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 (MPDR) beamformer instead of an MVDR beamformer such that not only the target speaker is known. In this paper, we consider the structure proposed in [16] and reformulate it in terms of an MPDR beamformer. 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., frontoral directional).

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 sig-

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 $M = M_a + M_e$ microphones. In the short time Fourier transform (STFT) domain the $m$-th microphone signal is given by

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

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 $y$, we distinguish between the $M_a$ local microphones and the $M_e$ eMics, i.e.,

$$y = [Y_{a,1}, Y_{a,2}, \ldots, Y_{a,M_a}, Y_{e,1}, \ldots, Y_{e,M_e}]^T = [y_a^T, y_e^T]^T, \ (2)$$

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

$$y = x + i + n. \ (3)$$

ISBN 978-3-8007-5627-8
© VDE VERLAG GMBH · Berlin · Offenbach
Since the target speaker is assumed to be a coherent source, the target speech component vector can be written as
\[ x = hX_1, \]
where \( X_1 \) is the target speech component in the first microphone and \( h \) denotes the RTF vector of the target speaker between all (local and external) microphones and the first microphone, i.e.
\[ h = [1, H_{a,2}, \ldots, H_{a,M}, H_{e,1}, \ldots, H_{e,M}] = [h^T_1, h^T_e]^T, \]
where \( h_a \) denotes the local RTF vector and \( h_e \) contains the external RTFs. Similarly, the interferer component vector can be written as
\[ i = bI_1, \]
where \( I_1 \) is the interferer component in the first microphone and \( b \) 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 \( R_y = \mathcal{E} \{ y y^H \} \), with \( \mathcal{E} \{ \cdot \} \) the expectation operator and \( \{ \cdot \}^H \) the Hermitian transpose operator, can be written as
\[ R_y = R_x + R_i + R_n, \]
with \( R_x \) the target speech covariance matrix, \( R_i \) the interferer covariance matrix and \( R_n \) the noise covariance matrix. Using (4) and (6), the rank-1 target speech and interferer covariance matrices can be written as
\[ R_x = \phi_x hh^H, \quad R_i = \phi_i bb^H, \]
where \( \phi_x = \mathcal{E} \{ |X_1|^2 \} \) and \( \phi_i = \mathcal{E} \{ |I_1|^2 \} \) 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
\[ R_n = \phi_n \Gamma, \]
where \( \phi_n \) denotes the noise PSD and \( \Gamma \) denotes the spatial coherence matrix of the noise field. The signal-to-interference ratio (SIR) and signal-to-noise ratio (SNR) are defined as
\[ SIR = \frac{\phi_x}{\phi_i}, \quad SNR = \frac{\phi_x}{\phi_n}. \]
The \( M_x \times M_x \)-dimensional noisy covariance matrix for the LMA can only be extracted from \( R_y \) as
\[ R_y,a = E_a R_y R_a^T, \]
where \( E_a = [I_{M_x \times M_x}, 0_{M_x \times M_e}] \) 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 \( \tilde{R}_n \) is computed, i.e.,
\[ \tilde{R}_n = \tilde{R}_n^{1/2} \tilde{R}_n^{H/2}. \]
A whitening operation is then applied to \( \tilde{R}_y - \tilde{R}_n \), which yields the pre-whitened covariance matrix
\[ \tilde{R}_y^w = \tilde{R}_n^{1/2} (\tilde{R}_y - \tilde{R}_n) \tilde{R}_n^{H/2}. \]
Based on the principal eigenvector \( \nu_{\max} \) of \( \tilde{R}_y^w \), the RTF vector is then estimated as
\[ h^C W = \frac{\tilde{R}_n^{1/2} \nu_{\max}}{\mathbf{e}_1^T \tilde{R}_n^{1/2} \nu_{\max}}, \]
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 \( \tilde{R}_y \) and \( \tilde{R}_n \). Using (8), the pre-whitened covariance matrix in (13) is given by
\[ R_y^w = \phi_x R_n^{-1/2} hh^H R_n^{-H/2} + \phi_i R_n^{-1/2} bb^H R_n^{-H/2}. \]
In case no interferer is present (\( \phi_i = 0 \)), i.e. only the target speaker and noise are present, \( R_y^w \) in (15) is a rank-1 matrix and the principal eigenvector \( \nu_{\max} \) is a scaled version of the pre-whitened target RTF vector \( R_n^{-1/2} h \). However, if the interferer is present, \( R_y^w \) in (15) is a rank-2 matrix spanned by the pre-whitened RTF vectors \( R_n^{-1/2} h \) and \( R_n^{-1/2} b \), such that the principal eigenvector \( \nu_{\max} \) is a linear combination of these vectors, i.e.,
\[ \nu_{\max} = \alpha_x R_n^{-1/2} h + \alpha_i R_n^{-1/2} b. \]
It can be shown that the weighting factors \( \alpha_x \) and \( \alpha_i \) 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 \( M_x \) 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 \( f_x \), generating a so-called speech reference, (2) a blocking matrix \( C_x \), generating so-called noise-and-interferer references, and (3) a filter \( v_x \), 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 \( h_x \) is required, which will be referred to here as the a-priori RTF vector \( h_a \), since it is assumed to be known.

For the fixed beamformer, we will use the \( M_x \)-dimensional matched filter \( f_x = h_a / |h_a|^2 \), for which \( f_x^H f_x = 1 \). This matched filter preserves sounds arriving from the direction associated with the a-priori RTF vector \( h_a \), while passively cancelling sounds arriving from other directions. Applying the fixed beamformer to the LMA signals yields the speech reference \( Y_f \), i.e.,
\[ Y_f = f_x^H y_a. \]
The $M_a \times (M_a - 1)$-dimensional blocking matrix $C_a$ is designed to be orthogonal to $\mathbf{h}_a$, i.e. $C_a^H \mathbf{h}_a = 0_{(M_a-1)\times 1}$, and can therefore be constructed as [7, 8]

$$C_a = \begin{bmatrix} -\mathbf{f}_{a,2} & -\mathbf{f}_{a,3} & \cdots & -\mathbf{f}_{a,M_a} \\ \mathbf{I}_{(M_a-1)\times (M_a-1)} \\ \end{bmatrix}. \quad (18)$$

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 $(M_a - 1)$ noise-and-interferer references, i.e.

$$\mathbf{u}_a = C_a^H \mathbf{y}_a. \quad (19)$$

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

$$\mathbf{v}_a = \left( C_a^H \hat{\mathbf{R}}_{y,a} C_a \right)^{-1} C_a^H \hat{\mathbf{R}}_{y,a} \mathbf{u}_a. \quad (20)$$

When minimizing the power of the noise component, i.e. using $\hat{\mathbf{R}}_{y,a}$ instead of $\mathbf{R}_{y,a}$ 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 $\mathbf{h}_a$ and the (true) local RTF vector $\hat{\mathbf{h}}_a$. The output signal of the L-GSC is then given by

$$Z_a = \mathbf{y}_f - \mathbf{v}_a^H \mathbf{u}_a. \quad (21)$$

### 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 targets 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 $\mathbf{u}_a$ may lead to target speech cancellation in the pre-processed eMic signals $\mathbf{z}_e$. The main idea is to pre-filter the noise-and-interferer references $\mathbf{u}_a$ of the L-GSC using the filters $\mathbf{v}_{e,m_e}$, aiming at cancelling correlated components between $\mathbf{u}_a$ and the eMic signals $\mathbf{y}_e$, i.e.

$$\mathbf{v}_{e,m_e} = \left( C_a^H \hat{\mathbf{R}}_{y,a} C_a \right)^{-1} C_a^H \hat{\mathbf{R}}_{y,a} \mathbf{e}_{e,m_e}, \quad m_e \in \{1, \ldots, M_e\}, \quad (22)$$

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

$$\mathbf{z}_e = \mathbf{y}_e - [\mathbf{v}_{e,1}, \ldots, \mathbf{v}_{e,M_e}]^H \mathbf{u}_a \quad (23)$$

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

$$\mathbf{w} = \frac{\hat{\mathbf{R}}_{y,a}^{-1} \hat{\mathbf{h}}_a}{\hat{\mathbf{h}}_a^H \hat{\mathbf{R}}_{y,a}^{-1} \hat{\mathbf{h}}_a}, \quad (24)$$

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

$$\mathbf{R}_{y,a} = \mathbf{E} \left[ \mathbf{z}_a \mathbf{z}_a^H \right], \quad (25)$$

and $\hat{\mathbf{h}}_a$ 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 $\hat{\mathbf{R}}_{y,a}$ 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

$$Z = \mathbf{w}^H \mathbf{z}_a. \quad (26)$$

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

#### 5.2 GSC-ER

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

$$\mathbf{z}_{e,m_e} = \mathbf{y}_{e,m_e}. \quad (27)$$

On the one hand, this means that no interferer is suppressed in the eMic signals $\mathbf{y}_e$, resulting in estimation errors for the RTF vector $\hat{\mathbf{h}}_a$ (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 $\mathbf{u}_a$.

### 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. $M_e = 4$ 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, $M_a = 2$ 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 $x, n$ and $\mathbf{w}$ were recorded separately at a sampling frequency of 16 kHz and subsequently mixed at input SIR $\text{SNR}_{\text{in}} = [10, 20, 0, 10]$ dB and input SNR $\text{SNR}_{\text{in}} = [10, 0, 10]$ 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
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 $\hat{h}_z$ in the extended GSC structures was estimated using CW.

The performance of the considered algorithms was evaluated in terms of broadband SNR improvement ($\Delta\text{SNR}$) and broadband SIR improvement ($\Delta\text{SIR}$), 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 $u_h$ 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 $\text{SIR}_{\text{in}} = 10$ 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 $\hat{h}_z$. The subsequent MPDR beamformer therefore mostly cancels the target and preserves the interferer, compared to the L-GSC. The GSC-ER 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 $v_{e_m}$ are able to suppress the interferer rather well from the eMics, such that the SIR in the pre-processed eMic signals $x_e$ is larger than in the unprocessed eMic signals $y_e$. This leads to a more accurate estimate of RTF vector $\hat{h}_z$ 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 $u_h$ and the eMic signals $y_e$.

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 $u_h$ has the most severe consequences: if there is so much speech leakage that the target speech is the most coherent source between $u_h$ and the speech reference $Y_f$ (or the eMic signals $y_e$ 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 $\Delta\text{SIR}$ and $\Delta\text{SNR}$ at $\text{SIR}_{\text{in}} = 10$ dB. On the other hand, the target speech cancellation in the pre-processed eMic signals $x_e$ leads to a less accurate estimation of the RTF vector $\hat{h}_z$. 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 $\Delta\text{SIR}$ and $\Delta\text{SNR}$ for the GSC-ESR at $\text{SIR}_{\text{in}} = 10$ dB. Nevertheless, in adverse conditions ($\Delta\text{SIR}_z < 0$ dB, $\Delta\text{SNR}_z < 0$ 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-ER 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.
References


