Comparison of Generalized Sidelobe Canceller Structures Incorporating External Microphones for Joint Noise and Interferer Reduction

Wiebke Middelberg, Simon Doclo

University of Oldenburg, Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Oldenburg, Germany
Email: {wiebke.middelberg, simon.doclo}@uni-oldenburg.de
Web: www.sigproc.uni-oldenburg.de

Abstract

In this paper, we compare two extended generalized sidelobe canceller (GSC) structures, which exploit external microphones in conjunction with a local microphone array to improve the noise and interferer reduction. As a baseline algorithm we consider a local GSC using only the local microphones, for which the relative transfer function (RTF) vector of the target speaker is known. To incorporate the external microphones in a minimum power distortionless response beamformer, the RTF vector of the target speaker needs to be estimated. Since the estimation accuracy of this RTF vector depends on the signal-to-interferer ratio, the GSC with external speech references (GSC-ESR) pre-processes the external microphone signals to reduce the interferer. In a simplified extended structure, namely the GSC with external references (GSC-ER) no such pre-filtering operation is performed. Simulation results show that the GSC-ESR structure yields the best results in terms of noise and interferer reduction, especially in adverse conditions.

1 Introduction

In assistive listening devices such as hearing aids or cochlear implants, speech quality and speech intelligibility are often degraded by background noise and competing speakers. Hence, single- and multi-microphone speech enhancement algorithms are used, aiming at joint noise and interferer reduction [1–3]. Popular multi-microphone speech enhancement algorithms are based on minimum variance distortionless response (MVDR) or minimum power distortionless response (MPDR) beamforming [4, 5], which can be implemented using the generalized sidelobe canceller (GSC) structure [6–8]. To implement these beamformers, the direction-of-arrival (DoA) or more general the relative transfer function (RTF) vector of the target speaker between the microphones is required. Although several RTF vector estimation methods have been proposed [7–11], these methods yield biased estimates of the target RTF vector when one or more competing speakers are present. In hearing aid applications, it is hence commonly assumed that the DoA or the RTF vector of the target speaker between the head-mounted microphones is known (e.g., frontal direction).

It has been shown that external microphones (eMics) enable to improve the speech enhancement performance compared to only using the head-mounted microphones [12–18]. This can be explained by the fact that the eMics allow for a more diverse sampling of the sound field than the head-mounted microphones, which will be referred to as the local microphone array (LMA) in this paper (see Fig. 1). In [16], a promising structure was proposed which allows to incorporate eMics into a local GSC (L-GSC) using only the LMA, assuming that the local RTF vector of the target speaker is known. In this paper, we consider the structure proposed in [16] and reformulate it in terms of an MPDR beamformer instead of an MVDR beamformer such that not only noise but also interferer can be cancelled. This structure, referred to as the GSC with external speech references (GSC-ESR), pre-filters the noise-and-interferer references of the L-GSC to reduce the interferer in the eMic signals. Using these pre-processed signals instead of the unprocessed eMic signals allows to improve the estimation accuracy of the external RTF vector of the target speaker, which is subsequently used in an MPDR beamformer for joint noise and interferer reduction. To assess the benefit of the pre-filtering operation, also the structure using the unprocessed eMic signals is considered, which is referred to as the GSC with external references (GSC-ER).

In the experimental evaluation, we compare the performance of both extended GSC structures and the L-GSC in terms of noise and interferer reduction using reverberant recordings. The results show that the pre-filtering operation in the GSC-ESR indeed increases the performance in terms of interferer reduction compared to the GSC-ER. In addition, we investigate the sensitivity against a mismatch of the - assumed to be known - local RTF vector. When using an approximate anechoic local RTF vector, the results show that the pre-filtering operation induces target speech cancellation, which decreases the performance of the GSC-ESR in comparison to the GSC-ER at high signal-to-interferer ratios (SIRs).

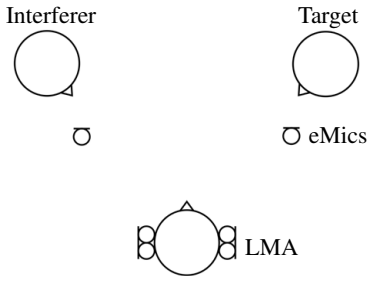

Figure 1: Top view of the considered acoustic scenario and microphone configuration with M_a = 4 head-mounted microphones of the LMA and M_e = 2 external microphones (eMics).

2 Signal Model and Notation

We consider an acoustic scenario with one target speaker, one interferer and background noise (see Fig. 1). The LMA is equipped with M_a head-mounted microphones and M_e additional eMics are present, resulting in a total of $M = M_\mathrm{a} + M_\mathrm{e}$ microphones. In the short time Fourier transform (STFT) domain the m-th microphone signal is given by

$$Y_m(k,l) = X_m(k,l) + I_m(k,l) + N_m(k,l), m \in \{1,\dots,M\}, \quad (1)$$

where k and l denote the frequency bin index and the frame index, respectively, $X_m(k,l)$ denotes the target speech component, $I_m(k,l)$ denotes the interferer component and $N_m(k,l)$ denotes the noise component in the m-th microphone. In the following, we neglect the indices k and l for conciseness. For the stacked signal vector $\mathbf{y}$, we distinguish between the M_a local microphones and the M_e eMics, i.e.

$$\mathbf{y} = [Y_{\mathrm{a},1}, Y_{\mathrm{a},2}, \dots, Y_{\mathrm{a},M_\mathrm{a}}, Y_{\mathrm{e},1}, \dots, Y_{\mathrm{e},M_\mathrm{e}}]^T = [\mathbf{y}_\mathrm{a}^T, \mathbf{y}_\mathrm{e}^T]^T, \quad (2)$$

where $\{\cdot\}^T$ denotes the transpose operator. By defining the signal component vectors $\mathbf{x}$, $\mathbf{i}$ and $\mathbf{n}$, the vector $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{x} + \mathbf{i} + \mathbf{n}. \quad (3)$$

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 352015383 (SFB 1330 B2) and Project ID 390895286 (EXC 2177/1).

Since the target speaker is assumed to be a coherent source, the target speech component vector can be written as

$$\mathbf{x} = \mathbf{h}X_1\,, \tag{4}$$

where X_1 is the target speech component in the first microphone and $\mathbf{h}$ denotes the RTF vector of the target speaker between all (local and external) microphones and the first microphone, i.e.

$$\mathbf{h} = [1, H_{\mathrm{a},2}, \ldots, H_{\mathrm{a},M_\mathrm{a}}, H_{\mathrm{e},1}, \ldots, H_{\mathrm{e},M_\mathrm{e}}] = [\mathbf{h}_\mathrm{a}^T, \mathbf{h}_\mathrm{e}^T]^T\,, \tag{5}$$

where $\mathbf{h}_\mathrm{a}$ denotes the local RTF vector and $\mathbf{h}_\mathrm{e}$ contains the external RTFs. Similarly, the interferer component vector can be written as

$$\mathbf{i} = \mathbf{b}I_1\,, \tag{6}$$

where I_1 is the interferer component in the first microphone and $\mathbf{b}$ denotes the RTF vector of the interferer.
Assuming all signal components to be uncorrelated with each other, the $M \times M$-dimensional noisy covariance matrix $\mathbf{R}_y = \mathscr{E}\{\mathbf{y}\mathbf{y}^H\}$, with $\mathscr{E}\{\cdot\}$ the expectation operator and $\{\cdot\}^H$ the Hermitian transpose operator, can be written as

$$\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_i + \mathbf{R}_n\,, \tag{7}$$

with $\mathbf{R}_x$ the target speech covariance matrix, $\mathbf{R}_i$ the interferer covariance matrix and $\mathbf{R}_n$ the noise covariance matrix. Using (4) and (6), the rank-1 target speech and interferer covariance matrices can be written as

$$\mathbf{R}_x = \phi_x \mathbf{h}\mathbf{h}^H\,, \quad \mathbf{R}_i = \phi_i \mathbf{b}\mathbf{b}^H\,, \tag{8}$$

where $\phi_x = \mathscr{E}\{|X_1|^2\}$ and $\phi_i = \mathscr{E}\{|I_1|^2\}$ denote the target speech power spectral density (PSD) and the interferer PSD in the first microphone, respectively. Assuming the noise field to be homogeneous, the noise covariance matrix is full-rank and given by

$$\mathbf{R}_n = \phi_n \mathbf{\Gamma}\,, \tag{9}$$

where ϕ_n denotes the noise PSD and $\mathbf{\Gamma}$ denotes the spatial coherence matrix of the noise field. The signal-to-interferer ratio (SIR) and signal-to-noise ratio (SNR) are defined as

$$SIR = \frac{\phi_x}{\phi_i}\,, \quad SNR = \frac{\phi_x}{\phi_n}\,. \tag{10}$$

The $M_\mathrm{a} \times M_\mathrm{a}$-dimensional noisy covariance matrix for the LMA only can be extracted from $\mathbf{R}_y$ as

$$\mathbf{R}_{y,\mathrm{a}} = \mathbf{E}_\mathrm{a}\mathbf{R}_y\mathbf{E}_\mathrm{a}^T\,, \tag{11}$$

where $\mathbf{E}_\mathrm{a} = [\mathbf{I}_{M_\mathrm{a}\times M_\mathrm{a}}, \mathbf{0}_{M_\mathrm{a}\times M_\mathrm{e}}]$ is a selection matrix.

3 RTF Vector Estimation

A commonly used method for RTF vector estimation is covariance whitening (CW), which has been thoroughly analyzed [10] and used for several applications [8, 19, 20]. First, a square-root decomposition (e.g., Cholesky decomposition) of the estimated noise covariance matrix $\hat{\mathbf{R}}_n$ is computed, i.e.

$$\hat{\mathbf{R}}_n = \hat{\mathbf{R}}_n^{1/2}\hat{\mathbf{R}}_n^{H/2}\,. \tag{12}$$

A whitening operation is then applied to $\hat{\mathbf{R}}_y - \hat{\mathbf{R}}_n$, which yields the pre-whitened covariance matrix

$$\hat{\mathbf{R}}_y^\mathrm{w} = \hat{\mathbf{R}}_n^{-1/2}(\hat{\mathbf{R}}_y - \hat{\mathbf{R}}_n)\hat{\mathbf{R}}_n^{-H/2}\,. \tag{13}$$

Based on the principal eigenvector $\mathbf{v}_\mathrm{max}$ of $\hat{\mathbf{R}}_y^\mathrm{w}$, the RTF vector is then estimated as

$$\hat{\mathbf{h}}^\mathrm{CW} = \frac{\hat{\mathbf{R}}_n^{1/2}\mathbf{v}_\mathrm{max}}{\mathbf{e}_1^T\hat{\mathbf{R}}_n^{1/2}\mathbf{v}_\mathrm{max}}\,, \tag{14}$$

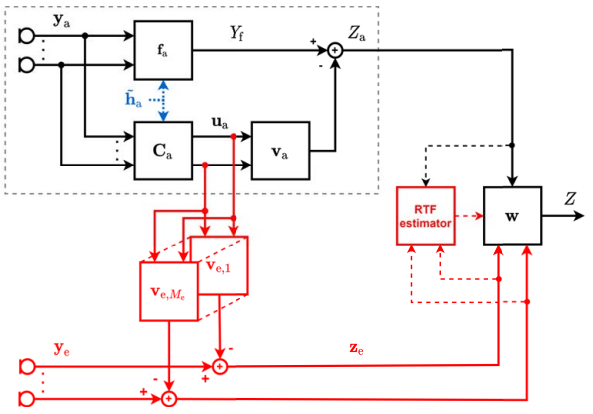

Figure 2: Processing scheme encompassing all considered structures. Upper branch (black) in gray box: L-GSC exploiting a-priori RTF vector $\tilde{\mathbf{h}}_\mathrm{a}$. Including the lower branch (red): GSC-ESR with pre-filters $\mathbf{v}_{\mathrm{e},m_\mathrm{e}}$ or GSC-ER without $\mathbf{v}_{\mathrm{e},m_\mathrm{e}}$.

where $\mathbf{e}_1$ denotes the M-dimensional selection vector, which contains all zeros except for the entry corresponding to the first microphone, which is equal to 1.

We now analyze the CW method for the considered acoustic scenario assuming no estimation errors in $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{R}}_n$. Using (8), the pre-whitened covariance matrix in (13) is given by

$$\mathbf{R}_y^\mathrm{w} = \phi_x\mathbf{R}_n^{-1/2}\mathbf{h}\mathbf{h}^H\mathbf{R}_n^{-H/2} + \phi_i\mathbf{R}_n^{-1/2}\mathbf{b}\mathbf{b}^H\mathbf{R}_n^{-H/2}\,. \tag{15}$$

In case no interferer is present ($\phi_i = 0$), i.e. only the target speaker and noise are present, $\mathbf{R}_y^\mathrm{w}$ in (15) is a rank-1 matrix and the principal eigenvector $\mathbf{v}_\mathrm{max}$ is a scaled version of the pre-whitened target RTF vector $\mathbf{R}_n^{-1/2}\mathbf{h}$. However, if the interferer is present, $\mathbf{R}_y^\mathrm{w}$ in (15) is a rank-2 matrix spanned by the pre-whitened RTF vectors $\mathbf{R}_n^{-1/2}\mathbf{h}$ and $\mathbf{R}_n^{-1/2}\mathbf{b}$, such that the principal eigenvector $\mathbf{v}_\mathrm{max}$ is a linear combination of these vectors, i.e.

$$\mathbf{v}_\mathrm{max} = \alpha_x\mathbf{R}_n^{-1/2}\mathbf{h} + \alpha_i\mathbf{R}_n^{-1/2}\mathbf{b}\,. \tag{16}$$

It can be shown that the weighting factors α_x and α_i depend on the SIR [9], i.e. for high SIR the estimated RTF vector in (14) corresponds predominantly to the target RTF vector, for low SIR it corresponds predominantly to the interferer RTF vector.

4 Local GSC

In this section, we discuss the local GSC (L-GSC) using only the M_a head-mounted microphones of the LMA, i.e. the upper branch in Fig. 2. The GSC [6–8] can be considered as an alternative implementation of the MPDR beamformer [4, 5]. Despite its theoretical equivalence, the GSC structure is chosen here for its larger flexibility. The GSC consists of three processing blocks: (1) a fixed beamformer $\mathbf{f}_\mathrm{a}$, generating a so-called speech reference, (2) a blocking matrix $\mathbf{C}_\mathrm{a}$, generating so-called noise-and-interferer references, and (3) a filter $\mathbf{v}_\mathrm{a}$, aiming at minimizing the correlation between the noise-and-interferer references and the speech reference. For the fixed beamformer and the blocking matrix an estimate of the local RTF vector of the target speaker $\mathbf{h}_\mathrm{a}$ is required, which will be referred to here as the a-priori RTF vector $\tilde{\mathbf{h}}_\mathrm{a}$, since it is assumed to be known.

For the fixed beamformer, we will use the M_a-dimensional matched filter $\mathbf{f}_\mathrm{a} = \tilde{\mathbf{h}}_\mathrm{a}/||\tilde{\mathbf{h}}_\mathrm{a}||_2^2$, for which $\mathbf{f}_\mathrm{a}^H\tilde{\mathbf{h}}_\mathrm{a} = 1$. This matched filter preserves sounds arriving from the direction associated with the a-priori RTF vector $\tilde{\mathbf{h}}_\mathrm{a}$, while passively cancelling sounds arriving from other directions. Applying the fixed beamformer to the LMA signals yields the speech reference Y_f, i.e.

$$Y_\mathrm{f} = \mathbf{f}_\mathrm{a}^H\mathbf{y}_\mathrm{a}\,. \tag{17}$$

ISBN 978-3-8007-5627-8

The $M_\mathrm{a} \times (M_\mathrm{a}-1)$-dimensional blocking matrix $\mathbf{C}_\mathrm{a}$ is designed to be orthogonal to $\tilde{\mathbf{h}}_\mathrm{a}$, i.e. $\mathbf{C}_\mathrm{a}^H \tilde{\mathbf{h}}_\mathrm{a} = \mathbf{0}_{(M_\mathrm{a}-1)\times 1}$, and can therefore be constructed as [7, 8]

$$\mathbf{C}_\mathrm{a} = \begin{bmatrix} -\tilde{H}_{\mathrm{a},2}^*, -\tilde{H}_{\mathrm{a},3}^*, \ldots, -\tilde{H}_{\mathrm{a},M_\mathrm{a}}^* \\ \mathbf{I}_{(M_\mathrm{a}-1)\times(M_\mathrm{a}-1)} \end{bmatrix}. \tag{18}$$

The blocking matrix aims at blocking out sounds arriving from the direction associated with the a-priori RTF vector. Applying the blocking matrix to the head-mounted microphone signals yields $(M_\mathrm{a}-1)$ noise-and-interferer references, i.e.

$$\mathbf{u}_\mathrm{a} = \mathbf{C}_\mathrm{a}^H \mathbf{y}_\mathrm{a}. \tag{19}$$

The filter $\mathbf{v}_\mathrm{a}$ aims at minimizing the correlated parts between the speech references Y_f and the noise-and-interferer references $\mathbf{u}_\mathrm{a}$. It should be noted that contrary to [16], the filter $\mathbf{v}_\mathrm{a}$ is designed to minimize the total power, corresponding to an MPDR implementation, instead of minimizing the power of the noise component only, i.e.

$$\mathbf{v}_\mathrm{a} = \left(\mathbf{C}_\mathrm{a}^H \hat{\mathbf{R}}_{y,\mathrm{a}} \mathbf{C}_\mathrm{a}\right)^{-1} \mathbf{C}_\mathrm{a}^H \hat{\mathbf{R}}_{y,\mathrm{a}} \mathbf{f}_\mathrm{a}. \tag{20}$$

When minimizing the power of the noise component, i.e. using $\hat{\mathbf{R}}_{n,\mathrm{a}}$ instead of $\hat{\mathbf{R}}_{y,\mathrm{a}}$ in (20), the interferer is hardly reduced. When minimizing the total power instead, the interferer will be reduced, but it comes with the risk of target speech cancellation in case of speech leakage in the noise-and-interferer references, which occurs in case of a mismatch between the a-priori RTF vector $\tilde{\mathbf{h}}_\mathrm{a}$ and the (true) local RTF vector $\mathbf{h}_\mathrm{a}$. The output signal of the L-GSC is then given by

$$Z_\mathrm{a} = Y_\mathrm{f} - \mathbf{v}_\mathrm{a}^H \mathbf{u}_\mathrm{a}. \tag{21}$$

5 Extended GSC Structures

In this section, we present and discuss two extended GSC structures incorporating the eMics. The first structure, the GSC with external speech references (GSC-ESR) is adopted from [16] and modified in order to achieve joint noise and interferer reduction. The GSC-ESR aims at suppressing the interferer component in the eMic signals by pre-filtering the noise-and-interferer references of the L-GSC. The second structure, the GSC with external references (GSC-ER) is a simpler structure without this pre-filtering operation. The RTF vector of the target speaker between the eMics and the first microphone is then estimated based on the output signal of the L-GSC and the pre-processed/unprocessed eMic signals, which is subsequently used in an MPDR beamformer.

5.1 GSC-ESR

The GSC-ESR structure was introduced in [16] (there referred to as GEVD-based method) and is depicted in Fig. 2. Similarly as for the L-GSC, we propose to use an MPDR implementation instead of an MVDR implementation (used in [16]), since otherwise the interferer cannot be be actively suppressed in the eMic signals. It should however be noted that speech leakage in the noise-and-interferer references $\mathbf{u}_\mathrm{a}$ may lead to target speech cancellation in the pre-processed eMic signals $\mathbf{z}_\mathrm{e}$. The main idea is to pre-filter the noise-and-interferer references $\mathbf{u}_\mathrm{a}$ of the L-GSC using the filters $\mathbf{v}_{\mathrm{e},m_\mathrm{e}}$, aiming at cancelling correlated components between $\mathbf{u}_\mathrm{a}$ and the eMic signals $\mathbf{y}_\mathrm{e}$, i.e.

$$\mathbf{v}_{\mathrm{e},m_\mathrm{e}} = \left(\mathbf{C}_\mathrm{a}^H \hat{\mathbf{R}}_{y,\mathrm{a}} \mathbf{C}_\mathrm{a}\right)^{-1} \mathbf{C}_\mathrm{a}^H \mathbf{E}_\mathrm{a} \hat{\mathbf{R}}_y \mathbf{e}_{\mathrm{e},m_\mathrm{e}}, m_\mathrm{e} \in \{1, \ldots, M_\mathrm{e}\}, \tag{22}$$

with $\mathbf{e}_{\mathrm{e},m_\mathrm{e}}$ a selection vector for the m_e-th external microphone. The pre-processed eMic signals $\mathbf{z}_\mathrm{e}$ are given by

$$\mathbf{z}_\mathrm{e} = \mathbf{y}_\mathrm{e} - [\mathbf{v}_{\mathrm{e},1}, \ldots, \mathbf{v}_{\mathrm{e},M_\mathrm{e}}]^H \mathbf{u}_\mathrm{a} \tag{23}$$

Subsequently, the output signal Z_a of the L-GSC and the pre-processed eMic signals $\mathbf{z}_\mathrm{e}$ are combined using an MPDR beamformer, i.e.

$$\mathbf{w} = \frac{\hat{\mathbf{R}}_{y,z}^{-1} \hat{\mathbf{h}}_z}{\hat{\mathbf{h}}_z^H \hat{\mathbf{R}}_{y,z}^{-1} \hat{\mathbf{h}}_z}, \tag{24}$$

where $\hat{\mathbf{R}}_{y,z}$ is an estimate of the pre-processed covariance matrix $\mathbf{R}_{y,z}$, defined as

$$\mathbf{R}_{y,z} = \mathscr{E}\left\{ \begin{bmatrix} Z_\mathrm{a} \\ \mathbf{z}_\mathrm{e} \end{bmatrix} [Z_\mathrm{a}^*, \mathbf{z}_\mathrm{e}^H] \right\}, \tag{25}$$

and $\hat{\mathbf{h}}_z$ is the RTF vector estimate of the target speaker between the external microphones and the first microphone. This estimate is obtained by applying the CW method discussed in Section 3 on $\hat{\mathbf{R}}_{y,z}$ and $\hat{\mathbf{R}}_{n,z}$, where $\mathbf{R}_{n,z}$ is defined similarly to $\mathbf{R}_{y,z}$ in (25). The output signal of the GSC-ESR is obtained as

$$Z = \mathbf{w}^H \begin{bmatrix} Z_\mathrm{a} \\ \mathbf{z}_\mathrm{e} \end{bmatrix}. \tag{26}$$

Based on (16), the RTF vector estimate $\hat{\mathbf{h}}_z$ will be more accurate when more interferer is suppressed by the L-GSC and by the pre-filters $\mathbf{v}_{\mathrm{e},m_\mathrm{e}}$.

5.2 GSC-ER

The GSC-ER can be seen as a simplified version of the GSC-ESR, where the pre-filters $\mathbf{v}_{\mathrm{e},m_\mathrm{e}}$ are set to zero (see Fig. 2), such that

$$Z_{\mathrm{e},m_\mathrm{e}} = Y_{\mathrm{e},m_\mathrm{e}}. \tag{27}$$

On the one hand, this means that no interferer is suppressed in the eMic signals $\mathbf{y}_\mathrm{e}$, resulting in estimation errors for the RTF vector $\hat{\mathbf{h}}_z$ (depending on the SIR). On the other hand, no target speech cancellation in the eMic signals occurs due to speech leakage in the noise-and-interferer references $\mathbf{u}_\mathrm{a}$.

6 Experimental Results

In this section, we experimentally evaluate the performance of the considered GSC structures, i.e. the L-GSC using only the head-mounted microphones and the GSC-ESR and GSC-ER incorporating the eMics, at different input SIR and SNR. Furthermore, we investigate their robustness against mismatches of the local RTF vector.

6.1 Recording Setup and Implementation

The investigated algorithms are evaluated using real-world signals, recorded in a laboratory at the University of Oldenburg with a reverberation time of approximately 350 ms. The LMA consisted of binaural hearing aids with two microphones per ear, i.e. M_a = 4 local microphones. The hearing aids were mounted on a KEMAR head-and-torso simulator (HATS). The reference microphone was the front microphone on the left side. In addition, M_e = 2 eMics were placed in the room as depicted in Fig. 1, approximately 1.5 m from the HATS. The target speaker was a male English speaker played back via a loudspeaker, placed about 35° to the right of the HATS. The interferer was a female English speaker played back via a loudspeaker, placed about 35° to the left of the HATS. Both loudspeakers were placed about 2 m from the HATS and 0.5 m from the eMics. Diffuse-like noise was generated with four loudspeakers facing the corners of the laboratory, playing back different versions of multi-talker babble noise. The signal components $\mathbf{x}$, $\mathbf{i}$ and $\mathbf{n}$ were recorded separately at a sampling frequency of 16 kHz and subsequently mixed at input SIR SIR_in = {-10, 0, 10} dB and input SNR SNR_in= {-10, 0, 10} dB.

For the STFT framework, the following parameters were used: a frame length of 1024 samples (corresponding to 64 ms), a frame overlap of 50% and a square-root Hann window as analysis and

ISBN 978-3-8007-5627-8

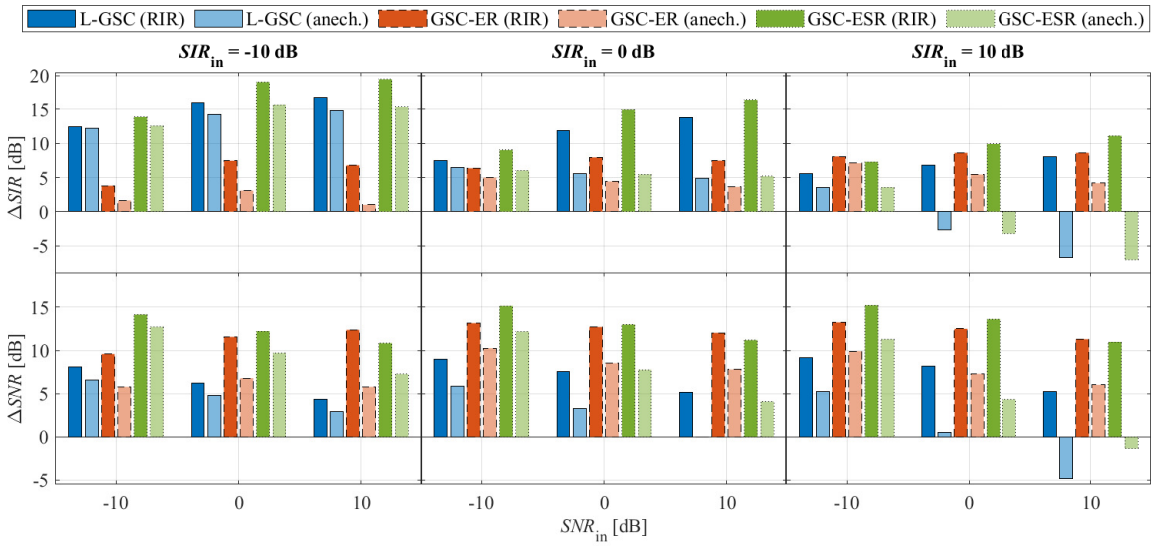

Figure 3: Broadband SIR and SNR improvement for the considered GSC structures (L-GSC, GSC-ESR, GSC-ER) for different input SIRs and SNRs, either using an ideal RTF vector (RIR) or an anechoic RTF vector (anech.) for the L-GSC.

synthesis windows. To allow for a more controlled investigation of the considered algorithms without the interplay of different SIRs, SNRs and a voice activity detection algorithm, the covariance matrices were estimated in a batch implementation assuming oracle knowledge about the noise. For the L-GSC, two different a-priori RTF vectors were considered: (1) an ideal RTF vector obtained from the measured room impulse response of the target speaker, denoted by "RIR" and (2) an anechoic RTF vector from a database using the same hearing aids [21], in the direction of the target speaker (denoted by "anech."). As described in Section 5, the external RTF vector $\hat{\mathbf{h}}_z$ in the extended GSC structures was estimated using CW.

The performance of the considered algorithms was evaluated in terms of broadband SNR improvement (ΔSNR) and broadband SIR improvement (ΔSIR), where the powers of the signal components were computed in the time domain in sequences where the target speaker and the interferer were active simultaneously.

6.2 Results

For the considered GSC structures, Fig. 3 depicts the broadband SIR and SNR improvement for different input SIRs and SNRs.

In case of an ideal a-priori RTF vector ("RIR"), it can be observed that the baseline system, i.e. the L-GSC, performs as expected according to the theoretical results from [22, 23]. At high SNRs, a larger SIR improvement and lower SNR improvement is obtained than at low SNRs. Realizing that the noise is diffuse, this can be explained by the fact that at high SNRs the interferer can be suppressed more due to a higher correlation between the noise-and-interferer references $\mathbf{u}_\mathrm{a}$ and the speech reference Y_f. Compared to the L-GSC, it can be observed that the SIR improvement of the GSC-ER is worse (except at SIR_in = 10 dB) and the SNR improvement is better. This can be explained by the fact that at low SIRs, the interferer is the dominant source (also in the eMics) and therefore strongly influences the estimation of the RTF vector $\hat{\mathbf{h}}_z$. The subsequent MPDR beamformer therefore mostly cancels the target and preserves the interferer, compared to the L-GSC. The GSC-ESR outperforms the baseline system for all considered scenarios in terms of both SNR improvement (by up to 6 dB) and as well as SIR improvement (by up to 4 dB). At low SIR, the pre-filters $\mathbf{v}_{\mathrm{e},m_\mathrm{e}}$ are able to suppress the interferer rather well from the eMics, such that the SIR in the pre-processed eMic signals $\mathbf{z}_\mathrm{e}$ is larger than in the unprocessed eMic signals $\mathbf{y}_\mathrm{e}$. This leads to a more accurate estimate of RTF vector $\hat{\mathbf{h}}_z$ in comparison to the GSC-ER, which uses the unprocessed eMic signals. At high SIR, the performance differences between the GSC-ESR and the GSC-ER become smaller, since the noise component can be assumed to be uncorrelated between the eMics and the LMA which leads to a lower correlation between the noise-and-interferer references $\mathbf{u}_\mathrm{a}$ and the eMic signals $\mathbf{y}_\mathrm{e}$.

In case of RTF mismatch, i.e. when using the anechoic local RTF vector ("anech."), it can be observed that the SIR improvement and the SNR improvement decrease for all algorithms, especially at high SIR and SNR. At high SIR, target speech leakage into the noise-and-interferer references $\mathbf{u}_\mathrm{a}$ has the most severe consequences: if there is so much speech leakage that the target speech is the most coherent source between $\mathbf{u}_\mathrm{a}$ and the speech reference Y_f (or the eMic signals $\mathbf{y}_\mathrm{e}$ respectively), the target can partly be cancelled due to the MPDR implementation in (20) and (22). On the one hand, this leads to a decreased performance of the L-GSC, even leading to negative ΔSIR and ΔSNR at SIR_in = 10 dB. On the other hand, the target speech cancellation in the pre-processed eMic signals $\mathbf{z}_\mathrm{e}$ leads to a less accurate estimation of the RTF vector $\hat{\mathbf{h}}_z$. Hence, at high SIR the performance drop in terms of SIR improvement and SNR improvement is much larger for the GSC-ESR than for the GSC-ER, even leading to negative ΔSIR and ΔSNR for the GSC-ESR at SIR_in = 10 dB. Nevertheless, in adverse conditions ($SIR_\mathrm{in} \leqslant 0$ dB, $SNR_\mathrm{in} \leqslant 0$ dB) the performance of the GSC-ESR is better or comparable to the performance of the L-GSC and the GSC-ER.

7 Conclusions

In this paper, we compared two extended GSC structures which use eMics in conjunction with an LMA in terms of noise and interferer reduction. The GSC-ESR uses the noise-and-interferer references of the L-GSC to pre-process the eMic signals aiming at reducing the interferer component. A simplified version of the GSC-ESR is the GSC-ER, where no pre-processing of the eMic signals is performed. Aiming at achieving joint noise and interferer reduction, we proposed to use an MPDR implementation instead of an MVDR implementation for all processing blocks.

Experimental results with reverberant signals showed that the GSC-ESR yields the best results in terms of noise and interferer reduction, especially in scenarios where the input SIR is low. In case of an ideal local RTF vector the GSC-ESR outperformed the L-GSC and the GSC-ER in all conditions. In case of RTF mismatch, speech leakage occurred which caused target speech cancellation in the L-GSC and the pre-processing in the GSC-ESR, heavily affecting the performance of all algorithms at high SIR and SNR. Nevertheless, in adverse conditions the GSC-ESR still outperformed the L-GSC and the GSC-ER.

ISBN 978-3-8007-5627-8

References

[1] V. Hamacher, U. Kornagel, T. Lotter, and H. Puder, "Binaural signal processing in hearing aids: Technologies and algorithms," in *Advances in Digital Speech Transmission*, ch. 14, pp. 401–429, Wiley, 2008.

[2] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, pp. 18–30, Mar. 2015.

[3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, pp. 692–730, Apr. 2017.

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, pp. 4–24, Apr. 1988.

[5] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, pp. 269–302, Wiley, 2010.

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

[7] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Processing*, vol. 49, pp. 1614–1626, Aug. 2001.

[8] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, pp. 1071–1086, Aug. 2009.

[9] M. Taseska and E. A. P. Habets, "Relative transfer function estimation exploiting instantaneous signals and the signal subspace," in *Proc. European Signal Processing Conference*, (Nice, France), pp. 404–408, Aug. 2015.

[10] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *Proc. European Signal Processing Conference*, (Rome, Italy), pp. 544–548, Sep. 2018.

[11] M. Tammen, I. Kodrasi, and S. Doclo, "Joint estimation of RETF vector and power spectral densities based on alternating least squares," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, (Brighton, UK), pp. 795–799, May 2019.

[12] A. Bertrand and M. Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Jan. 2009.

[13] J. Szurley, A. Bertrand, B. van Dijk, and M. Moonen, "Binaural noise cue preservation in a binaural noise reduction system with a remote microphone signal," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, pp. 952–966, May 2016.

[14] D. Yee, H. Kamkar-Parsi, R. Martin, and H. Puder, "A noise reduction post-filter for binaurally-linked single-microphone hearing aids utilizing a nearby external microphone," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, pp. 5–18, Jan. 2017.

[15] N. Gößling and S. Doclo, "Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field," in *Proc. International Workshop on Acoustic Signal Enhancement*, (Tokyo, Japan), pp. 146–150, Sep. 2018.

[16] R. Ali, G. Bernardi, T. van Waterschoot, and M. Moonen, "Methods of extending a generalized sidelobe canceller with external microphones," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, pp. 1349–1364, Sep. 2019.

[17] N. Gößling, W. Middelberg, and S. Doclo, "RTF-steered binaural MVDR beamforming incorporating multiple external microphones," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, USA), pp. 368–372, Oct. 2019.

[18] N. Gößling, D. Marquardt, and S. Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 462–476, Dec. 2020.

[19] I. Kodrasi and S. Doclo, "EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, (San Francisco, USA), pp. 116–120, Mar. 2017.

[20] R. Serizel, M. Moonen, B. van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, pp. 785–799, Apr. 2014.

[21] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel In-Ear and Behind-The-Ear Head-Related and Binaural Room Impulse Responses," *Eurasip Journal on Advances in Signal Processing*, vol. 2009, p. 10 pages, 2009.

[22] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, (Phoenix, AZ, USA), pp. 2965–2968, Mar. 1999.

[23] S. E. Nordholm and Y. Hong Leung, "Performance limits of the broadband generalized sidelobe cancelling structure in an isotropic noise field," *Journal of the Acoustical Society of America*, vol. 107, pp. 1057–1060, Feb. 2000.

ISBN 978-3-8007-5627-8