute values. This can be seen from the presence of $\zeta(k)$ in (38). A unique solution is achieved by postulating a distortionless response to the desired signal

$$\mathbf{F}_{\mathrm{SNR,D}}^H(k)\mathbf{H}_{\mathrm{D}}(k,m) \stackrel{!}{=} 1 \tag{39}$$

which results in a block-dependent scaling constant

$$\zeta(k,m) = \frac{H_{1,\mathrm{D}}(k,m)}{\mathbf{H}_{\mathrm{D}}^H(k,m)\mathbf{\Phi}_{\mathbf{NN}}^{-1}(k)\mathbf{H}_{\mathrm{D}}(k,m)}. \tag{40}$$

Furthermore, it can be seen that the multiplication of $\mathbf{\Phi}_{\mathbf{NN}}(k)$ by the ambiguous optimal beamformer transfer functions results in scaled versions of the TFRs

$$\mathbf{\Phi}_{\mathbf{NN}}(k)\mathbf{F}_{\mathrm{SNR,D}}(k) = \zeta(k)\bar{\mathbf{H}}_{\mathrm{D}}(k). \tag{41}$$

In the following, we denote these undetermined transfer functions by $\tilde{\mathbf{H}}_{\mathrm{D}}(k)$, i.e.,

$$\tilde{\mathbf{H}}_{\mathrm{D}}(k) := \zeta(k)\bar{\mathbf{H}}_{\mathrm{D}}(k). \tag{42}$$

Note that the ratios of the components of $\tilde{\mathbf{H}}_{\mathrm{D}}(k)$ to, say, $\tilde{H}_{1,\mathrm{D}}(k)$ are identical to the true TFRs

$$\frac{\tilde{\mathbf{H}}_{\mathrm{D}}(k)}{\tilde{H}_{1,\mathrm{D}}(k)} = \bar{\mathbf{H}}_{\mathrm{D}}(k). \tag{43}$$

In practice, the transfer functions $\mathbf{H}_{\mathrm{D}}(k,m)$ are unknown and $\zeta(k,m)$ cannot be computed from (40). Thus, to compensate for the distortions of the desired speech signal introduced by the wrong choice of $\zeta(k,m)$ different kinds of postfilters have been presented in [14] and [15].

It can be seen that the Max-SNR beamformer transfer functions $\mathbf{F}_{\mathrm{SNR,D}}(k)$ are related to the desired TFRs $\bar{\mathbf{H}}_{\mathrm{D}}$ by (41)–(43). An analogous relationship between the beamformer transfer functions $\mathbf{F}_{\mathrm{SNR,I}}(k)$ and the TFRs $\bar{\mathbf{H}}_{\mathrm{I}}$ is obtained if periods with stationary noise and nonstationary interferer only are considered. In the following, we will show how this relationship can be exploited for the construction of a blocking matrix for a GSC, where the scaling problem does not occur.

D. Generalized Eigenvector Blocking Matrix and Fixed Beamformer

1) Stationary Noise Only: We first deal with the case when there is no nonstationary interferer such that the desired speaker and the stationary noise only are present. The purpose of the blocking matrix is then to conduct a projection of the microphone signals into the orthogonal complement of $\mathbf{H}_{\mathrm{D}}(k)$. The following development for the construction of the blocking matrix is closely related to the adaptive blocking matrix (ABM) by Hoshuyama [5], where the goal is to create noise reference signals $u_i(l), i \in \{1, \ldots, M\}$ orthogonal to a speech reference signal. Let

$$Y_{\mathrm{SNR,D}}(k,m) := \mathbf{F}^H_{\mathrm{SNR,D}}(k)\mathbf{X}(k,m) \tag{44}$$

be the speech reference signal which is to be blocked by the blocking matrix. Note that this reference signal $Y_{\mathrm{SNR,D}}(k,m)$ is only an auxiliary construction and will not be directly used for the blocking matrix. We are now looking for a projection vector $\mathbf{P}_{\mathrm{D}}(k,m)$ such that the noise references at the output of the blocking matrix defined by

$$\mathbf{U}_{\mathrm{D}}(k,m) := \mathbf{X}(k,m) - \mathbf{P}_{\mathrm{D}}(k,m)Y_{\mathrm{SNR,D}}(k,m) \tag{45}$$

are orthogonal to the speech reference signal

$$E[\mathbf{U}_{\mathrm{D}}(k,m)Y^*_{\mathrm{SNR,D}}(k,m)] \stackrel{!}{=} \mathbf{0}. \tag{46}$$

The subscript D in $\mathbf{P}_{\mathrm{D}}(k,m)$ and $\mathbf{U}_{\mathrm{D}}(k,m)$ indicates that the desired component is to be blocked in the noise references. Using (33) we readily find

$$\mathbf{P}_{\mathrm{D}}(k) = \frac{\mathbf{\Phi_{NN}}(k)\mathbf{F}_{\mathrm{SNR,D}}(k)}{\mathbf{F}^H_{\mathrm{SNR,D}}(k)\mathbf{\Phi_{NN}}(k)\mathbf{F}_{\mathrm{SNR,D}}(k)} \tag{47}$$

which is obviously independent of the block index m. The blocking matrix $\mathbf{B}_{\mathrm{D}}(k,m)$ can now be easily found by using (44) in (45)

$$\mathbf{U}_{\mathrm{D}}(k,m) = \left[\mathbf{I}_M - \mathbf{P}_{\mathrm{D}}(k)\mathbf{F}^H_{\mathrm{SNR,D}}(k)\right]\mathbf{X}(k,m) \tag{48}$$

and comparing (48) with (21) to be

$$\mathbf{B}^H_{\mathrm{D}}(k) = \mathbf{I}_M - \mathbf{P}_{\mathrm{D}}(k)\mathbf{F}^H_{\mathrm{SNR,D}}(k). \tag{49}$$

Here, $\mathbf{I}_M$ denotes the identity matrix of dimension $M \times M$. Note that the result for $\mathbf{B}^H_{\mathrm{D}}(k)$ is block independent and can be computed from (49) and (47) if estimates of $\mathbf{\Phi_{NN}}(k)$ and $\mathbf{F}_{\mathrm{SNR,D}}(k)$ are available.

The projection into the orthogonal complement of $\mathbf{H}_{\mathrm{D}}(k)$, which is equal to the orthogonal complement of $\tilde{\mathbf{H}}_{\mathrm{D}}(k)$, becomes obvious if the projection vector $\mathbf{P}_{\mathrm{D}}(k)$ in (47) is rewritten using (41)

$$\mathbf{P}_{\mathrm{D}}(k) = \frac{\tilde{\mathbf{H}}_{\mathrm{D}}(k)}{\mathbf{F}^H_{\mathrm{SNR,D}}(k)\tilde{\mathbf{H}}_{\mathrm{D}}(k)}. \tag{50}$$

Using this in (49) we obtain for the blocking matrix

$$\mathbf{B}^H_{\mathrm{D}}(k) = \mathbf{I}_M - \frac{\tilde{\mathbf{H}}_{\mathrm{D}}(k)\mathbf{F}^H_{\mathrm{SNR,D}}(k)}{\mathbf{F}^H_{\mathrm{SNR,D}}(k)\tilde{\mathbf{H}}_{\mathrm{D}}(k)} \tag{51}$$

$$= \mathbf{I}_M - \frac{\mathbf{H}_{\mathrm{D}}(k)\mathbf{F}^H_{\mathrm{SNR,D}}(k)}{\mathbf{F}^H_{\mathrm{SNR,D}}(k)\mathbf{H}_{\mathrm{D}}(k)} \tag{52}$$

which illustrates its independence of the scaling constant $\zeta(k)$. It can be easily verified that the noise reference signals

$$\mathbf{U}_{\mathrm{D}}(k,m) = \mathbf{B}^H_{\mathrm{D}}(k)\mathbf{X}(k,m) = \left[\mathbf{I}_M - \frac{\mathbf{H}_{\mathrm{D}}(k)\mathbf{F}^H_{\mathrm{SNR,D}}(k)}{\mathbf{F}^H_{\mathrm{SNR,D}}(k)\mathbf{H}_{\mathrm{D}}(k)}\right]\mathbf{N}(k,m) \tag{53}$$

do not contain any desired speech signal components. In practice, however, there may be some signal leakage due to imperfect estimates of the noise PSD $\mathbf{\Phi_{NN}}(k)$ and the principal eigenvector $\mathbf{F}_{\mathrm{SNR,D}}(k)$ (see the experimental Section IV).

For obvious reasons, the blocking matrix (51) will be denoted as a generalized eigenvector blocking matrix (GEVBM) in the following. Note that the form of the transfer function ratios blocking matrix (TFRBM) by Gannot [8] is obtained if instead of $\mathbf{F}(k)_{\mathrm{SNR,D}}$ the first unit vector

$$\mathbf{E}(k) = (1, 0, \ldots, 0)^T \tag{54}$$

is used in (51). Accordingly, GSCs using either the GEVBM or the TFRBM, will be denoted GEVBM-GSC or TFRBM-GSC, respectively.

It is further worth noting that the number of GEVBM outputs is M in contrast to $M-1$ in [8]. Nevertheless, the rank of $\mathbf{B}^H(k)$ is $M-1$, i.e., one output is linearly dependent on the others. That means that it would be actually sufficient to use only $M-1$ noise references as input to the successive ANC. However, our experimental results showed that using all M blocking matrix output signals resulted in an improved SNR gain and comparable speech quality. The additional adaptive filter, which is required for the ANC in that case, leads to a slightly increased computational effort, but the convergence rate of the ANC is not noticeably affected [7].

One option for the design of the fixed beamformer is to employ a matched filter beamformer which applies a projection of the input signals $\mathbf{X}(k,m)$ onto the subspace spanned by $\bar{\mathbf{H}}_{\mathrm{D}}(k)$, which is accomplished by

$$\mathbf{F}^H_{\mathrm{FB,D}}(k) = \gamma_{\mathrm{D}}(k)\bar{\mathbf{H}}^H_{\mathrm{D}}(k). \tag{55}$$

The complex normalization factor $\gamma_{\mathrm{D}}(k)$ is chosen such that the resulting desired signal component at the output of the matched

filter beamformer is equal to the desired signal component at the first microphone $S_{1,\mathrm{D}}(k,m)$

$$\mathbf{F}_{\mathrm{FB,D}}^{H}(k)\mathbf{H}_{\mathrm{D}}(k,m) \stackrel{!}{=} H_{1,\mathrm{D}}(k,m) \tag{56}$$

thus avoiding speech distortions. From this requirement we obtain

$$\gamma_{\mathrm{D}}(k) = \frac{1}{||\bar{\mathbf{H}}_{\mathrm{D}}(k)||^2}. \tag{57}$$

Alternatively, a delay-and-sum beamformer (DSB) can be employed, which is known to provide some dereverberation to its input signals [24]. However, the use of a DSB requires the knowledge of the direction of arrival (DoA) of the desired signal and information about the microphone arrangement. In contrast to this, the matched filter beamformer can be realized without the aforementioned knowledge.

2) Additional Nonstationary Interferer: In this subsection, we assume the general signal model (15), including a nonstationary directional interferer. To deal with that case, we use the structure of the dual source transfer function generalized sidelobe canceler [16], whose idea is as follows. While the fixed beamformer is intended to suppress the nonstationary interferer, the goal of the noise canceling branch consisting of the BM and ANC is to remove the stationary noise from the microphone signals. For the construction of the FB and BM the TFRs $\bar{\mathbf{H}}_{\mathrm{D}}$ and $\bar{\mathbf{H}}_{\mathrm{I}}$ are required which are assumed to be time-invariant for the considered period of time.

In contrast to the least squares-based method according to [25], we propose here to compute the principal eigenvectors $\mathbf{F}_{\mathrm{SNR,D}}(k)$ and $\mathbf{F}_{\mathrm{SNR,I}}(k)$ of the GEVPs (33) and

$$\begin{aligned}\boldsymbol{\Phi}_{\mathbf{X}_{\mathrm{I}}\mathbf{X}_{\mathrm{I}}}(k,m)\mathbf{F}_{\mathrm{SNR,I}}(k,m) \\ = \lambda_{\mathrm{MAX,I}}(m,k)\boldsymbol{\Phi}_{\mathbf{NN}}(k)\mathbf{F}_{\mathrm{SNR,I}}(k,m)\end{aligned} \tag{58}$$

for the estimation of the TFRs. The PSD matrix

$$\boldsymbol{\Phi}_{\mathbf{X}_{\mathrm{I}}\mathbf{X}_{\mathrm{I}}}(k,m) := E[\mathbf{X}(k,m)\mathbf{X}^{H}(k,m)]|_{\mathbf{X}=\mathbf{S}_{\mathrm{I}}+\mathbf{N}} \tag{59}$$

is obtained in periods of time with interferer signal and stationary noise components only.

From the principal eigenvectors $\mathbf{F}_{\mathrm{SNR,D}}(k)$ and $\mathbf{F}_{\mathrm{SNR,I}}(k)$ the TFRs $\bar{\mathbf{H}}_{\mathrm{D}}(k)$ and $\bar{\mathbf{H}}_{\mathrm{I}}(k)$ can be computed according to (41)–(43). The dual-source blocking matrix can then be obtained by the projection into the orthogonal complement of the vector space spanned by $\bar{\mathbf{H}}_{\mathrm{D}}(k)$ and $\bar{\mathbf{H}}_{\mathrm{I}}(k)$. Using the Gram–Schmidt–Orthogonalization, we obtain the component

$$\mathbf{V}_{\mathrm{D}}(k) := \bar{\mathbf{H}}_{\mathrm{D}}(k) - \frac{\left[\bar{\mathbf{H}}_{\mathrm{I}}^{H}(k)\bar{\mathbf{H}}_{\mathrm{D}}(k)\right]\bar{\mathbf{H}}_{\mathrm{I}}(k)}{||\bar{\mathbf{H}}_{\mathrm{I}}(k)||^2} \tag{60}$$

of $\bar{\mathbf{H}}_{\mathrm{D}}(k)$ which is orthogonal to $\bar{\mathbf{H}}_{\mathrm{I}}(k)$. We thus propose to compute the blocking matrix by

$$\mathbf{B}_{\mathrm{DS}}^{H}(k) := \mathbf{I}_M - \frac{\bar{\mathbf{H}}_{\mathrm{I}}(k)\bar{\mathbf{H}}_{\mathrm{I}}^{H}(k)}{\bar{\mathbf{H}}_{\mathrm{I}}^{H}(k)\bar{\mathbf{H}}_{\mathrm{I}}(k)} - \frac{\mathbf{V}_{\mathrm{D}}(k)\mathbf{V}_{\mathrm{D}}^{H}(k)}{\mathbf{V}_{\mathrm{D}}^{H}(k)\mathbf{V}_{\mathrm{D}}(k)}. \tag{61}$$

We refer to this blocking matrix as the dual-source generalized eigenvector blocking matrix (DS-GEVBM), as it is based on estimations of generalized eigenvectors.

The fixed beamformer transfer functions are obtained by a projection into the subspace orthogonal to $\bar{\mathbf{H}}_{\mathrm{I}}(k)$ followed by a projection into the subspace spanned by $\bar{\mathbf{H}}_{\mathrm{D}}(k)$

$$\mathbf{F}_{\mathrm{FB,DS}}^{H}(k) := \gamma_{\mathrm{DS}}(k)\bar{\mathbf{H}}_{\mathrm{D}}^{H}(k)\left[\mathbf{I}_M - \frac{\bar{\mathbf{H}}_{\mathrm{I}}(k)\bar{\mathbf{H}}_{\mathrm{I}}^{H}(k)}{\bar{\mathbf{H}}_{\mathrm{I}}^{H}(k)\bar{\mathbf{H}}_{\mathrm{I}}(k)}\right]. \tag{62}$$

The complex normalization factor $\gamma_{\mathrm{DS}}(k)$ is chosen such that the resulting desired signal component at the output of the fixed beamformer is equal to the desired signal component at the first microphone $S_{1,\mathrm{D}}(k,m)$:

$$\mathbf{F}_{\mathrm{FB,DS}}^{H}(k)\mathbf{H}_{\mathrm{D}}(k,m) \stackrel{!}{=} H_{1,\mathrm{D}}(k,m) \tag{63}$$

thus avoiding speech distortions. From this requirement we obtain

$$\gamma_{\mathrm{DS}}(k) = \frac{||\bar{\mathbf{H}}_{\mathrm{I}}(k)||^2}{||\bar{\mathbf{H}}_{\mathrm{D}}(k)||^2||\bar{\mathbf{H}}_{\mathrm{I}}(k)||^2 - \left|\bar{\mathbf{H}}_{\mathrm{D}}^{H}(k)\bar{\mathbf{H}}_{\mathrm{I}}(k)\right|^2}. \tag{64}$$

Substituting the result for $\gamma_{\mathrm{DS}}(k)$ into (62) leads to the final expression for the filter beamformer [16]

$$\mathbf{F}_{\mathrm{FB,DS}}^{H}(k) = \frac{||\bar{\mathbf{H}}_{\mathrm{I}}(k)||^2\bar{\mathbf{H}}_{\mathrm{D}}^{H}(k) - \left[\bar{\mathbf{H}}_{\mathrm{D}}^{H}(k)\bar{\mathbf{H}}_{\mathrm{I}}(k)\right]\bar{\mathbf{H}}_{\mathrm{I}}^{H}(k)}{||\bar{\mathbf{H}}_{\mathrm{D}}(k)||^2||\bar{\mathbf{H}}_{\mathrm{I}}(k)||^2 - \left|\bar{\mathbf{H}}_{\mathrm{D}}^{H}(k)\bar{\mathbf{H}}_{\mathrm{I}}(k)\right|^2} \tag{65}$$

which will be referred to as the dual-source generalized eigenvector fixed beamformer (DS-GEVFB). As the computation of the DS-GEVBM in (61) and the DS-GEVFB in (65) is based on the determination of generalized eigenvectors, we will denote the new approach by DS-GEV-GSC. On the other hand, the approach described in [16] based on the least squares TFR estimation [8] will be referred to by DS-TFR-GSC and the corresponding fixed beamformer by DS-TFRFB.

III. Adaptive Eigenvector Tracking

The acoustic transfer function ratios (TFRs) are neither known nor constant in a practical setting. It is therefore important to have algorithms that estimate and track these ratios from the microphone signals. In the proposed system, the TFRs are computed from the principal eigenvectors of GEVPs. In the following, we concentrate on the solution of the GEVP (33) in its formulation as SEVP (34) as the other GEVP (58) is of the same structure and differs from (33) only by a different signal source. From the many algorithms for solving a GEVP or SEVP, we are only interested in those which achieve fast adaptation and low latency, as the beamforming system is to be used in a communication scenario.

We propose to apply the power iteration for the eigenvector computation [26]. The reason for the chosen method is the fast convergence behavior resulting from the rank one property of the matrix defining the SEVP (34). For the application of the power iteration the iteration matrix $\hat{\mathbf{A}}(k,m)$, which in our application is defined by

$$\hat{\mathbf{A}}(k,m) := \hat{\boldsymbol{\Phi}}_{\mathbf{NN}}^{-1}(k)\hat{\boldsymbol{\Phi}}_{\mathbf{XX}}(k,m) \tag{66}$$

has to be determined. The estimation of $\mathbf{\Phi_{NN}}(k)$ must be performed prior to the execution of the power iteration in periods when only stationary noise is present in the microphone signals. If we assume without loss of generality that these periods correspond to the first K_N blocks, an estimate of $\mathbf{\Phi_{NN}}(k)$ can be recursively computed by

$$\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k,0) := \sigma_u^2(k)\mathbf{I}_M \tag{67}$$

$$\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k,m) := \frac{1}{m+1}[\mathbf{X}(k,m)\mathbf{X}^H(k,m)]|_{\mathbf{X}=\mathbf{N}} + \frac{m}{m+1}\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k,m-1), \quad 1 \le m \le K_N \tag{68}$$

with a very small positive constant $\sigma_u^2(k)$. Basically, this operation corresponds to an average of the instantaneous estimates $\mathbf{N}(k,m)\mathbf{N}^H(k,m)$ if the initial value $\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k,0)$ is neglected. As at each estimation step in (68) a rank one update is performed, it is possible to apply the matrix inversion lemma (MIL) to recursively compute an estimate for the inverse $\hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(k,m)$. After the observation of K_N noise frames, we obtain an estimate

$$\hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(k) := \hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(k,K_N) \tag{69}$$

that may be used for the power iteration. The usage of the MIL avoids the computation of the inverse of $\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k,K_N)$ which is of order M^3. Thereby the computational expense is spread over all K_N frames where the complexity for each recursion is of the order of M^2. Note that in practice exponential weighting of the instantaneous estimates is also possible to track potential changes in the noise characteristics, but as stationary noise is assumed here, the equal weighting is more convenient.

Further, the PSD matrix $\mathbf{\Phi_{XX}}(k,m)$ is required. Since this expectation (31) is not available in practice, it has to be estimated from the captured microphone signals. An estimate can be computed by

$$\hat{\mathbf{\Phi}}_{\mathbf{XX}}(k,m) := \frac{1}{m-K_N}\sum_{\substack{m'=\\K_N+1}}^{m}\mathbf{X}(k,m')\mathbf{X}^H(k,m')|_{\mathbf{X}=\mathbf{S}_\mathrm{D}+\mathbf{N}} \tag{70}$$

from the observation of K_X frames when the microphone signals contain both speech and noise signal components. We assume here without loss of generality that these frames correspond to the frame indices $m = K_N+1,\ldots,K_N+K_X$ during which the TFRs are assumed to be time-invariant.

The justification for the time averaging operation according to (70) is that for $m \gg K_N$ the matrix $\hat{\mathbf{\Phi}}_{\mathbf{XX}}(k,m)$ assumes the same form as $\mathbf{\Phi_{XX}}(k,m)$ in (35) except for the constant $\Phi_{S_1S_1}(k,m)$

$$\hat{\mathbf{\Phi}}_{\mathbf{XX}}(k,m) \approx \left[\frac{1}{m-K_N}\sum_{m'=K_N+1}^{m}|S_1(k,m')|^2\right]\bar{\mathbf{H}}(k)\bar{\mathbf{H}}^H(k) + \mathbf{\Phi_{NN}}(k) \quad \text{for } m \gg K_N. \tag{71}$$

But obviously, this constant has no effect on the eigenvectors of the GEVP (33).

Instead of estimating $\hat{\mathbf{\Phi}}_{\mathbf{XX}}(k,m)$ according to (70) and then computing $\hat{\mathbf{A}}(k,m)$ from (66), the iteration matrix can be directly computed from the following recursion ($K_N+1 \le m \le K_N+K_X$)

$$\mathbf{b}(k,m) := \hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(k)\mathbf{X}(k,m) \tag{72}$$

$$\hat{\mathbf{A}}(k,m) = \frac{(m-K_N)-1}{m-K_N}\hat{\mathbf{A}}(k,m-1) + \frac{1}{m-K_N}\mathbf{b}(k,m)\mathbf{X}^H(k,m). \tag{73}$$

The estimate (73) can be used to apply a certain number of power iterations to a current eigenvector estimate at each frame m. The power iteration recursion for determining the principle eigenvector of $\hat{\mathbf{A}}(k,m)$ is given by

$$\hat{\mathbf{F}}_{\mathrm{SNR}}^{(j)}(k,m) = \frac{\hat{\mathbf{A}}(k,m)\hat{\mathbf{F}}_{\mathrm{SNR}}^{(j-1)}(k,m)}{\left\|\hat{\mathbf{A}}(k,m)\hat{\mathbf{F}}_{\mathrm{SNR}}^{(j-1)}(k,m)\right\|} \tag{74}$$

where j is the iteration index. The convergence speed of the power iteration is primarily determined by the ratio of the largest to the second largest eigenvalue of the iteration matrix. The larger this value, the higher is the convergence speed. As in our application, the rank of $\hat{\mathbf{A}}(k,m)-\mathbf{I}_M$ is close to one, see (66) and (71), a very high convergence speed is achieved. We observed that it is mostly sufficient to perform only one iteration per frame such that the eigenvector estimate $\hat{\mathbf{F}}_{\mathrm{SNR}}(k,m)$ for the mth frame ($K_N+1 \le m \le K_N+K_X$) is computed by

$$\hat{\mathbf{F}}_{\mathrm{SNR}}(k,m) := \frac{\hat{\mathbf{A}}(k,m)\hat{\mathbf{F}}_{\mathrm{SNR}}(k,m-1)}{\|\hat{\mathbf{A}}(k,m)\hat{\mathbf{F}}_{\mathrm{SNR}}(k,m-1)\|} \tag{75}$$

where, compared to (74), we have replaced the eigenvector estimate from the last iteration step by the eigenvector estimate from the last block. To save computational effort the norm can also be replaced by the absolute value of the real or imaginary part of the vector component of $\tilde{\mathbf{F}}_{\mathrm{SNR}}(k,m)$ with the largest absolute value.

In the next section, we will show by simulations that the convergence speed of the power iteration is higher than the convergence speed of the PSD estimations. That means that it is not worthwhile to use more than one iteration per frame.

It should be finally noted that because of the Hermitian symmetry of the DFT the estimation of the PSD matrices and application of the power iteration has only to be carried out for the first $L/2+1$ frequency bins.

IV. Performance Analysis

In this section, we will evaluate the performance of the presented GEVBM-GSC and compare it with GSCs using different realizations of blocking matrices proposed in the literature. The analysis is performed for simulated reverberant enclosures with varying reverberation time.

For the simulations we used a linear array of $M = 5$ microphones with an inter-element distance of 0.04 m which was placed in a simulated reverberant enclosure of the size (6 m) × (5 m) × (3 m). The positions of the desired speech, noise, and interference sources within the enclosure are depicted in Fig. 3. The angles and distances were chosen to $\theta_\mathrm{D} = -20^\circ$ and

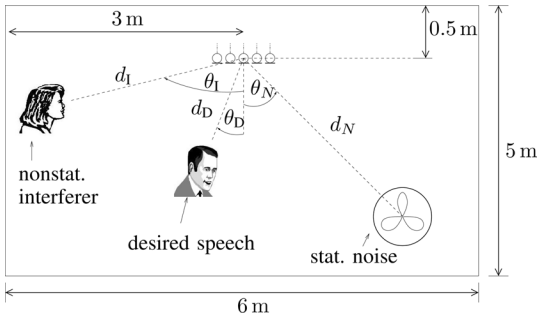

Fig. 3. Simulation setup.

$d_{\mathrm{D}} = 0.85$ m for the desired speaker, $\theta_{\mathrm{I}} = -76°$ and $d_{\mathrm{I}} = 2.07$ m for the nonstationary directional interferer, and $\theta_N = 45°$ and $d_N = 2.12$ m for the stationary directional noise source, respectively. All sources as well as all microphones were placed at a height of 1.5 m. Note, that the nonstationary interference was used only for simulations concerning the DS-GEV-GSC.

Ten sentences of the TIMIT database, each of a length of about 4 s, were used as desired speech and interference source signals. The directional stationary noise source signal was a recording of low-pass fan noise. The directional speech, noise, and interference signal components at the microphones were created using the image method [27] by Allen and Berkley and superposed with an SNR of 5 dB and a signal-to-interference ratio (SIR) of 0 dB. In addition, white Gaussian noise was added to each microphone with an SNR of 35 dB in order to take noncoherent noise into consideration. The sampling rate $1/T$ was 12 kHz.

A. Tracking of Principal Eigenvector

In this experiment we want to demonstrate the tracking capability of the proposed eigenvector determination.

For the computation of the STDFTs the microphone signals were windowed by a rectangular window with a length of $L = 512$ samples and an overlap of $L - B = L/2$. The noise PSD matrix was estimated according to (68) during a period of 10 s. Hereafter, in periods when the speech and noise components were both present, the iteration matrix was computed according to (73).

The time variability of the PSD estimates obviously leads to time variability in the optimal eigenvectors $\mathbf{F}_{\mathrm{SNR}}(k,m)$ of the corresponding estimated SEVP matrices (73). The relative temporal changes of the principal eigenvectors are measured by

$$\delta_{\mathrm{EV}}(k,m) := 10\log_{10}\left[\frac{||\bar{\mathbf{F}}_{\mathrm{SNR}}(k,m) - \bar{\mathbf{F}}_{\mathrm{SNR}}(k,m-1)||}{||\bar{\mathbf{F}}_{\mathrm{SNR}}(k,m-1)||}\right] \tag{76}$$

where

$$\bar{\mathbf{F}}_{\mathrm{SNR}}(k,m) := \frac{\mathbf{F}_{\mathrm{SNR}}(k,m)}{F_{\mathrm{SNR},1}(k,m)} \tag{77}$$

is a "true" principal eigenvector of the estimate $\hat{\mathbf{A}}(k,m)$ in (73) normalized by its first component which was computed by a MATLAB batch routine. The purpose of the normalization is to ensure uniqueness of the true eigenvectors of $\hat{\mathbf{A}}(k,m)$ to be able to measure changes. Fig. 4(a) and (b) illustrates the relative changes $\delta_{\mathrm{EV}}(k,m)$ measured on a noisy speech signal for two different reverberation time values. The gray scale indicates the value of $\delta_{\mathrm{EV}}(k,m)$ with light grey corresponding to large and dark grey to small changes. It can be seen that for both scenarios the changes gradually decrease with increasing estimation block index m. Stronger changes typically correlate with the presence of high power at certain time–frequency points. Further, the extent of the changes increases with the reverberation time. This is caused by the fact that the MTFA (15) for the STDFTs of the microphone signals becomes worse with an increasing length of the speaker-sensor impulse responses $\mathbf{h}(l)$ resulting in poorer PSD estimates.

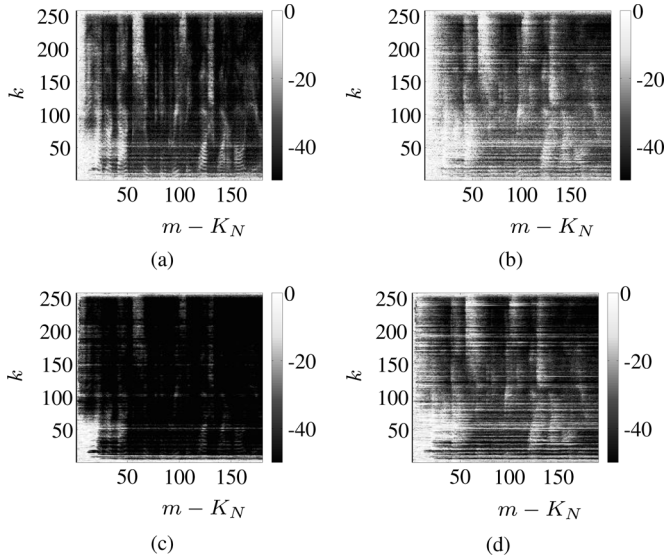

Fig. 4. Relative changes of optimal eigenvectors $\delta_{\mathrm{EV}}(k,m)$ during the estimation of PSD matrices for reverberation time $T_{60} = 100$ ms in (a) and $T_{60} = 800$ ms in (b) and corresponding relative error $e(k,m)$ between computed and exact eigenvectors for $T_{60} = 100$ ms in (c) and $T_{60} = 800$ ms in (d). The gray scale indicates the value of $\delta_{\mathrm{EV}}(k,m)$ and $e(k,m)$, respectively, ranging from 0 dB to -50 dB.

To show how fast these changes are tracked, one power iteration step according to (75) was performed for each frame simultaneously to the estimation of the iteration matrix $\hat{\mathbf{A}}(k,m)$. Then the relative error between the estimated and optimal eigenvector of the current SEVP

$$e(k,m) := 10\log_{10}\left[\frac{||\bar{\hat{\mathbf{F}}}_{\mathrm{SNR}}(k,m) - \bar{\mathbf{F}}_{\mathrm{SNR}}(k,m)||}{||\bar{\mathbf{F}}_{\mathrm{SNR}}(k,m)||}\right] \tag{78}$$

was examined, which is depicted in Fig. 4(c) and (d). Here we defined

$$\bar{\hat{\mathbf{F}}}_{\mathrm{SNR}}(k,m) := \frac{\hat{\mathbf{F}}_{\mathrm{SNR}}(k,m)}{\hat{F}_{\mathrm{SNR},1}(k,m)} \tag{79}$$

as the principal eigenvector estimate obtained from (75) normalized by its first component.

It should be stressed that for very low frequencies below 100 Hz and high frequencies near the Nyquist frequency with minor speech components the eigenvector $\bar{\mathbf{F}}_{\mathrm{SNR}}(k,m)$ of the GEVP (33) is not well defined as the PSD $\mathbf{\Phi}_{\mathbf{XX}}(k,m)$ reduces to $\mathbf{\Phi}_{\mathbf{NN}}(k,m)$. Errors at those frequencies are therefore not

significant and should be excluded from the following interpretation of the results.

For the case of the lower reverberation time $T_{60} = 100$ ms, the error $e(k,m)$ quickly decreases to values of about -30 dB for most frequency bins k already after about 50 blocks, which corresponds to about a second. It is evident that for $m \gg K_N$, the error is significantly smaller than the changes of the corresponding optimal eigenvector. This confirms the redundancy of the application of more than one power iteration step per frame. However, with increasing reverberation time the estimation errors in the PSD matrices worsen the ratio of the second largest to the principal eigenvalue of the estimated iteration matrix (66), which leads to a slower convergence speed of the power iteration. Fig. 4(d) shows that the resulting errors are higher than for $T_{60} = 100$ ms [Fig. 4(c)]. However, the errors are still smaller than the changes in the eigenvectors. That means that the convergence speed of the proposed eigenvector estimation method is mainly influenced by the convergence speed of the PSD estimations.

B. Performance of GEVBM-GSC

Now the performance of the proposed GEVBM-GSC shall be analyzed and compared with GSCs using different approaches for blocking matrices. First, the scenario with the stationary noise disturbing the speech signal is considered. A DSB was used as fixed beamformer. The multichannel noise cancellation in the ANC was implemented using the normalized least-mean-square (NLMS) method with a filter length of 1024 taps.

The following four different blocking matrices were compared: the blocking matrix by Griffiths and Jim (GJBM) [4], the transfer function ratios blocking matrix (TFRBM) by Gannot *et al.* [8], the adaptive blocking matrix (ABM) by Hoshuyama *et al.* [5], and the generalized eigenvector blocking matrix (GEVBM) proposed here. The implementation of these blocking matrices will be addressed in the following.

For the blocking matrix by Griffith and Jim (GJBM) the microphone signals were aligned the same way as in the DSB. The noise reference signals were computed by subtracting from one microphone signal the mean of the $M-1$ other aligned microphone signals. We extend the abbreviation GJBM by the addition (opt) to indicate that we assume a perfect knowledge of the direction of arrival.

The estimation of the TFRs in the TFRBM was conducted according to [8]: the microphone signals were windowed by a rectangular window with a length of $L = 512$ samples with an overlap of $L-B = L/2$ taps between successive windows. The estimation time corresponded to the length of each TIMIT sentence. The estimated transfer functions were then transformed to the time domain. As the resulting impulse responses were assumed to be noncausal and of finite length, they were cut off to the interval $[-127, 128]$.

Furthermore, the adaptive blocking matrix (ABM) [5] was incorporated into the simulations. It was realized in the frequency domain without constraints on the filter coefficients for improved robustness using filter lengths of 256 samples. Its performance was analyzed for two different conditions, namely for adaptation either in the presence or absence of noise. In the case of absence of noise [denoted by the addition (opt)] a perfect BM is achieved in the sense that the noise references are enforced to be orthogonal to speech by the adaptation process. While this is the preferred setup for adaptation [7], as the performance degrades if the noise is also present during adaptation (see the following results), this setup is not very practical and may be difficult to ensure in practice.

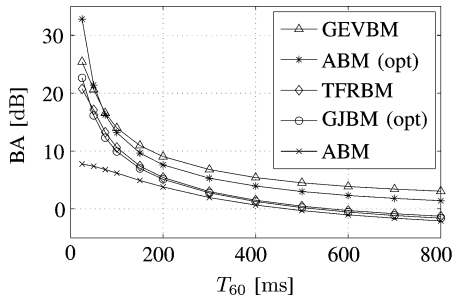

Fig. 5. Blocking ability for directional stationary noise.

Finally, for the GEVBM, the estimation of the principal eigenvector was performed as described in Section III. The resulting eigenvector was used to form the transfer functions of the GEVBM for each frequency bin as shown in (48). Finally, the FIR filter coefficients of the blocking matrix were determined in the same way as for the TFRBM. The filtering in the GEVBM and TFRBM was realized in the frequency domain using the overlap save method.

For the comparison of the blocking matrices all components of the GSC were analyzed in steady state. That was supposed to be achieved by choosing a high number of iterations for the ANC and the ABM. Different performance measures concerning signal leakage, noise reduction and speech quality are presented in the following.

First of all, the blocking ability (BA) of the different blocking matrices was computed as a function of the reverberation time T_{60}. It is defined as the difference between the input SNR, $\mathrm{SNR}^{(\mathrm{X})}$ [dB], and the SNR of the noise references, $\mathrm{SNR}^{(\mathrm{U})}$ [dB]

$$\mathrm{BA}\ [\mathrm{dB}] := \mathrm{SNR}^{(\mathrm{X})}\ [\mathrm{dB}] - \mathrm{SNR}^{(\mathrm{U})}\ [\mathrm{dB}] \tag{80}$$

all in logarithmic scale, obtained by the different blocking matrices. This is a commonly used measure of signal leakage [16], [17], [8]. Low values for BA indicate a large amount of signal leakage into the noise references which obviously results in speech distortion caused by the ANC. The results for BA are shown in Fig. 5. It can be seen that for reverberation time values greater than 75 ms the GEVBM provides the best blocking ability. Only for very low reverberation time values below 75 ms the ABM in the case of optimal adaptation provides better results than the GEVBM while its blocking ability is about 1 dB smaller compared to the GEVBM for greater reverberation time values. The TFRBM and GJBM (opt) perform similarly. For higher reverberation time values, they achieve about 3 dB lower blocking ability than the GEVBM. Finally, it can be observed from the noticeably lower values for BA of the ABM that its performance severely degrades if adaptation is carried out in the presence of noise.

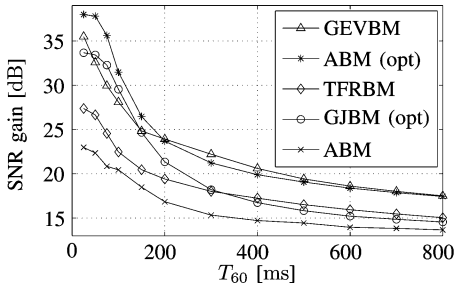

Fig. 6. SNR gain for directional stationary noise.

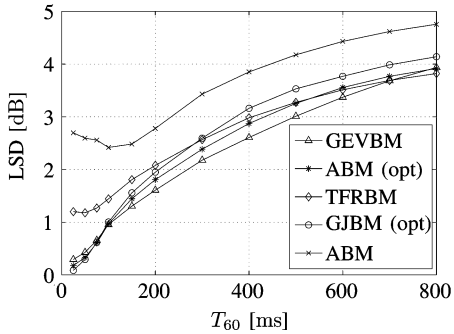

Fig. 7. Log spectral distance (LSD) for directional stationary noise.

Second, the SNR gain from the input to the output of the whole GSC structure was measured, which is defined by

$$\text{SNR gain [dB]} := \text{SNR}^{(\text{X})}\text{ [dB]} - \text{SNR}^{(\text{Y})}\text{ [dB]} \tag{81}$$

where $\text{SNR}^{(\text{Y})}$ [dB] denotes the SNR of the GSC output signal. As can be seen in Fig. 6, for small values of T_{60} the GJBM (opt) achieves very good results. Only the ABM under optimal, however unrealistic, adaptation conditions is able to achieve a larger SNR gain. These results are not surprising since for such low reverberation time the assumption of a simple direct path propagation in the GJBM (opt) is almost met. Remaining errors in the GJBM (opt) can be caused by the far-field assumption for the sound propagation and the approximation of the alignment filters by FIR filters. While the SNR gains of the GJBM (opt) significantly decrease for increased reverberation time, the decrease proceeds much slower with ABM (opt) and GEVBM such that about 3-dB SNR gain improvement compared to the GJBM (opt) is obtained for high reverberation time. Furthermore, it is important to mention that the GJBM (opt) would additionally lose in performance if the estimation of the DoA were not perfect.

Unfortunately, the SNR gains achieved with the TFRBM are considerably smaller for low reverberation times compared to the other approaches. Detailed investigations showed that these stem from less noise reduction in the lower frequencies below 500 Hz. For large values of T_{60}, the SNR improvements are only slightly higher than those obtained with the GJBM (opt). As expected, the ABM with adaptation in the presence of noise gives poor results for SNR gains for all considered reverberation time values.

To assess speech signal distortion we evaluated the log spectral distance (LSD) [28] between the speech signal component of the beamformer output $y_{\text{GSC},s}(l)$ and speech signal component of the delayed FB output $y_{\text{FB},s}(l)$. The FB output is chosen as reference because it is an estimate for the speech signal part right before signal cancellation introduced by the ANC can occur. That means that any log spectral distortion is caused by signal leakage into the noise references as the DSB used as fixed beamformer introduces only negligible distortion. The LSD was computed for a speech signal duration of K_S blocks by

$$\text{LSD [dB]} := \left(\frac{1}{K_S} \sum_{m=1}^{K_S} \frac{2}{L} \sum_{k=0}^{\frac{L}{2}-1} |\Delta_Y(k,m)|^2 \right)^{\frac{1}{2}} \tag{82}$$

where

$$\Delta_Y(k,m) := \mathcal{L}(Y_{\text{GSC},s}(k,m)) - \mathcal{L}(Y_{\text{FB},s}(k,m)) \tag{83}$$

denotes the difference between the log spectra of $y_{\text{GSC},s}(l)$ and $y_{\text{FB},s}(l)$ which were confined to about 50 dB dynamic range [29]

$$\mathcal{L}(Y_s(k,m)) := \max\{20\log_{10}|Y_s(k,m)|, P_{Y_{\text{FB},s},\min}\} \tag{84}$$

$$P_{Y_{\text{FB},s},\min} := \max_{k,m}\{20\log_{10}|Y_{\text{FB},s}(k,m)|\} - 50. \tag{85}$$

The results, which were averaged over the ten examined speech utterances, are displayed in Fig. 7.

It can be seen that the GEVBM, the ABM (opt), and the GJBM (opt) introduce similarly minor speech distortion for very low reverberation time values below $T_{60} = 100$ ms. However, as the reverberation time increases the distortion occurring with the GJBM (opt) is consistently higher than of the GEVBM and the ABM (opt). These differences arise from the fact that the direct path propagation assumption of the GJBM (opt) is violated for high reverberation time whereas the GEVBM and the ABM (opt) explicitely consider reverberation. We further observed from detailed investigations that for the TFRBM lower frequencies, especially between 100 and 500 Hz, were highly amplified for low reverberation time. That is partly the reason why for low reverberation time values below $T_{60} = 200$ ms the LSD values for the TFRBM are higher than those of the GEVBM. However, for higher reverberation time values the LSD values of the TFRBM approach those of the GEVBM. As expected, the ABM with adaptation under presence of noise leads to a spectral distance which is considerably larger than those of all other compared blocking matrices.

Finally, the speech distortion in the GSC output signal is measured by a perceptual speech quality measure (PSM) [30] in Fig. 8. PSM has been shown to give objective perceptual quality evaluation results comparable with the well-known PESQ measure. Here, the reference was the clean speech output signal of the DSB. A value of one for the PSM measure indicates absence of any measured deviation from the reference. The PSM values for the novel GEVBM are among the best for all reverberation time values. The TFRBM delivers somewhat inferior results, which can be explained by the fact that the TFRBM boosts low-frequency signal components. It comes as no surprise that the worst PSM results were expectedly achieved by the ABM

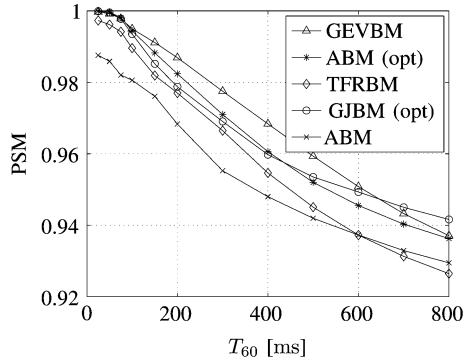

Fig. 8. Perceptual speech quality measure (PSM) of the GSC output signals compared to the DSB output signals using different blocking matrices in the case of directional stationary noise.

with adaptation under presence of noise. The displayed PSM results corresponded well to our informal listening tests.

C. Computational Complexity

The computational complexity of the adaptive eigenvector estimation is of order M^2 for each frequency bin due to the smoothing of the PSD matrices and application of the power iteration. Furthermore, the computation of the noise references requires $O(M^2)$ arithmetical operations as the GEVBM is dense for each frequency bin. In contrast to this, the TFR estimation according to [25] and computation of the TFRBM output [8] both have a complexity which is of the order of M since the TFRBM is sparse [see (54)]. Moreover, as a result of M noise references rather than $M-1$, an additional adaptive filter for the ANC is used for the realization of the GEVBM-GSC compared to the TFRBM-GSC. Still, the computational effort of the GEVBM-GSC is quite amenable to a real-time implementation. Our C/C++ implementation comprised the GEVBM-GSC with a DSB as FB, a VAD, DoA estimation by eigenvalue decomposition [31] and the multichannel audio input/output management. The computational effort for the whole system running on an Intel Quad-Core Xeon E5345/2.33-GHz processor resulted in a real-time factor of 0.3. This result confirms the feasibility of the GEVBM-GSC. Audio and video files of a demonstration of the above-mentioned beamforming system in a laboratory setup are available at our webpage [32].

D. Performance of DS-GEV-GSC

Now the performance of the DS-GEV-GSC will be analyzed by simulations and compared to that of the DS-TFR-GSC. Again, the scenario according to Fig. 3 is considered, which now includes the nonstationary interferer. Five sentences each served as source signals for the desired speech and nonstationary interferer. The presented simulation results are averages of the results for the 5×5 signal combinations for each reverberation time. While the stationary noise was present at all times, the utterances of the desired speaker and the interferer did not overlap in time. We assumed here to know the periods when each of the nonstationary signals was active. With that knowledge the TFRs $\bar{\mathbf{H}}_{\mathrm{D}}(k)$ and $\bar{\mathbf{H}}_{\mathrm{I}}(k)$ were estimated as described in Subsection II-D2 using the PSD estimations (59).

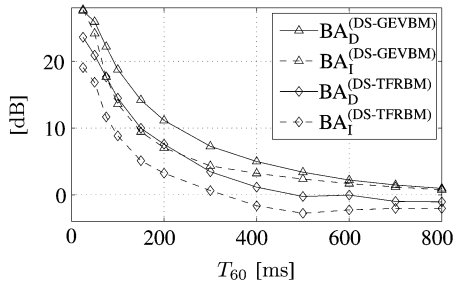

Fig. 9. Blocking ability of dual-source blocking matrices for directional stationary noise and directional nonstationary interferer.

In contrast to the simulations of the GSC in the previous subsection, we used here rectangular windows with a length of $L = 1024$ samples and an overlap of $L/2$ for the STDFT computation. The filter lengths for the FB and the BM were set to 512 taps. The purpose of the increased filter length of the FB is to improve the suppression of the nonstationary interferer for high reverberation times.

First, we evaluated the blocking ability of the DS-GEVBM given by (61) and the DS-TFRBM defined in [16]. The blocking ability was computed with respect to the desired source signal and the nonstationary interfering signal. This is denoted by BA_{D} and BA_{I}, respectively. Fig. 9 clearly shows that the DS-GEVBM gives a significant improvement in blocking ability compared to the DS-TFRBM for the whole range of considered reverberation times.

Furthermore, the SIR gain of the matched filter beamformer with respect to the nonstationary interferer was analyzed. It is defined by

$$\text{SIR gain}^{(\mathrm{FB})}\,[\mathrm{dB}] := \text{SIR}^{(\mathrm{FB})}\,[\mathrm{dB}] - \text{SIR}^{(\mathrm{IN})}\,[\mathrm{dB}] \qquad (86)$$

where $\text{SIR}^{(\mathrm{FB})}$ and $\text{SIR}^{(\mathrm{IN})}$ are the SIRs of the FB output and the microphone signals, respectively. The SIR and SNR gains $\text{SIR gain}^{(\mathrm{GSC})}$ and $\text{SNR gain}^{(\mathrm{GSC})}$ of the whole dual-source GSCs are defined accordingly and displayed in Fig. 10. For all reverberation times it can again be observed that the DS-GEV-GSC is superior to the DS-TFR-GSC concerning the attenuation of undesired stationary and directional nonstationary noises. Detailed investigations showed that the remarkably low values for the SIR gain of the TFR matched filter beamformer are partly caused by errors in the TFR estimation, especially in frequencies below 500 Hz. These errors result in an amplification of the interferer signal compared to the desired signal.

Further, we analyzed the speech distortion introduced by the DS-GEV-GSC and the DS-TFR-GSC. For this purpose, we measured the log spectral distance between the clean reverberated desired signal component $s_{1,\mathrm{D}}(l)$ at the first microphone and each desired signal component of the corresponding dual-source fixed beamformer (DS-FB) output $y_{\mathrm{FB},s}(l)$ and the corresponding desired signal component of the dual-source GSC output $y_{\mathrm{GSC},s}(l)$, respectively. The results are depicted in Fig. 11. Considerably smaller values for the LSD, for the whole range of considered reverberation times, of the DS-GEVFB compared to the DS-TFRFB as well as of the DS-GEV-GSC compared to the DS-TFR-GSC indicate less speech distortion.

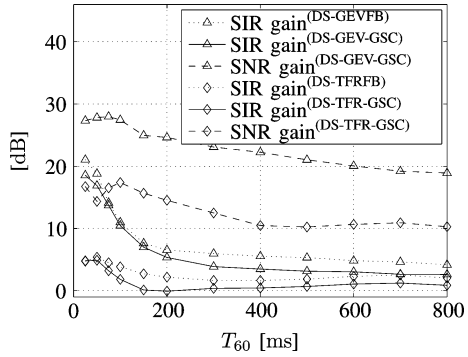

Fig. 10. SIR and SNR gain of fixed beamformer and whole dual-source GSCs.

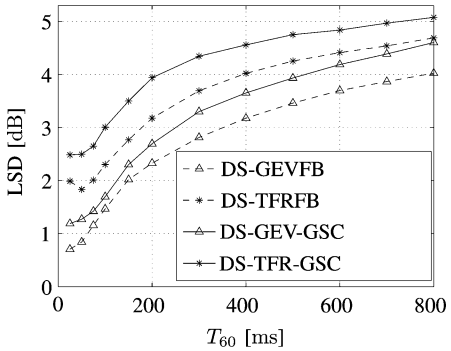

Fig. 11. LSD between the clean reverberated desired signal component $s_{1,\mathrm{D}}(l)$ at the 1. microphone and each the desired signal component of the corresponding dual-source fixed beamformer output $y_{\mathrm{FB},s}(l)$ and the corresponding desired signal component of the dual-source GSC output $y_{\mathrm{GSC},s}(l)$.

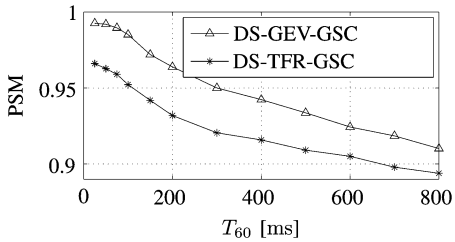

Fig. 12. Perceptual speech quality measure (PSM) for comparison of the desired speech signal component at the dual-source GSC output $y_{\mathrm{GSC},s}(l)$ and the clean reverberated desired speech signal component $s_{1,\mathrm{D}}(l)$ at the 1. microphone.

The latter result is confirmed by the PSM results depicted in Fig. 12.

Finally, we want to point out that in practice only a suboptimal performance of all considered algorithms is attainable due to imperfect estimation of the presence or absence of the desired speech signal, the interferer, or the noise.

V. Conclusion

In this paper, we have presented a new blocking matrix and fixed beamformer design for a generalized sidelobe canceler to be used in a reverberant hands-free communication environment. The design requires an estimate of the ratios of the transfer functions from a desired speaker to the sensors in the presence of stationary noise. The computation of the ratios is based on an adaptive determination of the principal eigenvector of a generalized eigenvalue problem. Simulation results for the proposed eigenvector tracking algorithm show that the convergence speed is mainly influenced by the estimation of the involved PSD matrices. Especially for lower reverberation time a very high convergence speed is achieved.

Furthermore, simulation results of the proposed GEV blocking matrix give evidence for the improved blocking ability compared to other blocking matrices. As a consequence, less speech distortion is introduced by the application of a GEVBM-GSC. Additionally, a higher SNR gain is achieved. The performance of the GEVBM is similar to that of the adaptive blocking matrix by Hoshuyama *et al.* if adaptation of the ABM is carried out in the absence of noise, a requirement which is difficult to meet in a practical setting. If adaptation of the ABM is carried out in the presence of stationary noise, the performance of the ABM severely degrades and falls well below that of the GEVBM.

Finally, we introduced the DS-GEV-GSC, which is based on the DS-TFR-GSC [16]. The differences to the DS-TFR-GSC are, first, the new generalized eigenvector-based technique employed for the TFR estimation needed for the blocking matrix and fixed beamformer computation and second, the new way of computing the blocking matrix. Simulation results show that the DS-GEV-GSC is superior to the DS-TFR-GSC in terms of SNR gain and resulting speech quality. However, these improvements can only be achieved with a higher computational effort which is increased by about one factor of M compared to the TFRBM.

References

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

[2] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

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

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

[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] W. Herbordt and W. Kellermann, "Efficient frequency-domain realization of robust generalized, sidelobe cancellers," in *Proc. IEEE 4rth Workshop Multimedia Signal Process.*, 2001, pp. 377–382.

[7] W. Herbordt and W. Kellermann, "Analysis of blocking matrices for generalized sidelobe cancellers for non-stationary broadband signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP'02)*, 2002, vol. 4, pp. IV-4187–.

[8] 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.

[9] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[10] 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.

[11] S. Doclo and M. Moonen, "Multimicrophone noise reduction using recursive GSVD-based optimal filtering with ANC postprocessing stage," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 53–69, Jan. 2005.

[12] J. Chen, J. Benesty, and Y. Huang, "A minimum distortion noise reduction algorithm with multiple microphones," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 481–493, Mar. 2008.

[13] B. Cornelis, M. Moonen, and J. Wouters, "Comparison of frequency domain noise reduction strategies based on multichannel wiener filtering and spatial prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP'09*, Apr. 2009, pp. 129–132.

[14] E. Warsitz and M. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.
[15] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP'07*, Apr. 2007, vol. 1, pp. I-41–I-44.
[16] G. Reuven, S. Gannot, and I. Cohen, "Dual-source transfer-function generalized sidelobe canceller," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 711–727, May 2008.
[17] G. Reuven, S. Gannot, and I. Cohen, "Performance analysis of dual source transfer-function generalized sidelobe canceller," *Speech Commun.*, vol. 49, no. 7–8, pp. 602–622, 2007.
[18] J. Schmalenstroeer and R. Haeb-Umbach, "Joint speaker segmentation, localization and identification for streaming audio," in *Proc. Interspeech'07*, 2007, pp. 570–573.
[19] G. Lathoud and J.-M. Odobez, "Short-term spatio-temporal clustering applied to multiple moving speakers," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1696–1710, Jul. 2007.
[20] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.
[21] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Appl. Signal Process.*, vol. 2006, 2006.
[22] E. Warsitz, A. Krueger, and R. Haeb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP'08*, Apr. 2008, pp. 73–76.
[23] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.
[24] N. D. Gaubitch and P. A. Naylor, "Analysis of the dereverberation performance of microphone arrays," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC'05)*, 2005, pp. 121–125.
[25] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 2055–2063, Aug. 1996.
[26] J. Karhunen, "Adaptive algorithms for estimating eigenvectors of correlation type matrices," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP'84*, Mar. 1984, vol. 9, pp. 592–595.
[27] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J.Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[28] J. A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 5, pp. 380–391, Oct. 1976.
[29] E. A. P. Habets, "Single-and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Univ. Eindhoven, Eindhoven, The Netherlands, Jun. 25, 2007.
[30] R. Huber, "Objective assessment of audio quality using an auditory processing model," Ph.D. dissertation, Univ. of Oldenburg, Oldenburg, Germany, 2003.
[31] E. Warsitz, "Mehrkanalige Sprachsignalverbesserung durch adaptive Lösung eines Eigenwertproblems im Frequenzbereich," Ph.D. dissertation, Univ. of Paderborn, Paderborn, Germany, 2009.
[32] [Online]. Available: http://nt.uni-paderborn.de/index.php?id=mic_array_proc_gsc&L=1_gsc_video

Alexander Krueger (S'08) was born in 1981. He received the Dipl.-Math. degree (*summa cum laude*) in technomathematics from the University of Paderborn, Paderborn, Germany, in 2007.

Since 2007, he has been a Research Staff Member with the Department of Communications Engineering, University of Paderborn. His research interests include statistical speech signal processing and recognition.

Ernst Warsitz (M'04) received the Dipl.-Ing. and Ph.D. degrees in electrical engineering from the University of Paderborn, Paderborn, Germany, in 2000 and 2008, respectively.

From 2001 to 2007, he joined the Department of Communications Engineering, University of Paderborn, as a Research Staff Member. There, he was involved in automatic speech recognition, microphone array beamforming, and acoustic scene analysis. He is currently with Hella KGaA Hueck & Co. in Lippstadt, Germany, engaged in advanced engineering for driver assistance systems. His research interests are in multichannel statistical signal processing in general.

Reinhold Haeb-Umbach (M'89–SM'09) received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from RWTH Aachen University, Aachen, Germany, in 1983 and 1988, respectively.

From 1988 to 1989, he was a Postdoctoral Fellow at the IBM Almaden Research Center, San Jose, CA, conducting research on coding and signal processing for recording channels. From 1990 to 2001, he was with Philips Research working on various aspects of automatic speech recognition, such as acoustic modeling, efficient search strategies, and mapping of algorithms on low-resource hardware. Since 2001, he has been a Professor in Communications Engineering at the University of Paderborn, Paderborn, Germany. His main research interests are in statistical speech signal processing and recognition and in signal processing for communications. He has published more than 100 papers in peer-reviewed journals and conferences.