Carl von Ossietzky
University of Oldenburg

Master Thesis

GSC-Based Noise and Interferer Reduction for Binaural Hearing Aids Exploiting External Microphones

---

Master of Science Program
Engineering Physics

Submitted by
Wiebke Middelberg, B.Eng.

Supervisor
Prof. Dr. ir. Simon Doclo

Co-Supervisor
Prof. Dr.-Ing. Jörg Bitzer

Oldenburg, June 15, 2021

Acknowledgements

In the course of this thesis, I had the pleasure of working on a great, versatile topic which allowed me to fully pursue my interests. For this, I would firstly like to thank Prof. Dr. Simon Doclo, who encouraged me to think critically, who gave me the opportunity to develop my own ideas and with whom I had many interesting discussions. I would also like to thank Prof. Dr. Jörg Bitzer for the co-supervision of my thesis and giving valuable feedback and helpful insights from a different perspective.

A big thanks goes to my colleagues. Especially Henri and Piero who morally supported me in times of Corona by being present in the NeSSy and lending a hand whenever needed, either by general advice, sessions on the whiteboard, or needed distractions like funny evenings with great playlists (and too much beer). I would also like to thank Klaus for proof reading my thesis.

Of course I would also like to thank my parents for always having my back and supporting me in every possible way, not only throughout this thesis, but rather throughout my entire studies, which allowed me to focus on achieving my goals.

Abstract

In this thesis, the problem of noise and interfering speech decreasing the intelligibility of a target speaker in hearing aid applications is addressed. Considering state-of-the-art beamformers like the minimum variance distortionless response (MVDR) beamformer or the minimum power distortionless response (MPDR) beamformer, an estimate of the target relative transfer function (RTF) vector is required to steer the beamformer. In this thesis, it is shown that in the presence of an interferer, blind RTF vector estimation using the covariance whitening method yields a biased estimate of the target RTF vector, whose accuracy depends on the signal-to-interferer ratio (SIR). Hence, at low SIRs, the stated goal of joint noise and interferer reduction cannot be achieved by means of blind beamforming approaches. Assuming that the the target RTF vector of the hearing aid microphones is known, a local GSC (L-GSC) steering towards the target speaker can be used. Since the performance of the hearing aid microphones is limited, additional external microphones (eMics) are considered, for which, however, the target RTF vector is unknown. For the incorporation of the eMics, four different extended GSC structures are presented, in which the external target RTF vector is estimated on the pre-filtered local microphone signals. In an experimental evaluation with real-world recordings it is shown that only one structure performs well in adverse scenarios with a low input SIR, namely the GSC with external speech references (GSC-ESR), which exploits the noise-and-interferer references of the hearing aid microphone signals to pre-process the eMic signals. This pre-processing aims at minimizing the correlated parts between the noise-and-interferer references and the eMic signals, which ideally leads to an increased SIR in the eMic signals, allowing for a better RTF vector estimation. The evaluation also shows the downside of all filters being implemented such that they aim at minimizing the total output power: If the RTF vector for the L-GSC is subject to a mismatch with the true target RTF vector, all considered structures, but especially the GSC-ESR suffer from target speech cancellation. This occurs particularly at high SIRs, while the performance at low SIR, i.e., in adverse conditions where a reliable processing is most valuable, still is the best among all considered algorithms.

Zusammenfassung

In dieser Arbeit wird das Problem des Rauschens und der störenden Sprache, die die Verständlichkeit des Zielsprechers in Hörgeräteanwendungen verringern, behandelt. Bei gängigen Beamformern wie dem Minimum-Varianz-Distortionless-Response- (MVDR) Beamformer oder dem Minimum-Power-Distortionless-Response- (MPDR) Beamformer wird eine Schätzung der relativen Übertragungsfunktion (RTF) des Zielsprechers benötigt, um den Beamformer zu steuern. In dieser Arbeit wird gezeigt, dass blinde RTF-Schätzung mit dem Covariance-Whitenign-Verfahren in Anwesenheit eines Störsprechers eine verzerrte Schätzung des Ziel-RTF-Vektors ergeben, deren Genauigkeit vom Signal-zu-Störer-Abstands (SIR) abhängt. Daher kann bei niedrigen SIRs das erklärte Ziel der gemeinsamen Rausch- und Störer-Reduzierung nicht durch blinde Beamforming-Ansätze erreicht werden. Unter der Annahme, dass der Ziel-RTF-Vektor der Hörgerätemikrofone bekannt ist, kann ein lokaler Generalized Sidelobe Canceller (L-GSC) verwendet werden, der auf den Zielsprecher ausgerichtet ist. Da die Leistungsfähigkeit der Hörgerätemikrofone begrenzt ist, werden zusätzlich externe Mikrofone (eMics) berücksichtigt, für die jedoch der Ziel-RTF-Vektor unbekannt ist. Für die Einbindung der eMics werden vier verschiedene erweiterte GSC-Strukturen vorgestellt, bei denen der externe Ziel-RTF-Vektor auf den vorgefilterten lokalen Mikrofonsignalen geschätzt wird. In einer experimentellen Auswertung mit realen Aufnahmen wird gezeigt, dass nur eine Struktur in schwierigen Szenarien mit einem niedrigen Eingangs-SIR gut abschneidet, nämlich der Generalized Sidelobe Canceller mit externen Sprachreferenzen (GSC-ESR), der die Rausch- und Störerreferenzen der Hörgerätemikrofonsignale zur Vorverarbeitung der eMic-Signale nutzt. Diese Vorverarbeitung zielt darauf ab, die korrelierten Anteile zwischen den Rausch- und-Störerreferenzen und den eMic-Signalen zu minimieren, was im Idealfall zu einem erhöhten SIR in den eMic-Signalen führt und eine bessere RTF-Vektorschätzung ermöglicht. Die Auswertung zeigt aber auch die Nachteile aller Filter, die so implementiert sind, dass sie auf eine Minimierung der Gesamtausgangsleistung abzielen: Wenn der RTF-Vektor für den L-GSC einer Diskrepanz zum wahren Ziel-RTF-Vektor unterliegt, leiden alle betrachteten Strukturen, insbesondere aber der GSC-ESR, unter Auslöschung des Zielsprechers. Dies tritt vor allem bei hohen SIRs auf, während die Leistung bei niedrigen SIRs, also unter schwierigen Bedingungen, bei denen eine zuverlässige Verarbeitung am wertvollsten ist, immer noch die beste unter allen betrachteten Algorithmen ist.

List of Figures

List of Abbreviations

ATF	acoustic transfer function
CPSD	cross-power spectral density
CW	covariance whitening
DoA	direction of arrival
eMic	external microphone
EVD	eigenvalue decomposition
GEVD	generalized eigenvalue decomposition
GSC	generalized sidelobe canceller
GSC-ENR-1	GSC with external noise references type 1
GSC-ENR-2	GSC with external noise references type 2
GSC-ER	GSC with external references
GSC-ESR	GSC with external speech references
HATS	head-and-torso simulator
INR	interferer-to-noise ratio
ISTFT	inverse short-time Fourier transform
INR	interferer-to-noise ratio
LCMV	linearly constraint minimum variance
L-GSC	local GSC
LMA	local microphone array
MPDR	minimum power distortionless response
MSC	magnitude-squared coherence
MVDR	minimum variance distortionless response
NLMS	normalized least-mean squares

NR	noise reduction
PSD	power spectral density
RIR	room impulse response
RTF	relative transfer function
SC	spatial coherence
SD	speech distortion
SIR	signal-to-interferer ratio
SINR	signal-to-interferer-and-noise ratio
SNR	signal-to-noise ratio
SPP	speech presence probability
STFT	short-time Fourier transform
ULA	uniform linear array
VAD	voice activity detection

Contents

1 Introduction

In assistive listening devices like hearing aids or cochlear implants, speech enhancement is an important task to increase speech intelligibility. Besides suppressing background noise, also the reduction of undesired, interfering speakers is essential [1–3], since the presence of multiple speech sources leads to a lower intelligibility of the target speaker [4].

To achieve noise and interferer reduction, many algorithms based on the minimum variance distortionless response (MVDR) beamformer and the minimum power distortionless response (MPDR) beamformer [5,6] or respectively the generalized sidelobe canceller (GSC) structure exist [7,8]. One well studied beamforming algorithm which allows to treat different speakers differently, e.g., preserve one and suppress another one, is the linearly constraint minimum variance (LCMV) beamformer [5,6,9]. However, in practice, the LCMV beamformer is rarely used since it requires the relative transfer function (RTF) vectors of all present speakers to treat them accordingly. This assumption is rather unrealistic and impractical, as it is considered cumbersome or even impossible to estimate the RTF vectors of two speakers simultaneously. In fact, not only estimating two RTF vectors at once, but even estimating one, i.e., the RTF vector of the target speaker, is an unsolved problem. Most RTF vector estimation methods proposed in literature [10–13] are not designed for the case of a second speaker being present in the acoustic scenario or if designed for this case, it is assumed that either all speakers or the dominant one are desired [11]. In [14], the spatial coherence (SC) estimator was investigated for the case of two coherent speakers, where it was shown that the presence of the interferer leads to a bias of the estimated RTF vector. A similar analysis is performed in this thesis for the covariance whitening (CW) estimator [10]. The results show that the estimated RTF vector is a weighted linear combination of the target and the interferer RTF vector, which depends on the multi-channel signal-to-interferer ratio (SIR). The consequence is that the dominant speaker is mostly preserved, while the other is suppressed. When both speakers are approximately equally loud, both speakers are cancelled equally much, which does not allow to actively suppress one speaker while preserving the other when considering a blind RTF vector estimation. This problem mostly arises from the inseparability of the activity of both speakers by means of a classical voice activity detection (VAD) as it is usually required for RTF vector estimation.

These findings motivate a major assumption made in this thesis: In hearing aid applications, it is often assumed that the direction of arrival (DoA) of the target is known (e.g., to the front) and hence, also an approximate (often anechoic) RTF

vector is known [15, 16]. In this thesis, this assumption is taken further and it is assumed that the target RTF vector for the hearing aid microphones (here referred to as the local microphone array (LMA)) is known, similarly to [17].

It has been shown in literature that incorporating external microphones (eMics) into the processing of an LMA can significantly improve the performance of speech enhancement algorithms [13, 16–20]. However, since the target RTF vector corresponding to the eMics cannot be assumed to be known as the eMics might be placed anywhere in the acoustic scene and their relative position is therefore unknown, their respective RTFs must be estimated. In [17], two different structures were proposed, which exploit the intermediate signal stages of a GSC applied to the LMA, referred to as the local GSC (L-GSC), steered by the known local target RTF vector. In this thesis, these structures are adopted and changed in their formulation such that not only noise but also the interferer is cancelled. Furthermore, alternative structures are proposed which allow for a detailed investigation of benefits or downsides coming with certain pre-processing operations. The four investigated structures are the GSC with external noise references type 1 (GSC-ENR-1), the GSC with external noise references type 2 (GSC-ENR-2), the GSC with external references (GSC-ER) and the GSC with external speech references (GSC-ESR). The first two structures use the estimated external RTF vector to complete a blocking matrix which allows for creating additional external noise-and-interferer references. The two structures differ in terms of the signals used for the RTF vector estimation and the filter optimization, which is done either jointly with the the local filter or cascaded. Another structure is the GSC-ER, which uses the output signal of the L-GSC to improve the RTF vector estimation and subsequently uses the output signal of the L-GSC together with the eMic signals in a joint MPDR beamformer. The GSC-ER can be regarded as a simplified version of the GSC-ESR. The difference of the GSC-ESR to the GSC-ER is that the GSC-ESR (proposed in [17]) pre-processes the eMic signals by means of the local noise-and-interferer references, aiming at minimizing correlated components between these signals. This ideally leads to an improved SIR in the pre-processed eMic signals, allowing for a better RTF vector estimation and a decoupling of noise and interferer reduction.

In the experimental evaluation of the considered algorithms, their performance in terms of noise and interferer reduction is evaluated using real-world recordings. It is shown that the GSC-ENR-1 and the GSC-ENR-2 are not suitable for joint noise and interferer reduction, as they strongly suffer from speech leakage and hence tar-

get speech cancellation due to imperfect RTF vector estimation. Also the GSC-ER does not perform well, especially in terms of interferer reduction, at low input SIRs, but allows for a clear investigation of the pre-processing of the eMic signals in the GSC-ESR. Especially in adverse conditions with a low input SIR, the GSC-ESR performs well and seems suitable for joint noise and interferer reduction. In case of a mismatch between the given RTF vector of the LMA and the true target RTF vector for the LMA, the GSC-ESR performs poorly at high SIRs, since speech leakage leads to speech cancellation in the pre-processed eMic signals. This can be deduced from the fact that the GSC-ER performs reasonably well in these conditions, i.e., is not affected that strongly by target speech cancellation effects. However, in conditions where noise - and particularly interferer - reduction is crucial, i.e., when the interferer is the dominant speech source, the pre-processing of the eMic signals is advantageous over the un-processed eMic signals and leads to good results in terms of noise and interferer reduction.

This thesis is structured as follows: In Section 2, the signal model and notation used throughout the thesis are introduced. In Section 3, the MVDR and the MPDR beamformer are introduced. In Section 4, firstly, the VAD and its consequences for the parameter estimation are discussed, secondly, the CW algorithm is presented in detail (Section 4.2) and a bias analysis is performed for the case of two coherent speakers (Section 4.3). The GSC structure is introduced in Section 5 and subsequently extended to incorporate the eMics in Section 6. In Section 7, the experimental evaluation is presented.To visualize the structure of this thesis, an overview scheme is shown in Fig. 1.1 (to be read from the center "Objective" following the arrows clockwise).

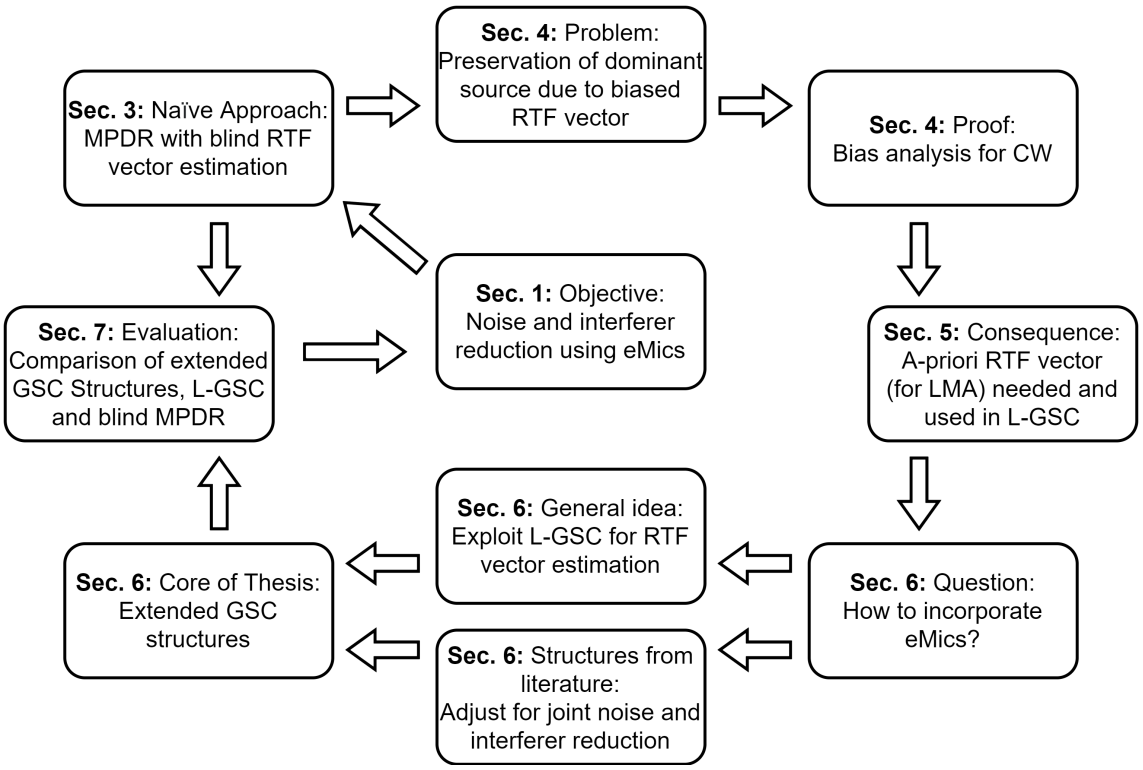

Fig. 1.1: Overview of contents and their connections. To be read from center following the arrows clockwise.

2 Signal Model and Notation

In this section, the signal model and the used notation are introduced. A configuration with M_a hearing aid microphones, i.e., the LMA, and M_e eMics, giving a total of $M = M_\mathrm{a} + M_\mathrm{e}$ microphones, is considered. The noisy input signal of the m-th microphone in the time-domain is denoted by $y_m(t)$ (with $m \in \{1, 2, \ldots, M\}$), where t is the discrete sample index. The frequency-domain representation of the complex-valued noisy input signal $Y_m(k, l)$ is obtained by applying the short-time Fourier transform (STFT) to $y_m(t)$, i.e.,

$$Y_m(k,l) = \sum_{t_\mathrm{d}=0}^{T_\mathrm{d}-1} y_m(lT_\mathrm{s} + t_\mathrm{d}) w(t_\mathrm{d}) \exp\left(\frac{-j2\pi k t_\mathrm{d}}{T_\mathrm{d}}\right), \tag{2.1}$$

with $w(t_\mathrm{d})$ a window function of length T_d which is the STFT length in samples. T_s denotes the frame shift in samples, t_d the sample index of the l-th frame and j the imaginary unit (i.e., $j^2 = -1$). In the following, all frequency bins are processed separately, as they are assumed to be independent of each other. For conciseness, the frequency-bin index k and the frame index l are omitted in the remainder of this thesis, wherever possible.

All microphones are stacked in the M-dimensional noisy input signal vector $\mathbf{y}$, where the first M_a entries correspond to the LMA, denoted as $\mathbf{y}_\mathrm{a}$, and the last M_e entries correspond to the eMics, denoted as $\mathbf{y}_\mathrm{e}$, respectively, i.e.,

$$\begin{aligned}\mathbf{y} &= [\, Y_1 \,,\, Y_2 \,,\, \ldots \,,\, Y_M \,]^T \\ &= [\, Y_{\mathrm{a},1} \,,\, Y_{\mathrm{a},2} \,,\, \ldots \,,\, Y_{\mathrm{a},M_\mathrm{a}} \,,\, Y_{\mathrm{e},1} \,,\, \ldots \,,\, Y_{\mathrm{e},M_\mathrm{e}} \,]^T \\ &= [\, \mathbf{y}_\mathrm{a}^T \,,\, \mathbf{y}_\mathrm{e}^T \,]^T\end{aligned} \tag{2.2}$$

where $\{\cdot\}^T$ denotes the transpose operator. In the considered acoustic scenario, where a target speaker, an interferer and noise are present, the noisy input signal in the m-th microphone can be decomposed into its components, assuming an additive signal model, i.e.,

$$Y_m = X_m + I_m + N_m \,, \tag{2.3}$$

where X_m denotes the target speech component, I_m denotes the interferer component and N_m denotes the noise component in the m-th microphone. Hence, for the noisy signal vector, it follows that

$$\mathbf{y} = \mathbf{x} + \underbrace{\mathbf{i} + \mathbf{n}}_{\mathbf{v}}, \tag{2.4}$$

where $\mathbf{x}$ is the target speech vector, $\mathbf{i}$ is the interferer vector and $\mathbf{n}$ is the noise vector, defined similarly to $\mathbf{y}$. The vector $\mathbf{v}$ denotes the total undesired signal component. Similarly to (2.2), the signal vectors $\mathbf{x}$, $\mathbf{i}$, $\mathbf{n}$ and $\mathbf{v}$ can be separated into their local and external parts.

For coherent sources as the target or the interferer, the signal components can be written in terms of their RTF vectors $\mathbf{h}$ and $\mathbf{b}$, respectively. The RTF vector relates the source's acoustic transfer function (ATF) of all microphones to a reference microphone. Here, the reference microphone is chosen to be the first of the LMA, without loss of generality, such that the target speech component vector $\mathbf{x}$ and interferer component vectors $\mathbf{i}$ can be written in terms of the target speech component X_1 and the interferer component I_1 in the first microphone, i.e.,

$$\mathbf{x} = \mathbf{h}X_1 \tag{2.5}$$

and

$$\mathbf{i} = \mathbf{b}I_1 \,. \tag{2.6}$$

The target RTF vector $\mathbf{h}$ can furthermore be decomposed into the local RTF vector $\mathbf{h}_\mathrm{a}$ (the first M_a entries) and $\mathbf{h}_\mathrm{e}$ containing the external RTFs (the last M_e entries), i.e.,

$$\begin{aligned} \mathbf{h} &= [\, \mathbf{h}_\mathrm{a}^T \,,\, \mathbf{h}_\mathrm{e}^T \,]^T \\ &= [\, 1 \,,\, H_{\mathrm{a},2} \,,\, \ldots \,,\, H_{\mathrm{a},M_\mathrm{a}} \,,\, H_{\mathrm{e},1} \,,\, \ldots \,,\, H_{\mathrm{e},M_\mathrm{e}} \,]^T \,, \end{aligned} \tag{2.7}$$

and similarly for the interferer RTF vector $\mathbf{b}$. Since the noise is not assumed to be a coherent source, it cannot be written in terms of an RTF vector.

At this point, a distinction can be made about the accessibility of a-priori knowledge about the RTF vector components. In hearing aid applications it is often assumed that the target DoA is known (e.g., the target speaker is located in the frontal direction of the hearing aid user) [15, 16]. Since the geometry of the hearing aids and hence the positioning of the microphones is usually known, the target RTF vector can be approximated by a steering vector for the target DoA. Taking this assumption even further, it is assumed that the target RTF vector for the LMA is known, i.e., the a-priori RTF vector for the LMA $\tilde{\mathbf{h}}_\mathrm{a}$ is assumed to be given. For the interferer, no such assumption is made, since its DoA is not as simple to anticipate as the target DoA. Also for the target RTFs corresponding to the eMics, no a-priori knowledge is given, since the eMics could be placed anywhere in the acoustic scene.

The $M \times M$-dimensional noisy covariance matrix is defined as

$$\mathbf{R}_y = \mathcal{E}\{\mathbf{y}\mathbf{y}^H\}\,, \tag{2.8}$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation value operator and $\{\cdot\}^H$ is the Hermitian transpose operator. Assuming that all signal components are uncorrelated with each other, the noisy covariance matrix can be written as

$$\mathbf{R}_y = \mathbf{R}_x + \underbrace{\mathbf{R}_n + \mathbf{R}_i}_{\mathbf{R}_v}\,, \tag{2.9}$$

where the rank-1 target speech covariance matrix $\mathbf{R}_x$, the rank-1 interferer covariance matrix $\mathbf{R}_i$, the noise covariance matrix $\mathbf{R}_n$, and the undesired covariance matrix $\mathbf{R}_v$ are defined similarly to $\mathbf{R}_y$, i.e.,

$$\begin{aligned}\mathbf{R}_x &= \mathcal{E}\{\mathbf{x}\mathbf{x}^H\} \\ &= \phi_x \mathbf{h}\mathbf{h}^H\,,\end{aligned} \tag{2.10}$$

$$\begin{aligned}\mathbf{R}_i &= \mathcal{E}\{\mathbf{i}\mathbf{i}^H\} \\ &= \phi_i \mathbf{b}\mathbf{b}^H\,,\end{aligned} \tag{2.11}$$

and

$$\begin{aligned}\mathbf{R}_n &= \mathcal{E}\{\mathbf{n}\mathbf{n}^H\} \\ &= \phi_n \mathbf{\Gamma}\,,\end{aligned} \tag{2.12}$$

where the target, interferer and noise power spectral densities (PSDs) in the first microphone are defined as $\phi_x = \mathcal{E}\{|X_1|^2\}$, $\phi_i = \mathcal{E}\{|I_1|^2\}$ and $\phi_n = \mathcal{E}\{|N_1|^2\}$, respectively and Γ is the normalized noise covariance matrix, characterizing the spatial properties of the noise field, for which no further explicit assumptions are made here. In the following, it is assumed that the target speech and interferer PSDs ϕ_x and ϕ_i are strongly time varying, i.e., these sources are considered to be spectro-temporally non-stationary, while the noise PSD ϕ_n is assumed to be relatively constant over time. These properties are important for a detection of speech sources via a VAD algorithm discussed further in Section 4.1.

When only the signal vector of the LMA $\mathbf{y}_\mathrm{a}$ is considered, the $M_\mathrm{a} \times M_\mathrm{a}$-dimensional covariance matrix $\mathbf{R}_{y,\mathrm{a}}$ is obtained similarly to (2.8) and can be written as

$$\mathbf{R}_{y,\mathrm{a}} = \mathbf{E}_\mathrm{a} \mathbf{R}_y \mathbf{E}_\mathrm{a}^T \tag{2.13}$$

where

$$\mathbf{E}_{\mathrm{a}} = [\mathbf{I}_{M_{\mathrm{a}} \times M_{\mathrm{a}}}, \, \mathbf{0}_{M_{\mathrm{a}} \times M_{\mathrm{e}}}] \tag{2.14}$$

is a selection matrix for the LMA signals only. $\mathbf{R}_{x,\mathrm{a}}$, $\mathbf{R}_{i,\mathrm{a}}$ and $\mathbf{R}_{n,\mathrm{a}}$ are defined similarly to (2.13).
All filtering operations in this thesis are linear and can be defined by means of a complex-valued filter vector $\mathbf{f}_{\mathrm{lin}}$, which is applied to the input signal vector $\mathbf{y}$, i.e.,

$$Z = \mathbf{f}_{\mathrm{lin}}^{H} \mathbf{y} \, . \tag{2.15}$$

To define the objective performance measures in Section 2.1, which are used to investigate the influence of a filtering operation as defined in (2.15) on the separate signal components, oracle knowledge about the input signal component vectors $\mathbf{x}$, $\mathbf{i}$ and $\mathbf{n}$ is required. For the so-called shadow filtered signals, the outputs X_{out}, I_{out} and N_{out} are given by applying the same linear filter $\mathbf{f}_{\mathrm{lin}}$ to the separate input component vectors, i.e.,

$$X_{\mathrm{out}} = \mathbf{f}_{\mathrm{lin}}^{H} \mathbf{x} \, , \tag{2.16}$$

$$I_{\mathrm{out}} = \mathbf{f}_{\mathrm{lin}}^{H} \mathbf{i} \, , \tag{2.17}$$

and

$$N_{\mathrm{out}} = \mathbf{f}_{\mathrm{lin}}^{H} \mathbf{n} \, , \tag{2.18}$$

respectively.

2.1 Objective Performance Measures

In this thesis, purely technical objective performance measures (in contrast to psycho-acoustically motivated objective measures) are used to quantify the performance of the considered algorithms. The measures of interest are the noise and interferer reduction, the amount of distortions introduced by a filtering operation and the accuracy of the RTF vector estimation. Firstly, the narrow-band measures, i.e., per frequency-bin, are introduced, which only occur in theoretical derivations and which are therefore defined on a linear scale. The frequency-bin index k is included here to stress the difference between narrow- and broad-band measures. The narrow-band signal-to-noise ratio (SNR) at the input, i.e., in the reference microphone, is given by

$$\mathrm{SNR}(k) = \frac{\phi_x(k)}{\phi_n(k)} \, , \tag{2.19}$$

the narrow-band SIR at the input is given by

$$\mathrm{SIR}(k) = \frac{\phi_x(k)}{\phi_i(k)}\,, \tag{2.20}$$

and the narrow-band signal-to-interferer-and-noise ratio (SINR) at the input is given by

$$\mathrm{SINR}(k) = \frac{\phi_x(k)}{\phi_i(k) + \phi_n(k)}\,. \tag{2.21}$$

Precisely speaking, the narrow-band measures defined here do not quantify the performance of a filtering operation since (2.19) - (2.21) are merely defined using the input parameters. These metrics rather characterize the input signals and occur in theoretical derivations in this thesis. In principle, however, they could also be defined on the shadow filtered signals to characterize the output of a filter.

In the following, the broad-band measures are introduced, which are used for the evaluation of the performance of different processing structures in Section 7. In contrast to the narrow-band measures, broad-band measures, do not include a frequency bin index and are directly defined on a logarithmic scale and are therefore given in dB.

To compute the broad-band measures, the shadow filtered time-domain signals are used, meaning that an inverse short-time Fourier transform (ISTFT) is applied to the shadow filtered signals in (2.16) - (2.18) to obtain the shadow filtered time-domain signals $x_{\mathrm{out}}(t)$, $i_{\mathrm{out}}(t)$ and $n_{\mathrm{out}}(t)$, respectively. The input SNR, SIR and SINR are defined by the time-domain input signals in the first microphone, i.e.,

$$\mathrm{SNR}_{\mathrm{in}} = 20\log_{10}\left(\frac{\mathrm{rms}(x_1(t))}{\mathrm{rms}(n_1(t))}\right)\,, \tag{2.22}$$

$$\mathrm{SIR}_{\mathrm{in}} = 20\log_{10}\left(\frac{\mathrm{rms}(x_1(t))}{\mathrm{rms}(i_1(t))}\right)\,, \tag{2.23}$$

and

$$\mathrm{SINR}_{\mathrm{in}} = 20\log_{10}\left(\frac{\mathrm{rms}(x_1(t))}{\mathrm{rms}(i_1(t) + n_1(t))}\right)\,, \tag{2.24}$$

where $\mathrm{rms}(s(t))$ denoted the root-mean-square value of a time sequence $s(t)$ of samples, i.e.

$$\mathrm{rms}(s(t)) = \sqrt{\frac{1}{L}\sum_{t=1}^{L} s(t)^2}\,, \tag{2.25}$$

with L the number of samples over which the root-mean-square value is calculated. The output SNR, SIR and SINR are defined via the shadow filtered signals, i.e.,

$$\mathrm{SNR}_{\mathrm{out}} = 20\log_{10}\left(\frac{\mathrm{rms}(x_{\mathrm{out}}(t))}{\mathrm{rms}(n_{\mathrm{out}}(t))}\right), \tag{2.26}$$

$$\mathrm{SIR}_{\mathrm{out}} = 20\log_{10}\left(\frac{\mathrm{rms}(x_{\mathrm{out}}(t))}{\mathrm{rms}(i_{\mathrm{out}}(t))}\right), \tag{2.27}$$

and

$$\mathrm{SINR}_{\mathrm{out}} = 20\log_{10}\left(\frac{\mathrm{rms}(x_{\mathrm{out}}(t))}{\mathrm{rms}(i_{\mathrm{out}}(t) + n_{\mathrm{out}}(t))}\right). \tag{2.28}$$

To assess the performance of an algorithm, the difference of in- and output is taken into account, leading to the definition of the SNR improvement ΔSNR, the SIR improvement ΔSIR and the SINR improvement ΔSINR, i.e.,

$$\Delta\mathrm{SNR} = \mathrm{SNR}_{\mathrm{out}} - \mathrm{SNR}_{\mathrm{in}}, \tag{2.29}$$

$$\Delta\mathrm{SIR} = \mathrm{SIR}_{\mathrm{out}} - \mathrm{SIR}_{\mathrm{in}}, \tag{2.30}$$

and

$$\Delta\mathrm{SINR} = \mathrm{SINR}_{\mathrm{out}} - \mathrm{SINR}_{\mathrm{in}}, \tag{2.31}$$

respectively. Since these measures only assess how much interferer/noise is suppressed by a filtering operation relative to the effect which the same filtering operation has on the target speech, one important factor is neglected when only regarding these metrics: speech distortion (SD) is another energy based measure which allows to quantify the distortions applied to the target speech. The broad-band SD is defined as

$$\mathrm{SD} = 20\log_{10}\left(\frac{\mathrm{rms}(x_{\mathrm{out}}(t))}{\mathrm{rms}(x_1(t))}\right). \tag{2.32}$$

An SD value of 0 dB indicates perfect target speech preservation by a filtering operation, negative values indicate cancellation of target speech and positive values indicate a gain applied to the target speech. A similar measure is defined using the interferer or the noise component. In the latter case, it is referred to as the amount of noise reduction (NR), given by

$$\mathrm{NR} = 20 \log_{10} \left(\frac{\mathrm{rms}(n_{\mathrm{out}}(t))}{\mathrm{rms}(n_1(t))} \right) . \tag{2.33}$$

Another measure considered in this thesis is the Hermitian angle, which is considered to access the quality of an RTF vector estimate [21], since it represents the similarity of two complex valued vectors (e.g., of an estimated RTF vector and the true RTF vector). The narrow-band Hermitian angle $\Theta_{\mathbf{p},\mathbf{q}}(k)$ between two complex-valued vectors $\mathbf{p}$ and $\mathbf{q}$ is defined as

$$\Theta_{\mathbf{p},\mathbf{q}}(k) = \arccos \left(\frac{|\mathbf{p}^H \mathbf{q}|}{\|\mathbf{p}\|_2 \, \|\mathbf{q}\|_2} \right) , \tag{2.34}$$

where $\|\mathbf{p}\|_2 = \sqrt{\mathbf{p}^H \mathbf{p}}$ is the 2-norm of $\mathbf{p}$. Note that usually the Hermitian angle is computed in every frequency-bin and subsequently averaged over all frequency to condense the measure into a scalar value.

3 MVDR and MPDR Beamforming

In this section, the so-called minimum variance distortionless response (MVDR) and minimum power distortionless response (MPDR) beamformer are introduced. Both structures are closely related and only differ in their objective function [6, 22, 23]. While the MVDR beamformer aims at minimizing the noise power at the output, the MPDR beamformer aims at minimizing the total power at the output Z, which is defined similarly to the linear filtering operation in (2.15), i.e.,

$$Z = \mathbf{w}^H \mathbf{y} \,. \tag{3.1}$$

Besides aiming at minimizing the objective function, both beamformers aim at preserving all sounds associated with the RTF vector estimate $\hat{\mathbf{h}}$, where $\{\hat{\cdot}\}$ denotes an estimate of a quantity. The complex-valued M-dimensional filter vector $\mathbf{w}$ is obtained by solving the constrained optimization problem, here formulated for the MVDR beamformer, i.e.,

$$\min_{\mathbf{w}} \mathbf{w}^H \hat{\mathbf{R}}_n \mathbf{w} \,, \qquad \text{s.t. } \mathbf{w}^H \hat{\mathbf{h}} = 1 \,, \tag{3.2}$$

where the optimization problem of the MPDR beamformer uses an estimate of the noisy covariance matrix $\mathbf{R}_y$ instead of an estimate of the noise covariance matrix $\mathbf{R}_n$. The closed form solution to (3.2) is given by

$$\mathbf{w}_{\text{MVDR}} = \frac{\hat{\mathbf{R}}_n^{-1} \hat{\mathbf{h}}}{\hat{\mathbf{h}}^H \hat{\mathbf{R}}_n^{-1} \hat{\mathbf{h}}} \,, \tag{3.3}$$

for the MVDR beamformer and respectively

$$\mathbf{w}_{\text{MPDR}} = \frac{\hat{\mathbf{R}}_y^{-1} \hat{\mathbf{h}}}{\hat{\mathbf{h}}^H \hat{\mathbf{R}}_y^{-1} \hat{\mathbf{h}}} \,, \tag{3.4}$$

for the MPDR beamformer. The filter vectors of the MVDR and MPDR only rely on an estimate of the noise covariance matrix $\mathbf{R}_n$ or respectively the noisy covariance matrix $\mathbf{R}_y$ and an estimate of the target RTF vector $\mathbf{h}$. Typical estimators for these quantities, their limitations in the case of two coherent speech sources and the consequences for blind beamforming algorithms are discussed in Section 4. For the case of the CW method used for RTF vector estimation, it can be shown that the MVDR and the MPDR beamformer are identical (see Appendix A).

The processing scheme of the beamformer $\mathbf{w}$, which can either be implemented as an MVDR or MPDR beamformer, respectively, is shown in Fig. 3.1. The RTF estimation block represents a blind, data driven RTF vector estimation using all

available signals.

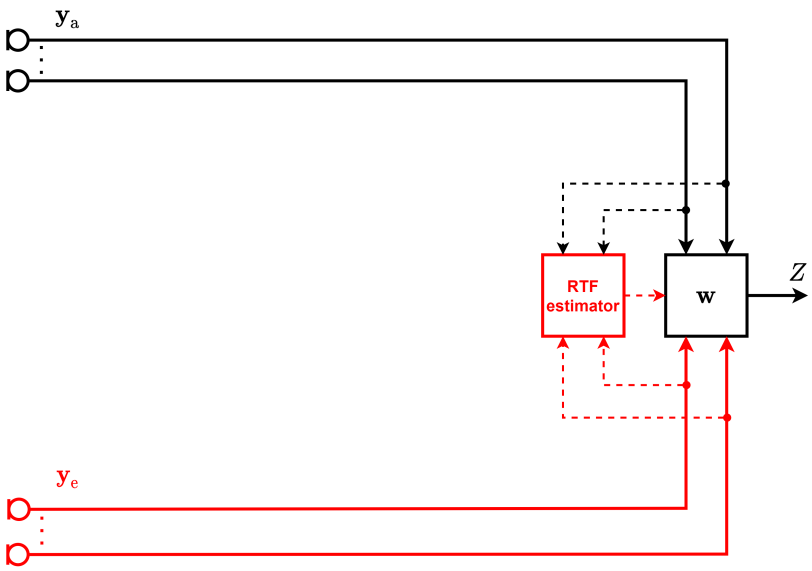

Fig. 3.1: Schematic of blind MVDR/MPDR beamforming using all local and external microphone signals, including an RTF estimation block.

In the remainder of this thesis, only the MPDR is considered, since (as mentioned above) both beamformers are identical if the CW method is used for the RTF vector estimation. Furthermore, the MPDR beamformer as depicted in Fig. 3.1 is always treated as a blind algorithm, using an estimate of the RTF vector, estimated on all available signals. This property is stressed here to distinguish between blind and informed beamformers (using the a-priori RTF vector $\tilde{\mathbf{h}}_\mathrm{a}$), like the L-GSC, which is introduced in Section 5.

4 Standard Estimators for Model Parameters

In this section, limitations of state-of-the-art algorithms for the detection of different signal components, the so-called voice activity detection (VAD), and RTF vector estimators are discussed. These algorithms are required to estimate the quantities needed for many beamformers, like the MPDR beamformer discussed in Section 3. A VAD algorithm is often used to estimate the noisy and noise covariance matrices, which are used to estimate the target RTF vector. However, most state-of-the-art VAD algorithms are not designed for the case of two coherent speakers. Hence, their applicability in the considered acoustic scenario with a target and an interferer is limited. In the following, the limitations and consequences of VAD algorithms (Section 4.1) and blind RTF vector estimation using the CW method (Sections 4.2 - 4.5) are presented.

4.1 Voice Activity Detection and Covariance Matrix Estimation

State-of-the-art voice activity detection (VAD) algorithms like the speech presence probability (SPP) estimator in [24] are designed to detect the presence of all speech sources in a noisy background. These algorithms are usually either energy based (like the SPP) or they exploit speech properties like periodicity [25]. Most classical VAD algorithms cannot distinguish between different speakers. Hence, if at least one speaker is active, a binary VAD will detect speech, as depicted in Fig. 4.1.

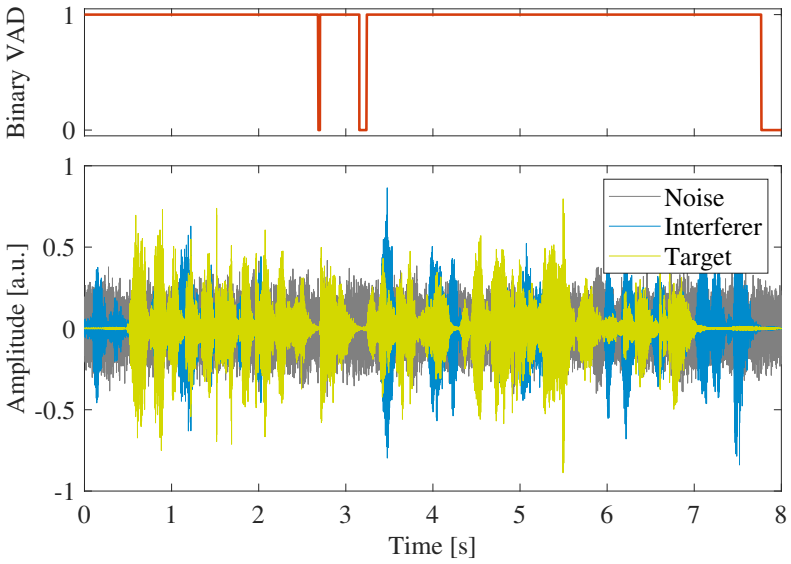

Fig. 4.1: Effects of two speakers on an energy based VAD algorithm. Upper panel: Corresponding binary VAD (red). Lower panel: Two speech sources, target (green) and interferer (blue) with background noise (gray).

The problem arising from the detection of both speech sources becomes clear, when the recursive, VAD-based estimation of the covariance matrices is considered. Where the estimated covariance matrices $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{R}}_n$ are given by

$$\hat{\mathbf{R}}_y(k,l) = \begin{cases} \alpha_y \hat{\mathbf{R}}_y(k,l-1) + (1-\alpha_y)\mathbf{y}(k,l)\mathbf{y}(k,l)^H & \text{, in speech plus noise frames} \\ \hat{\mathbf{R}}_y(k,l-1) & \text{, in noise only frames} \end{cases} \tag{4.1}$$

and

$$\hat{\mathbf{R}}_n(k,l) = \begin{cases} \alpha_n \hat{\mathbf{R}}_n(k,l-1) + (1-\alpha_n)\mathbf{n}(k,l)\mathbf{n}(k,l)^H & \text{, in noise only frames} \\ \hat{\mathbf{R}}_n(k,l-1) & \text{, in speech plus noise frames} \end{cases}, \tag{4.2}$$

where α_y and α_n are smoothing constants, corresponding to the time constants τ_y and τ_n, for which the relations

$$\alpha_y = e^{-\frac{T_\mathrm{d}-T_\mathrm{s}}{f_\mathrm{s}\tau_y}} \tag{4.3}$$

and

$$\alpha_n = e^{-\frac{T_\mathrm{d}-T_\mathrm{s}}{f_\mathrm{s}\tau_n}} \tag{4.4}$$

hold, where f_s denotes the sampling frequency. Note that the estimate $\hat{\mathbf{R}}_y$ of $\mathbf{R}_y$ might not satisfy all assumptions made for the signal model in (2.9), while $\mathbf{R}_y$ fulfills all stated properties by definition.

From the update rules in (4.1) and (4.2), it becomes clear that the estimated noise covariance matrix $\hat{\mathbf{R}}_n$ only contains information about the noise and that the estimate of the noisy covariance matrix $\hat{\mathbf{R}}_y$ contains all acoustic sources, i.e. target, noise and interferer. The undesired covariance matrix $\mathbf{R}_v$ cannot be accessed using VAD-driven update rules, since the VAD cannot distinguish between the target speaker and the interferer.

Note that the VAD-based estimation method for $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{R}}_n$, are often used in practice, but not in this thesis. The focus here is the investigation of the potential of different processing strategies and not the entire system including a blind VAD algorithm and covariance matrix estimation. Hence, in the evaluation, no VAD algorithm is included and the covariance matrices are obtained using oracle knowledge, since blind estimation might introduce estimation errors that make the results more difficult to interpret. This can be justified by the fact that most VAD algorithms are prone to errors at low SNRs and their performance also strongly depends on the SIR. To eliminate these factors of misestimation and to make the results as

comparable as possible, a different implementation without a VAD is chosen here (see Section 7).

4.2 RTF Vector Estimation Using Covariance Whitening

In this section, one RTF estimator, namely the covariance whitening (CW) method from [10], is investigated, which has been thoroughly analyzed in [26] and has been used in various applications [11, 27, 28]. Firstly, the general steps of this method are shown. Subsequently, the CW method is investigated for the ideal case of only one speaker and no estimation errors. Lastly, a bias analysis is performed for the case of two (for a VAD algorithm) indistinguishable coherent speakers (Sections 4.3 - 4.5). As discussed above, it is assumed that noisy covariance matrix $\mathbf{R}_y$ and the noise covariance matrix $\mathbf{R}_n$ are accessible (by means of a VAD), while the undesired covariance matrix $\mathbf{R}_v$ is not.

For the whitening operation, which aims at reducing the influence of the noise component on the estimated noisy covariance matrix, the matrix $\hat{\mathbf{R}}_n$ is written in terms of the lower triangular matrix $\hat{\mathbf{R}}_n^{1/2}$ using a square-root decomposition (e.g., Cholesky decomposition), i.e.,

$$\hat{\mathbf{R}}_n = \hat{\mathbf{R}}_n^{1/2}\hat{\mathbf{R}}_n^{H/2}\,, \tag{4.5}$$

such that $\hat{\mathbf{R}}_n^{-1/2}\hat{\mathbf{R}}_n\hat{\mathbf{R}}_n^{-H/2} = \mathbf{I}_{M\times M}$. The whitening operation is then applied to the noisy covariance matrix $\hat{\mathbf{R}}_y$, which yields the whitened covariance matrix $\hat{\mathbf{R}}_y^{\mathrm{w}}$, i.e.,

$$\hat{\mathbf{R}}_y^{\mathrm{w}} = \hat{\mathbf{R}}_n^{-1/2}\hat{\mathbf{R}}_y\hat{\mathbf{R}}_n^{-H/2}\,. \tag{4.6}$$

To estimate the whitened RTF vector, a rank-1 approximation is done by performing an eigenvalue decomposition (EVD) on $\hat{\mathbf{R}}_y^{\mathrm{w}}$ and taking the principal eigenvector, i.e., the eigenvector corresponding to the largest eigenvalue, that is

$$\mathbf{v}_{\max} = \mathcal{P}\{\hat{\mathbf{R}}_y^{\mathrm{w}}\}\,, \tag{4.7}$$

where $\mathcal{P}\{\cdot\}$ denotes the principal eigenvector operator. To obtain the RTF vector estimate $\hat{\mathbf{h}}$, $\mathbf{v}_{\max}$ in (4.7) is de-whitened and normalized to the entry corresponding to the first microphone (selected by means of the selection vector $\mathbf{e}_1$, containing only zeros, except for the entry corresponding to the first microphone), i.e.,

$$\hat{\mathbf{h}}^{\mathrm{CW}} = \frac{\hat{\mathbf{R}}_n^{1/2}\mathbf{v}_{\max}}{\mathbf{e}_1^T\hat{\mathbf{R}}_n^{1/2}\mathbf{v}_{\max}}\,. \tag{4.8}$$

Up to this point, a very general formulation of the CW method was used, which relies on the estimates $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{R}}_n$. So far, the signal model was not explicitly used to motivate the steps done in CW. However, in terms of the model parameters defined in Section 2, it can be distinguished between the case of one coherent speaker (which is the underlying assumption of the CW method) or the case of two coherent speakers (see Section 4.3).
When now considering the case that only one coherent speaker is present (i.e., $\phi_i = 0$), which is the case for which the CW estimator was proposed, the signal model in (2.9) reads

$$\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_n \,, \tag{4.9}$$

where $\mathbf{R}_x$ is assumed to be a rank-1 matrix as defined in (2.10). The whitened noisy covariance matrix in (4.6) can be written in terms of the model parameters, that is

$$\begin{aligned} \mathbf{R}_y^{\mathrm{w}} &= \mathbf{R}_n^{-1/2}(\mathbf{R}_x + \mathbf{R}_n)\mathbf{R}_n^{-H/2} \\ &= \mathbf{R}_n^{-1/2}\mathbf{R}_x\mathbf{R}_n^{-H/2} + \mathbf{I}_{M\times M} \\ &= \phi_x \underbrace{\mathbf{R}_n^{-1/2}\mathbf{h}}_{\mathbf{v}_{\max}} \underbrace{\mathbf{h}^H\mathbf{R}_n^{-H/2}}_{\mathbf{v}_{\max}^H} + \mathbf{I}_{M\times M} \,. \end{aligned} \tag{4.10}$$

It should be noted that the addition of an identity matrix does not affect the eigenvectors of $\mathbf{R}_y^{\mathrm{w}}$, but increases all eigenvalues by 1. Therefore, as indicated in (4.10), the principal eigenvector $\mathbf{v}_{\max}$ is a scaled version of $\mathbf{R}_n^{-1/2}\mathbf{h}$. Hence, in the case of only one coherent speech source, the estimated RTF vector equals the true target RTF vector (when using model parameters instead of estimated quantities), i.e., $\hat{\mathbf{h}}^{\mathrm{CW}}$ becomes $\mathbf{R}_n^{1/2}\mathbf{R}_n^{-1/2}\mathbf{h} = \mathbf{h}$.
Note that this is the case for which the CW method is designed. Inherently, it is assumed that the covariance matrices of all undesired sources can be estimated, i.e., that even in the case of multiple speakers, the undesired covariance matrix $\mathbf{R}_v$ can be estimated, as in [26].
In the following section, a derivation is presented which shows the influence of an interfering speaker, i.e., where $\mathbf{R}_v$ is not available, on the RTF vector estimation using the CW method.

4.3 Bias Analysis of Covariance Whitening for Two Coherent Speakers

As discussed in Section 4.1, it is considered impossible to separate two speakers using a VAD. The consequences of a second speaker being present and indistinguishable for a VAD, to the best of the author's knowledge, have not been analyzed in theory so far. In this section, the influence of a second coherent speaker on the RTF estimation using CW is investigated theoretically. Furthermore, the theoretical findings are supported by simulations using artificial data in Section 4.5.

For this bias analysis, the interferer is included in the signal model, i.e., as defined in (2.9). If it is now assumed that the whitening is still done with $\mathbf{R}_n$, but $\mathbf{R}_y$ includes the interferer covariance matrix, the matrix $\mathbf{R}_y^{\mathrm{w}}$ becomes

$$\begin{aligned}\mathbf{R}_y^{\mathrm{w}} &= \mathbf{R}_n^{-1/2}(\mathbf{R}_x + \mathbf{R}_i + \mathbf{R}_n)\mathbf{R}_n^{-H/2} \\ &= \mathbf{R}_n^{-1/2}(\mathbf{R}_x + \mathbf{R}_i)\mathbf{R}_n^{-H/2} + \mathbf{I}_{M\times M} \\ &= \mathbf{R}_n^{-1/2}(\phi_x \mathbf{h}\mathbf{h}^H + \phi_i \mathbf{b}\mathbf{b}^H)\mathbf{R}_n^{-H/2} + \mathbf{I}_{M\times M}\,.\end{aligned} \tag{4.11}$$

Applying an EVD to (4.11), the identity matrix $\mathbf{I}_{M\times M}$ can be neglected and therefore be cancelled out by subtraction. Hence, the eigenvectors of the rank-2 matrix

$$\begin{aligned}\overline{\mathbf{R}} &= \mathbf{R}_y^{\mathrm{w}} - \mathbf{I}_{M\times M} \\ &= \phi_x \mathbf{R}_n^{-1/2}\mathbf{h}\mathbf{h}^H\mathbf{R}_n^{-H/2} + \phi_i \mathbf{R}_n^{-1/2}\mathbf{b}\mathbf{b}^H\mathbf{R}_n^{-H/2} \\ &= \overline{\phi}_x \overline{\mathbf{h}\mathbf{h}}^H + \overline{\phi}_i \overline{\mathbf{b}\mathbf{b}}^H\end{aligned} \tag{4.12}$$

can be considered instead of $\mathbf{R}_y^{\mathrm{w}}$, where the normalized whitened RTF vectors $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$ are defined as

$$\overline{\mathbf{h}} = \frac{\mathbf{R}_n^{-1/2}\mathbf{h}}{\left\|\mathbf{R}_n^{-1/2}\mathbf{h}\right\|_2}\,, \tag{4.13}$$

and

$$\overline{\mathbf{b}} = \frac{\mathbf{R}_n^{-1/2}\mathbf{b}}{\left\|\mathbf{R}_n^{-1/2}\mathbf{b}\right\|_2}\,, \tag{4.14}$$

respectively, such that they have a norm of 1 by definition, i.e., $\left\|\overline{\mathbf{h}}\right\|_2 = \left\|\overline{\mathbf{b}}\right\|_2 = 1$, and the target and interferer PSDs $\overline{\phi}_x$ and $\overline{\phi}_i$ include the normalization factors from (4.13) and (4.14), i.e.,

$$\overline{\phi}_x = \phi_x \left\| \mathbf{R}_n^{-1/2} \mathbf{h} \right\|_2^2 , \tag{4.15}$$

and

$$\overline{\phi}_i = \phi_i \left\| \mathbf{R}_n^{-1/2} \mathbf{b} \right\|_2^2 . \tag{4.16}$$

The PSDs $\overline{\phi}_x$ and $\overline{\phi}_i$ can be interpreted as multi-channel PSDs in the whitened domain.
From the structure of the matrix $\overline{\mathbf{R}}$ in (4.12) it can be seen that the principal eigenvector of $\overline{\mathbf{R}}$ is a linear combination of the vectors $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$, weighted with the complex-valued factors α_x and α_i, respectively, as in

$$\mathbf{v}_{\max} = \alpha_x \overline{\mathbf{h}} + \alpha_i \overline{\mathbf{b}} . \tag{4.17}$$

In [11], the weighting factors α_x and α_i were written in terms of the inner product of the principal eigenvector $\mathbf{v}_{\max}$ and the whitened RTFs vectors $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$, respectively. In the following, a formulation is found which does not explicitly depend on $\mathbf{v}_{\max}$, which allows for a better interpretability, since it directly depends on the input characteristics of the two speech sources.
For the principal eigenvector of $\overline{\mathbf{R}}$, it holds that

$$\overline{\mathbf{R}} \mathbf{v}_{\max} = \lambda_1 \mathbf{v}_{\max} , \tag{4.18}$$

where λ_1 is the principal, i.e., the largest, eigenvalue of $\overline{\mathbf{R}}$. Using the definition of $\overline{\mathbf{R}}$ in (4.12) and the definition of $\mathbf{v}_{\max}$ in (4.17) in the eigenvalue equation in (4.18), it yields

$$\alpha_x \overline{\phi}_x (\overline{\mathbf{h}}^H \overline{\mathbf{h}}) \overline{\mathbf{h}} + \alpha_x \overline{\phi}_i (\overline{\mathbf{b}}^H \overline{\mathbf{h}}) \overline{\mathbf{b}} + \alpha_i \overline{\phi}_x (\overline{\mathbf{h}}^H \overline{\mathbf{b}}) \overline{\mathbf{h}} + \alpha_i \overline{\phi}_i (\overline{\mathbf{b}}^H \overline{\mathbf{b}}) \overline{\mathbf{b}} = \lambda_1 \alpha_x \overline{\mathbf{h}} + \lambda_1 \alpha_i \overline{\mathbf{b}} . \tag{4.19}$$

Explicitly assuming linear independence of $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$, meaning that these two vectors form a basis of a complex-valued two dimensional subspace, (4.19) can be separated into two equations, since all terms in $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$ can be separated. Therefore the two relations

$$\alpha_x \overline{\phi}_x (\overline{\mathbf{h}}^H \overline{\mathbf{h}}) + \alpha_i \overline{\phi}_x (\overline{\mathbf{h}}^H \overline{\mathbf{b}}) = \lambda_1 \alpha_x \tag{4.20}$$

and

$$\alpha_x \overline{\phi}_i (\overline{\mathbf{b}}^H \overline{\mathbf{h}}) + \alpha_i \overline{\phi}_i (\overline{\mathbf{b}}^H \overline{\mathbf{b}}) = \lambda_1 \alpha_i \tag{4.21}$$

are obtained, where $(\overline{\mathbf{h}}^H\overline{\mathbf{h}}) = (\overline{\mathbf{b}}^H\overline{\mathbf{b}}) = 1$. Reformulating (4.20) and (4.21) in terms of λ_1 and setting them equal gives

$$\overline{\phi}_x \left(1 + \overline{\mathbf{h}}^H\overline{\mathbf{b}}\frac{\alpha_i}{\alpha_x}\right) = \overline{\phi}_i \left(1 + \overline{\mathbf{b}}^H\overline{\mathbf{h}}\frac{\alpha_x}{\alpha_i}\right) . \tag{4.22}$$

Defining the weighting ratio r as

$$r = \frac{\alpha_x}{\alpha_i} , \tag{4.23}$$

and rearranging (4.22) to 0, the quadratic equation

$$r^2 + r\frac{\overline{\phi}_x - \overline{\phi}_i}{\overline{\phi}_i\overline{\mathbf{b}}^H\overline{\mathbf{h}}} - \frac{\overline{\phi}_x\overline{\mathbf{h}}^H\overline{\mathbf{b}}}{\overline{\phi}_i\overline{\mathbf{b}}^H\overline{\mathbf{h}}} = 0 \tag{4.24}$$

is obtained, which can be solved for r, that is

$$r = \frac{\overline{\phi}_x - \overline{\phi}_i}{2\overline{\phi}_i\overline{\mathbf{b}}^H\overline{\mathbf{h}}} \pm \sqrt{\left(\frac{\overline{\phi}_x - \overline{\phi}_i}{2\overline{\phi}_i\overline{\mathbf{b}}^H\overline{\mathbf{h}}}\right)^2 + \frac{\overline{\phi}_x\overline{\mathbf{h}}^H\overline{\mathbf{b}}}{\overline{\phi}_i\overline{\mathbf{b}}^H\overline{\mathbf{h}}}} . \tag{4.25}$$

By splitting up the inner products $\overline{\mathbf{h}}^H\overline{\mathbf{b}}$ and $\overline{\mathbf{b}}^H\overline{\mathbf{h}}$ into their magnitude and phase, respectively, i.e.,

$$\overline{\mathbf{h}}^H\overline{\mathbf{b}} = |\overline{\mathbf{h}}^H\overline{\mathbf{b}}|e^{j\angle\overline{\mathbf{h}}^H\overline{\mathbf{b}}} , \tag{4.26}$$

where $\angle$ denotes the phase angle of a complex number, the ratio r in (4.25) can be rewritten as

$$r = e^{j\angle\overline{\mathbf{h}}^H\overline{\mathbf{b}}}\frac{\overline{\phi}_x - \overline{\phi}_x}{2\overline{\phi}_i|\overline{\mathbf{b}}^H\overline{\mathbf{h}}|} \pm e^{j\angle\overline{\mathbf{h}}^H\overline{\mathbf{b}}}\sqrt{\left(\frac{\overline{\phi}_x - \overline{\phi}_i}{2\overline{\phi}_i|\overline{\mathbf{b}}^H\overline{\mathbf{h}}|}\right)^2 + \frac{\overline{\phi}_x}{\overline{\phi}_i}} . \tag{4.27}$$

Furthermore, the definition of the SIR in (2.20) can be considered and transferred to the whitened domain, which reads

$$\overline{\mathrm{SIR}} = \frac{\overline{\phi}_x}{\overline{\phi}_i} . \tag{4.28}$$

Note that in (4.28), it is not only the SIR in the whitened domain, but rather an SIR value averaged over channels, since $\overline{\phi}_x$ and $\overline{\phi}_i$ contain the norms of the whitened RTF vectors.

The Hermitian angle defined in (2.34) is considered, such that the absolute value of the inner product $|\overline{\mathbf{b}}^H\overline{\mathbf{h}}|$ can be replaced by it, i.e.,

$$|\overline{\mathbf{b}}^H \overline{\mathbf{h}}| = \left\|\overline{\mathbf{b}}\right\|_2 \left\|\overline{\mathbf{h}}\right\|_2 \cos\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}}\,, \tag{4.29}$$

where by definition $\left\|\overline{\mathbf{h}}\right\|_2 = \left\|\overline{\mathbf{b}}\right\|_2 = 1$. Using (4.28) and (4.29), the ratio r in (4.27) becomes

$$r = e^{j\angle \overline{\mathbf{h}}^H \overline{\mathbf{b}}} \frac{\overline{\mathrm{SIR}} - 1}{2\cos\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}}} \pm e^{j\angle \overline{\mathbf{h}}^H \overline{\mathbf{b}}} \sqrt{\left(\frac{\overline{\mathrm{SIR}} - 1}{2\cos\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}}}\right)^2 + \overline{\mathrm{SIR}}}\,. \tag{4.30}$$

The weighting ratio r, offers two solutions to the weighting in (4.17), which result from the fact that so far all steps are valid for both non-zero eigenvalues, and not just, as in the discussed case, for the principal eigenvalue. The property of the eigenvalue in (4.18) and (4.19) to correspond to the largest eigenvalue, has not been used explicitly in this derivation. Therefore, the solution for r contains the weighting for both, eigenvector vectors spanning the matrix $\overline{\mathbf{R}}$.

For a unique solution, determining the weighting of the principal eigenvector, and thus the whitened RTF vectors in (4.17), the ambiguity evoked by the plus-minus-sign can be lifted by discussing one extreme case: Considering the case where $\overline{\mathrm{SIR}} \to 0$, i.e, no target is speaker present, the weighting for the target RTF vector α_x should go towards 0, while the weighting for the interferer RTF vector α_i should go towards 1 and therefore r should go towards 0. If the solution in (4.30) with the minus-sign is now considered - note that the phase information is neglected and that $\cos\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}} > 0$ - the solution for r would become negative, which does not match the outcome discussed above. If, however, the solution with the plus-sign is used, the terms in front of the plus sign tend to equal the negative of the square-root term so they cancel, and hence also r goes to 0, as expected. From this, it finally follows that the ratio r for the weighting of the two whitened RTF vectors is given by

$$\boxed{r = e^{j\angle \overline{\mathbf{h}}^H \overline{\mathbf{b}}} \frac{\overline{\mathrm{SIR}} - 1}{2\cos\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}}} + e^{j\angle \overline{\mathbf{h}}^H \overline{\mathbf{b}}} \sqrt{\left(\frac{\overline{\mathrm{SIR}} - 1}{2\cos\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}}}\right)^2 + \overline{\mathrm{SIR}}}\,.} \tag{4.31}$$

Together with a normalization of α_x and α_i, that is

$$\alpha_x + \alpha_i = 1\,, \tag{4.32}$$

the ratio r in (4.31) fully defines the weighting of the two whitened RTF vectors $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$ in (4.17). Note that in general also a different normalization of α_x and α_i than in (4.32) is possible, since the EVD problem in (4.18) is invariant to scaling of $\mathbf{v}_{\max}$. However, the normalization of α_x and α_i is chosen as such due to the direct

interpretability as a weighing of the vectors $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$.

4.4 Visualization of Bias Analysis

To visualize the weighting of the target RTF vector when using the CW method in the presence of an interferer, a simulation of one frequency bin is performed. Two generic RTF vectors (in the whitened domain) are created, i.e., $\overline{\mathbf{h}} = 1/\sqrt{3}[1,\, 1,\, 1]^T$ and $\overline{\mathbf{b}} = 1/\sqrt{3}[1,\, e^{-j\Delta_p},\, e^{-j2\Delta_p}]^T$, which can be interpreted as the RTF vectors resulting from two sources impinging on a uniform linear array (ULA) with three microphones. One source is impinging from broadside (target) and the other from an arbitrary direction (interferer), which creates a phase shift of Δ_p and respectively $2\Delta_p$. The phase shift Δ_p is varied, such that different Hermitian angles $\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}}$ between $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$ ranging from 5° to 90° are created, which allows for investigating its influence on the weighting. The Hermitian angle $\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}}$ can - to some extent - be interpreted as the "spatial" similarity of the two sources. For the investigation of the SIR in the whitened domain, $\overline{\mathrm{SIR}}$ is varied from -20 dB to 20 dB. For visualization purposes, in Fig. 4.2, the real-valued weighting γ_x is considered, where the normalization is performed as $|\alpha_x| + |\alpha_i| = 1$, and thus

$$\begin{aligned} \gamma_x &= |\alpha_x| \\ &= \frac{1}{1 + 1/|r|}\,. \end{aligned} \tag{4.33}$$

The real-valued weighting factor γ_x is plotted over $\overline{\mathrm{SIR}}$ in the upper panel. Note that normalizing α_x and α_i using the absolute value is not identical with the normalization in (4.32), since the complex phase information is neglected here. However, it allows for a simpler visualization and intuitive understanding of the underlying processes since a complex-valued weighting would be more difficult to interpret. The weighting of $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$ is obtained by orthogonally projecting $\mathbf{v}_{\max}$ (obtained from the EVD of the matrix $\overline{\mathbf{R}}$) onto the subspace spanned by $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$, that is

$$\mathbf{c} = \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{v}_{\max}\,, \tag{4.34}$$

where the vector $\mathbf{c}$ contains the complex-valued weighting of $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$, which are stacked in the matrix $\mathbf{A} = [\overline{\mathbf{h}}, \overline{\mathbf{b}}]$. The vector $\mathbf{c}$ is subsequently normalized (i.e., by taking the absolute value of its entries and normalizing them to add up to 1) to obtain γ_x. It should be noted that the results shown here are obtained from the simulation as described above, where the weighting is computed after performing an EVD on the whitened rank-2 covariance matrix spanned by the artificially created

vectors $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$ and not directly from the formula shown in (4.31). However, identical curves are obtained by simulating the theoretical results.

The upper panel of Fig. 4.2 shows the real-valued weighting γ_x of $\overline{\mathbf{h}}$ in dependence on $\overline{\text{SIR}}$, i.e., the (multi-channel) input SIR in the whitened domain. The lower panel in Fig. 4.2 shows the resulting Hermitian angle $\Theta_{\mathbf{v}_{\max},\overline{\mathbf{h}}}$ (i.e., the actual estimation accuracy) between the estimated and the target RTF vector.

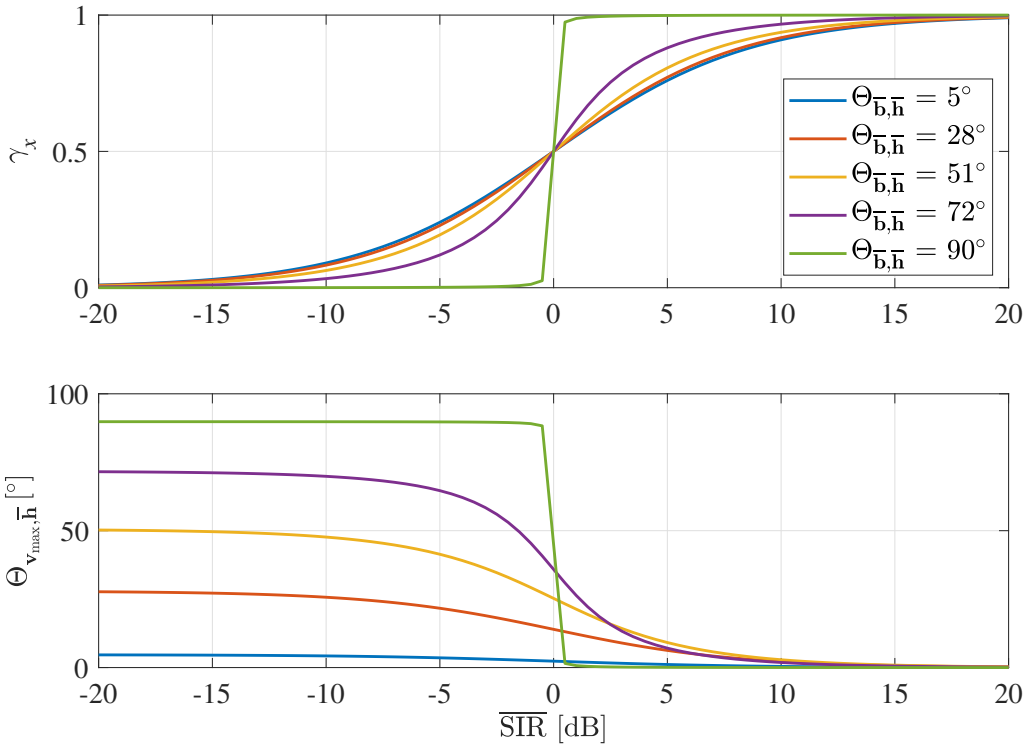

Fig. 4.2: Weighting γ_x of the target RTF vector and Hermitian angle between the estimated RTF vector and the target RTF vector for different input SIRs in the whitened domain and different Hermitian angles between target and interferer RTF vector.

The results show that, generally, a larger $\overline{\text{SIR}}$ leads to a stronger weighting of the target RTF vector, which can directly be deduced from the strictly positive influence of $\overline{\text{SIR}}$ on r in (4.31) (when neglecting the phase information). A larger Hermitian angle $\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}}$ between $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$ (see legend in Fig. 4.2) leads to a steeper curve. This can also be explained by taking an extreme cases into account, again neglecting the phase information: A Hermitian angle between $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$ going towards 90° will lead to a cosine term, for which $\cos\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}} \to 0$ and thus $1/\cos\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}} \to \infty$. Hence, for $\overline{\text{SIR}} > 1$ (or respectively $\overline{\text{SIR}} > 0$ dB), the total expression for r goes to ∞, which, when normalized as in (4.33), leads to a real-valued weighting of the target RTF vector $\gamma_x = 1$, i.e., the estimated RTF vector corresponds to the target RTF vector.

In the case of $\overline{\mathrm{SIR}} < 1$ (i.e., $\overline{\mathrm{SIR}} < 0$ dB), the term in front of the square-root in (4.31) goes to $-\infty$, while the first term inside the square-root goes to ∞ (due to the square). The second term inside the square-root becomes negligible. Then, both terms, in front of and inside the square-root, become identical but with opposite signs and thus $r \to 0$. In this extreme case, a toggling behavior between positive and negative SIRs can be seen in Fig. 4.2. This can also be explained by the fact that the matrix $\overline{\mathbf{R}}$ is a Hermitian matrix, which means that its eigenvectors are orthogonal to each other. In case of $\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}} = 90°$, the vectors $\overline{\mathbf{h}}$ and $\overline{\mathbf{b}}$ are already orthogonal and are therefore the eigenvectors of $\overline{\mathbf{R}}$. Thus, the EVD of $\overline{\mathbf{R}}$ "picks" the vector of the two which corresponds to the largest eigenvalue. For Hermitian angles $\Theta_{\overline{\mathbf{b}},\overline{\mathbf{h}}} < 90°$, this toggling behavior becomes more gradual and therefore yields a smoother weighting between low and high SIRs.
For very low SIRs, where the weighting of the target goes to zero, and as such the estimated RTF vector almost perfectly corresponds to the interferer RTF vector, the Hermitian angle $\Theta_{\mathbf{v}_{\max},\overline{\mathbf{h}}}$ between estimated and target RTF vector converges towards the Hermitian angle between interferer and target RTF vector. Therefore, the Hermitian angle between interferer and target RTF vector limits the "missteering" of a beamformer which uses this RTF vector estimate. At high SIRs, where $\mathbf{v}_{\max}$ mostly corresponds to the target RTF vector, the Hermitian angle $\Theta_{\mathbf{v}_{\max},\overline{\mathbf{h}}}$ goes towards 0.

4.5 Validation of Bias Analysis

In this section, it is investigated what influences the biased RTF vector estimation discussed in Section 4.3 has on the output of a beamformer. To this end, simulations with artificial data are conducted.

4.5.1 Configuration and Conditions

The configuration of microphones, target speaker, and interferer is depicted in Fig. 4.3. The room and the respective room impulse responses (RIRs) are simulated using the RIR generator by Habets et al. [29]. The parameters for this simulation are set accordingly: The room has dimensions of $(5 \times 5 \times 3)$ m, the speed of sound is set to 340 m/s and the sampling frequency is set to 16 kHz. To achieve a scenario that is as controlled as possible, the reverberation time is set to 0 ms, i.e., the walls, floor, and ceiling of the simulated room do not produce any reflections. As a target and interferer signal, two uncorrelated white noise signals are generated and convolved with their respective RIR. The noise signals are generated using the method in [30].

Spatially, the noise can be regarded as diffuse, while spectro-temporally it is a white noise signal. The LMA is a ULA with $M_\mathrm{a} = 4$ microphones, with a spacing of 2 cm. An eMic is placed 1.8 m away from the LMA in endfire direction. The target speaker is placed in endfire direction, 2 m away from the LMA. The interferer is placed in broadside direction, also at a distance of 2 m from the LMA.

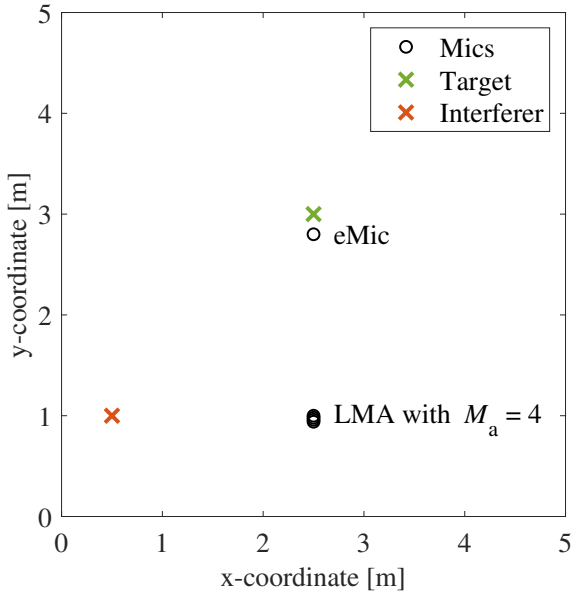

Fig. 4.3: Configuration of artificial scene for validation of CW bias analysis using a ULA as the LMA with a target placed in endfire direction and an interferer in broadside direction. An additional eMic is placed close to the target.

In this validation, the following input parameters are chosen: The input SNR in the reference microphone (here the first of the ULA) is set to a constant value of 0 dB. The input SIR in the reference microphone is varied from -20 dB to 20 dB. This also means that the input interferer-to-noise ratio (INR) is not constant over all conditions and it takes values between 20 dB and -20 dB.
Furthermore, three conditions with different microphone configurations are investigated:

Condition 1: Only the LMA is used for both, RTF vector estimation and filtering, i.e., the eMic is ignored in this condition (Fig. 4.4 - 4.6).

Condition 2: The eMic, in addition to the LMA, is used for both, the RTF vector estimation and the filtering (Fig. 4.7 - 4.9).

Condition 3: The eMic is only used for the RTF vector estimation, but its signal is not filtered. Which can be understood as the eMic "guiding" the RTF vector

estimation of the LMA (Fig. 4.7 - 4.9).

The three conditions allow to investigate the influence of the multi-channel SIR on the RTF vector estimation using CW.

4.5.2 Implementation and Parameters

For the filter, steered by the estimated RTF vector, an MPDR beamformer, defined in (3.4), is used. The parameters of the implementation and the STFT framework are set as follows: A frame length of 1024 samples is used with an overlap of 50%, corresponding to a frame shift of 512 samples. The frames are windowed with a square-root-Hann window for analysis and synthesis. The covariance matrices $\mathbf{R}_y$ and $\mathbf{R}_n$ are estimated batch (i.e., averaged over the complete signal with a length of 10 s) with oracle knowledge about the noise component.
One performance measure is the weighting of the two oracle RTF vectors $\mathbf{h}$ and $\mathbf{b}$, where a weighting going towards 1 indicates that the respective source dominates the RTF vector estimation, while a weighting going towards 0 indicates that the respective source barely influences the RTF vector estimation. In contrast to the weighting discussed in Section 4.3, the weighting here is computed in the de-whitened domain. To this end, the least-squares solution, similar to (4.34), is computed on the de-whitened RTF vectors, i.e., using the estimated RTF vector $\hat{\mathbf{h}}^{\mathrm{CW}}$ as in (4.8) (and not the principal eigenvector $\mathbf{v}_{\max}$ in the whitened domain) and the RTF vectors $\mathbf{h}$ and $\mathbf{b}$. The weighting factors obtained per frequency-bin are averaged subsequently. Another evaluated measure is the Hermitian angle between the target RTF vector (and respectively the interferer RTF vector) and the estimated RTF vector, which is computed per frequency bin as defined in (2.34) and subsequently averaged over all frequency bins. Furthermore, the SD as in (2.32) for both speech sources , i.e., the target and the interferer, is investigated as well as the resulting SIR improvement ΔSIR, as in (2.30). Similarly, the NR as in (2.33) and the resulting SNR improvement ΔSNR, as in (2.29), are considered. All metrics are computed broad-band via the root-mean-square value of the time-domain signals as defined in Section 2.1.

4.5.3 Results

In Fig. 4.4, the upper panel shows the weighting γ of the target and the interferer RTF vector, and the lower panel shows the resulting Hermitian angle between the estimated RTF vector and the target and interferer RTF vector, respectively, fir the first considered condition. Note that on the x-axis, the input SIR in the reference microphone (i.e., the first of the LMA) is depicted and not the multi-channel SIR which is the determining factor of the weighting ratio in (4.31).

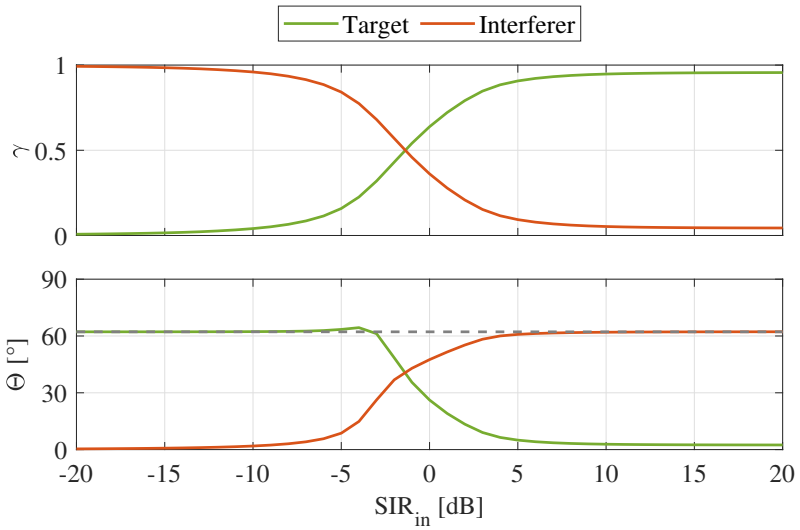

Fig. 4.4: Weighting and Hermitian angle for condition 1 (see Section 4.5.1) of CW bias analysis. Upper panel: Weighting of target and interferer RTF vector for the blindly estimated RTF vector using CW. Lower panel: Hermitian angle between target/interferer RTF vector and estimated RTF vector.

The two curves for γ in the upper panel of Fig. 4.4 clearly show a strong strong weighting of the target RTF vector (green curve) at positive input SIRs, going towards 1, and a low weighting at negative SIRs, going towards 0, as expected from the results shown in Section 4.4 and vice versa for the weighting of the interferer RTF vector (red curve). It should be noted that the weighting of 0.5 is not obtained at exactly 0 dB input SIR, as it would be expected from theory. This might be caused by the a-symmetry of the acoustic scenario and possible effects caused by the whitening operation. The resulting Hermitian angle is depicted in the lower panel, where an accordance with the findings in Section 4.4 can be seen: At high input SIRs, where the target RTF vector $\mathbf{h}$ is weighted more strongly, the Hermitian angle between $\mathbf{h}$ and the estimate obtained using CW $\hat{\mathbf{h}}^{\mathrm{CW}}$ goes towards 0, indicating a very precise estimate, while at low input SIRs, this angle increases up to 62° (gray dashed line), which is the Hermitian angle between the target and the interferer RTF vector. The Hermitian angle between the interferer RTF vector $\mathbf{b}$ and $\hat{\mathbf{h}}^{\mathrm{CW}}$ behaves

similarly, i.e., a small Hermitian angle for the interferer is obtained at low input SIRs, meaning that $\hat{\mathbf{h}}^{\mathrm{CW}}$ almost perfectly corresponds to $\mathbf{b}$, while a large Hermitian angle is found at high input SIRs.
So far, the results from Section 4.4 were confirmed. Now, the consequences of this RTF vector estimation on a beamformer, namely the MPDR beamformer, are investigated, i.e., the influence on the separate processing of the target speaker, the interferer and the noise is analyzed experimentally. Fig. 4.5 shows the SD scores for both speech sources, i.e., green for the target and red for the interferer, and the resulting SIR improvement ΔSIR (blue).

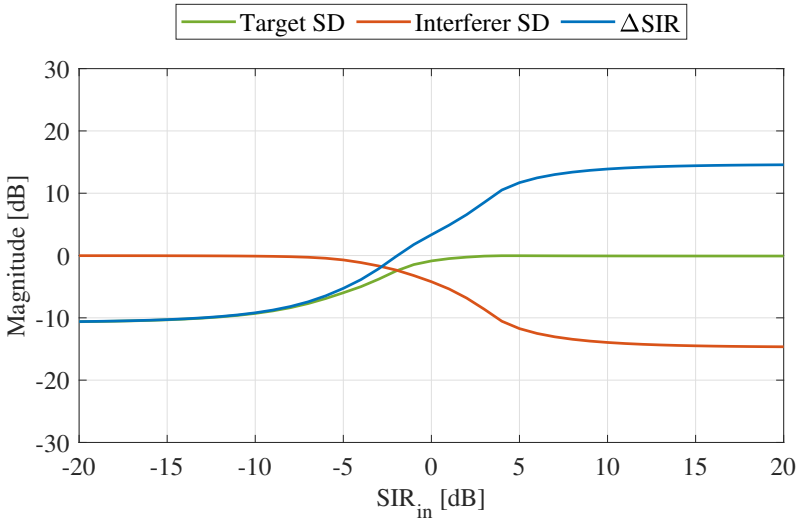

Fig. 4.5: Target and interferer SD and ΔSIR for condition 1 (see Section 4.5.1) of CW bias analysis.

It can be seen that for high input SIRs, the SD score for the target goes towards 0, which indicates a perfect preservation of the target speech, while low input SIRs lead to target speech distortions of about -10 dB, indicating a cancellation of target speech. Vice versa, the SD scores for the interferer go towards 0 at low input SIRs and decrease towards higher input SIRs down to about -15 dB. The blue curve in Fig. 4.5 depicts the SIR improvement, which is the difference between the target SD and the interferer SD scores, which is used more often than separate SD scores to characterize the interferer reduction performance, as it quantifies the interferer reduction relative to the distortions of the target speech. It can be seen that the SIR improvement shows good (i.e., positive) results up to about 15 dB at high input SIRs, which indicates that a blind RTF estimation using CW works in the case of a low interferer level, where the interferer is even cancelled leading to an even higher SIR at the output. However, at low input SIRs, i.e., in scenarios where interferer reduction is essential to gain intelligibility of the target, blind RTF vector estimation

leads to negative ΔSIR scores, down to about -10 dB, and thus an enhancement of the interferer (compared to the target).

Since the objective of the algorithms in this thesis is both, interferer and noise reduction, the influence of the beamformer on the present noise field is now investigated. Fig. 4.6 shows the NR scores (purple), the SD scores for the target (green), and the SNR improvement (blue).

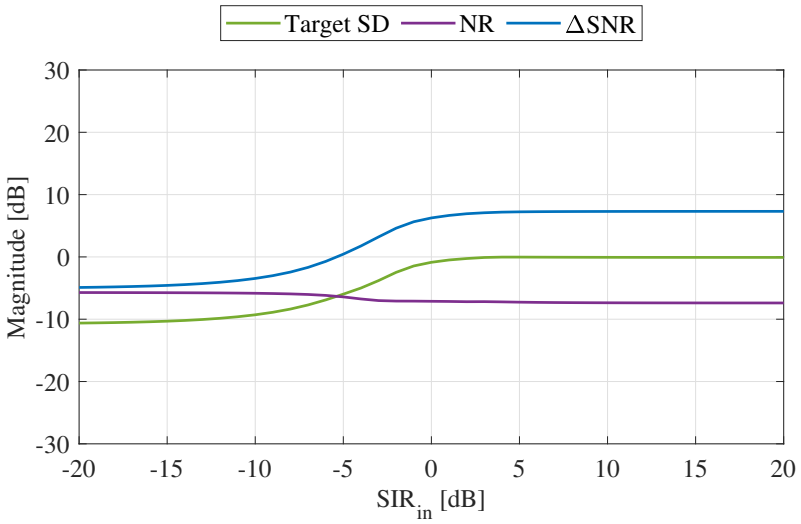

Fig. 4.6: Target SD, NR and ΔSNR for condition 1 (see Section 4.5.1) of CW bias analysis.

As the NR scores are rather constant (between -5.7 dB at low input SIRs and -7.4 dB at high input SIRs), the SIR improvement yields a similar curve as the SD scores, but shifted up by the amount of NR performed by the beamformer. Similarly to the ΔSIR curve in Fig. 4.5, it can be seen that for low input SIRs, the ΔSNR score becomes negative, meaning that the target is cancelled more than the noise and at high input SIRs, the ΔSNR scores take positive values, indicating that the target speech is enhanced compared to the noise.
The conclusion that can be drawn from this first condition is that blind RTF vector estimation using CW and the subsequent use of this RTF vector estimate in a state-of-the-art beamforming algorithm is not suitable for joint noise and interferer reduction in scenarios with low input SIRs, where is it a fundamental task to reduce the interferer to achieve intelligibility of the target speaker. It is shown here that always the dominant speaker is preserved (as also expected from the theoretical derivation in Section 4.3), which is not the stated goal of this thesis.
Though this naïve approach can be discarded to achieve the stated goal of joint noise and interferer reduction, two further conditions are evaluated to investigate

the influence of the multi-channel SIR on the RTF vector estimated using CW by incorporating an eMic placed close to the target speaker. The second condition investigated here is where the eMic is included in the estimation of the RTF vector and also in the MPDR beamformer. The results are shown in Fig. 4.7 - 4.9. Similarly to Fig. 4.4, Fig. 4.7 shows the weighting of the target and the interferer RTF vector in the upper panel and the Hermitian angle between the estimated RTF vector and the target and interferer RTF vector, respectively, in the lower panel.

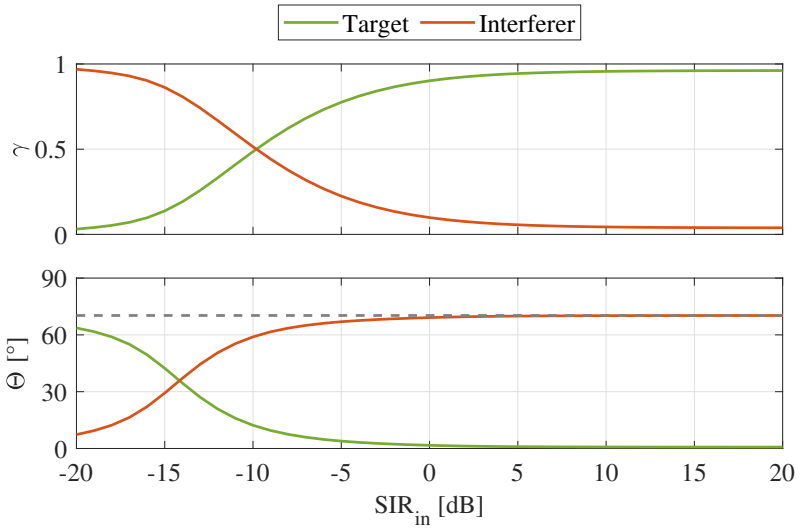

Fig. 4.7: Weighting and Hermitian angle for condition 2 (see Section 4.5.1) of CW bias analysis. Upper panel: Weighting of target and interferer RTF vector for the blindly estimated RTF vector using CW. Lower panel: Hermitian angle between target/interferer RTF vector and estimated RTF vector.

Generally, similar trends as before are observed: For high input SIR the target RTF vector is weighted more strongly and for low input SIRs the interferer RTF vector is weighted more strongly. However, a clear shift of the two curves towards lower input SIRs can be observed, as a weighting of 0.5 for the two RTF vectors is found at an input SIR of about -10 dB. Accordingly, also the curves for the Hermitian angle between the target RTF vector $\mathbf{h}$ or respectively the interferer RTF vector $\mathbf{b}$ and the estimated RTF vector $\hat{\mathbf{h}}^{\mathrm{CW}}$ shift towards lower input SIRs. Note that in contrast to the first condition, the RTF vector now includes the eMic and therefore has an additional entry, which changes the Hermitian angle between $\mathbf{h}$ and $\mathbf{b}$ from 62° to 70°, which now limits the "missteering" of the beamformer. However, in the shown range from -20 dB to 20 dB input SIR, the Hermitian angle between $\hat{\mathbf{h}}^{\mathrm{CW}}$ and $\mathbf{h}$ does not reach a value of 70°, while the Hermitian angle between $\hat{\mathbf{h}}^{\mathrm{CW}}$ and $\mathbf{b}$ reaches the 70° at an input SIR of about 0 dB.

The consequences of this shift of the weighting curves and the resulting RTF vector

on the beamformer's output are shown in Fig. 4.8 and Fig. 4.9. Similarly to Fig. 4.5, Fig. 4.8 shows the effect of the beamformer on the target and the interferer.

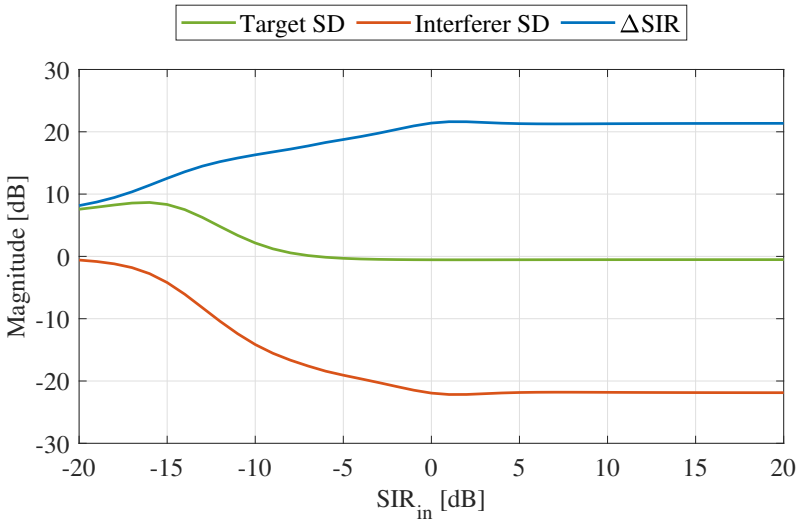

Fig. 4.8: Target and interferer SD and ΔSIR for condition 2 (see Section 4.5.1) of CW bias analysis.

For the interferer SD, it can be observed that much lower values of about -22 dB are obtained (at mid and high input SIRs), and that even for low input SIRs, negative SD scores are obtained, which converge towards 0 dB for very low input SIRs. The SD scores for the target are constantly at 0 dB for input SIRs between -5 dB to 20 dB, only for lower input SIRs, target SD scores unequal to zero are obtained, which, however, are positive (up to almost 10 dB), indicating a gain applied to the target speech by the beamformer. Following from the SD scores for target and interferer, the resulting SIR improvement ΔSIR is constantly positive, taking values of about 10 dB to up to 22 dB. These results can partially be explained when considering the fact that the eMic is placed so close to the target and therefore has an input SIR that is about 22 dB larger than the reference microphone (for which the input SIR is shown on the x-axis of the graphs). For a more detailed explanation, further investigations of the beamformer would be required which exceed the scope of this validation.

The results for the beamformer's influence on the noise are shown in Fig. 4.9, where the overall trend of an improvement compared the the configuration without the eMic, shown in Fig. 4.6, can be observed.

The NR scores in Fig. 4.9 go down to about -20 dB at high input SIRs and gradually increase, i.e., get less negative up to -8 dB, towards low input SIRs. Considering the high input SNR in the eMic of about 20 dB more compared to the reference microphone can explain these results. Following from the NR scores and the SD scores for

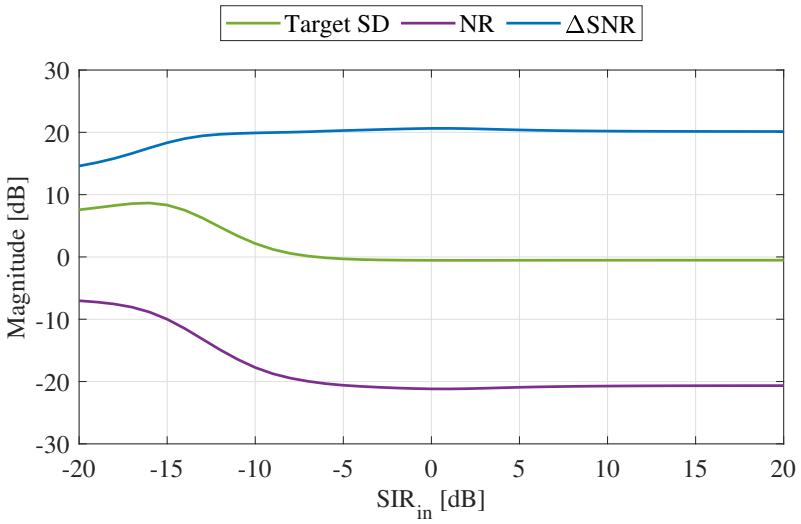

Fig. 4.9: Target SD, NR and ΔSNR for condition 2 (see Section 4.5.1) of CW bias analysis.

target, the resulting SNR improvement ΔSNR is constantly positive, taking values of about 15 dB to 20 dB.

From this second conditions, two major conclusions can be drawn: 1) The RTF vector estimation indeed depends on the multi-channel input SIR, meaning that one microphone with a particularly high SIR leads to an RTF vector estimate that favors the target over the interferer. And 2) a very well placed eMic in conjunction with an LMA (e.g., hearing aids) actually allows for joint noise and interferer reduction, even without any sophisticated processing schemes. This would for example be the case for a lapel microphone attached to the target speaker. However, if the eMic is not placed close to the target speaker, e.g., a table microphone in a restaurant that captures all speech signals or even being close to the interferer, the performance of a blind beamformer would strongly suffer and joint noise and interferer reduction would not be possible, perhaps even leading to an enhanced interferer signal. Though the assumption that the eMic is placed well, i.e., close to the target speaker, is often made in practice, it might not always hold and the systems based on this assumption strongly suffer. Thus, this assumption is not made in this thesis and the general case of arbitrarily placed eMics is considered.

In the last condition investigated here, the eMic is only used for the RTF vector estimation, but not in the beamformer, meaning that the entry of the RTF vector corresponding to the eMic is discarded after the estimation. This allows for the investigation of the multi-channel SIR on the entire RTF vector, to ensure that the

results found in the condition before were not purely caused by the entry of the estimated RTF vector corresponding to the eMic, but by the entire estimated RTF vector favoring the target over the interferer. Furthermore, this allows to investigate if (or how much of) the benefit coming with including the eMic is caused by a better RTF vector estimation or respectively by the incorporation of a very good microphone into the beamformer. The results for this third condition are shown in Fig. 4.10 - 4.12. The weighting in the upper panel of Fig. 4.10 and the Hermitian angle in the lower panel are now again only computed on the entries of the RTF vector used in the beamformer, i.e., the ones corresponding to the LMA, giving a maximal Hermitian angle of 62° as in the first condition (see Fig. 4.4).

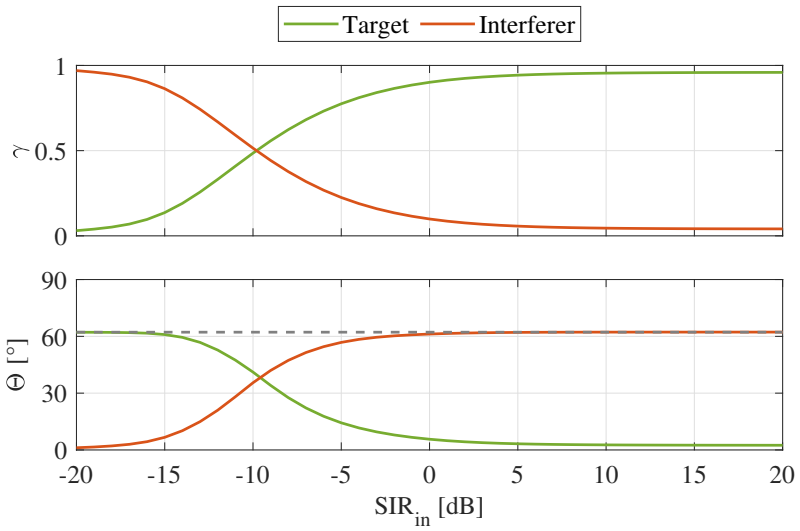

Fig. 4.10: Weighting and Hermitian angle for condition 3 (see Section 4.5.1) of CW bias analysis. Upper panel: Weighting of target and interferer RTF vector for the blindly estimated RTF vector using CW. Lower panel: Hermitian angle between target/interferer RTF vector and estimated RTF vector.

The weighting curves of the target RTF vector $\mathbf{h}$ and the interferer RTF vector $\mathbf{b}$, identical curves as in Fig. 4.7 are obtained, even when neglecting the entry corresponding to the eMic. This indicates that the entire RTF vector estimate is affected when including one microphone with a high input SIR and not only that particular entry. For the Hermitian angles between the estimated RTF vector $\hat{\mathbf{h}}^{\mathrm{CW}}$ and $\mathbf{h}$ and $\mathbf{b}$, respectively, slightly different curves than in Fig. 4.4 are obtained, since they, as already mentioned, converge to a different upper bound (namely 62°) and are also shifted slightly towards higher input SIR. However, the overall trend of an accurate RTF vector estimate which has a small Hermitian angle to the target RTF vector even at negative SIR is still present.
The effect of the estimated (and cropped) RTF vector on the target and the interferer

can be seen in Fig. 4.11.

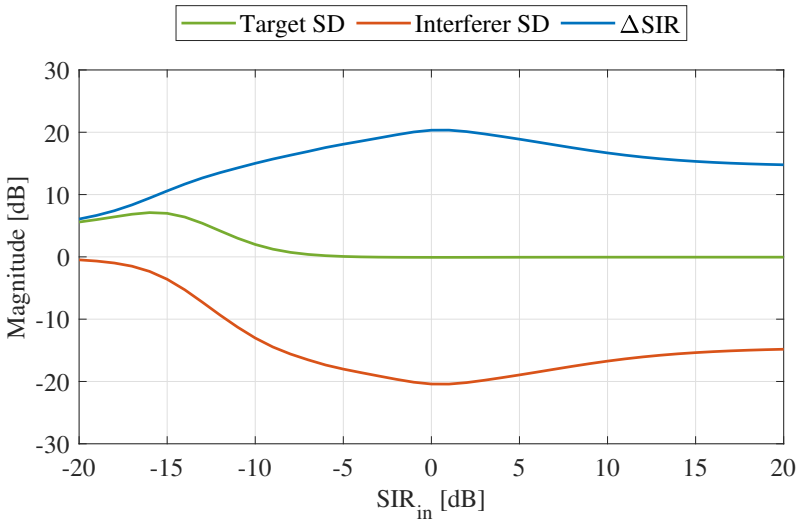

Fig. 4.11: Target and interferer SD and ΔSIR for condition 3 (see Section 4.5.1) of CW bias analysis.

Similarly to the results in Fig. 4.8, the SD score for the target is larger or equal to 0 dB for all input SIRs, although the maximal value is only slightly lower (a gain of about 7 dB). The SD scores for the interferer also resemble the results from Fig. 4.8, even though the eMic is not included in the beamformer. The SD scores for the interferer go down to about -20 dB at an input SIR of 0 dB, but go towards less negative values (to about -15 dB) for higher input SIRs. The resulting SIR improvement is therefore always positive and takes values of 6 dB to 20 dB. These results indicate that using a well placed eMic, or generally a signal with a high SIR, to "guide" the RTF vector estimation yields good results in terms of interferer reduction.

Similar to the interferer reduction, the results for the noise reduction, depicted in Fig. 4.12, indicate a clear benefit from using the eMic in the RTF vector estimation. The ΔSNR scores are positive for all input SIRs and have an almost constant value of 6 dB to 7 dB. The tendencies are the same as in the second condition where the eMic was also filtered (see Fig. 4.9), but the absolute values are quite different: When the eMic is included in the filtering, NR scores of down to -20 dB are obtained, while when only included in the RTF vector estimation, the NR scores only go down to about -7 dB. Nevertheless, since the resulting ΔSNR scores are positive for all input SIRs, good results are achieved when the RTF vector estimation is "guided" by a microphone with a very high SIR.

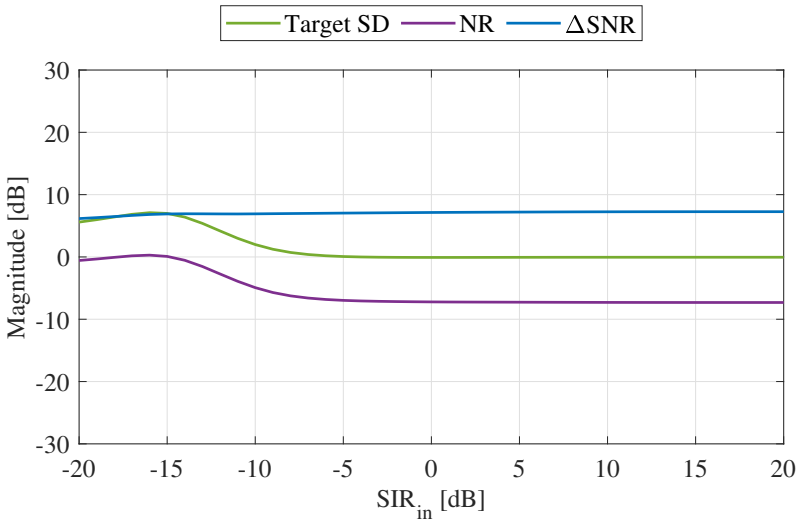

Fig. 4.12: Target SD, NR and ΔSNR for condition 3 (see Section 4.5.1) of CW bias analysis.

The main conclusion that can be drawn from the third considered condition, where the eMic is only used to estimate the RTF vector but is not included in the beamformer, is that the entire RTF vector is affected by the one particularly good channel and not just the entry corresponding to it. Similarly, the noise and interferer reduction scores obtained when using this estimated RTF vector leads to clear improvements compared to the first condition where the RTF vector was estimated only using the microphones of the LMA.

To summarize the performed validation of the bias analysis for the CW method, the main conclusions are recapitulated here:
In the first condition, i.e., where only the LMA was considered for both, the RTF vector estimation and the beamformer, it could clearly be seen that the dominant speaker was preserved while the other was suppressed. In this case, joint noise and interferer reduction is therefore not possible, since in case of a low input SIR, where interferer reduction is most important, the interferer is not reduced, but rather the target.
In the second condition, i.e., where additionally to the LMA an eMic placed close to the target was used for both, the RTF vector estimation and the beamformer, it could be seen that even for low input SIRs at the LMA, the estimated RTF vector corresponded rather to the target RTF vector than to the interferer RTF vector. This finding supports the theoretical results in Section 4.3 that the factor determining the weighing of the two RTF vectors in the CW method is the multi-channel SIR. The beamformer using this RTF vector is suitable for joint noise and interferer

reduction, since for a large range of input SIRs, no target SDs were introduced and the ΔSIR and ΔSNR scores were constantly positive. However, this finding also implies that an eMic that is placed close to the interferer leads to a beamformer which enhances the interferer and hence suppresses the target. For the general case of an eMic for which it is not explicitly assumed to be close to the target, including an eMic using a blind RTF estimation and beamforming approach is therefore also not suitable for joint noise and interferer reduction.
In the third condition, i.e., where the eMic was used only for the RTF vector estimation, but not in the beamformer, it could be seen that despite discarding the entry corresponding to the eMic, the accuracy of the estimated RTF vector was very high, indicated by a low Hermitian angle between the estimated RTF vector and the target RTF vector. This finding implies that using a particularly good channel for "guiding" the RTF vector estimation leads to an overall more precise RTF vector estimate. The resulting beamformer therefore also favors the target over the interferer, similarly to the second condition.
Generally, from this validation the conclusions can be drawn that 1) the theoretical findings in Section 4.3 appear to hold and 2) that in the general case of an unknown (possibly unfortunate) position of one or multiple eMics, a blind RTF vector estimation and beamforming approach is not suitable for joint noise and interferer reduction.

5 The Generalized Sidelobe Canceller Structure

In this section, the generalized sidelobe canceller (GSC) structure is introduced. The GSC is often regarded as an adaptive implementation of the MVDR or respectively MPDR beamformer [6–8]. In its closed form implementation, the GSC is - in fact - identical to the MVDR/MPDR beamformer. In this thesis, the GSC is therefore not considered for being an adaptive implementation of the MPDR beamformer, but rather for its larger flexibility and intermediate signal stages. Furthermore, introducing the GSC allows for a distinction of informed and un-informed beamforming structures. As discussed in Section 3, the MPDR is considered to be a un-informed beamformer, where the entire RTF vector blindly. As shown in Sections 4.3 and 4.5, blind beamforming approaches are not suitable in acoustic scenarios with interfering speakers. Therefore, the a-priori RTF vector $\tilde{\mathbf{h}}_{\mathrm{a}}$, which is assumed to be given for the hearing aid microphones (i.e., the LMA) can exploited. The resulting informed beamformer, namely the local GSC (L-GSC), is only applied to the LMA signals $\mathbf{y}_{\mathrm{a}}$ and exploits the a-priori RTF vector $\tilde{\mathbf{h}}_{\mathrm{a}}$. Using the GSC structure instead of a closed form MPDR for the local beamformer allows to later exploit the intermediate signal stages for the incorporation of the eMics (see Section 6).

The block diagram of the L-GSC is shown in Fig. 5.1. The filter blocks that are not data driven, but which exploit the a-priori RTF vector are the blocking matrix $\mathbf{C}_{\mathrm{a}}$ and the fixed beamformer $\mathbf{f}_{\mathrm{a}}$.

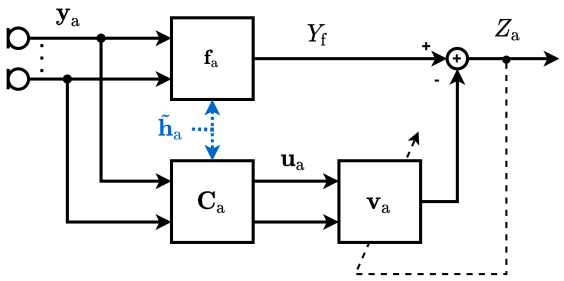

Fig. 5.1: Block diagram of the L-GSC, applied to the local microphones, where the fixed beamformer $\mathbf{f}_{\mathrm{a}}$ and the blocking matrix $\mathbf{C}_{\mathrm{a}}$ exploit the a-priori RTF vector $\tilde{\mathbf{h}}_{\mathrm{a}}$.

The upper branch of the L-GSC is meant to fulfill the distortionless constraint of the MPDR and is given by a fixed beamformer $\mathbf{f}_{\mathrm{a}}$ (here a matched filter), relying on the a-priori RTF vector $\tilde{\mathbf{h}}_{\mathrm{a}}$, that is

$$\mathbf{f}_{\mathrm{a}} = \frac{\tilde{\mathbf{h}}_{\mathrm{a}}}{||\tilde{\mathbf{h}}_{\mathrm{a}}||_2^2}\,, \tag{5.1}$$

such that $\mathbf{f}_{\mathrm{a}}^{H}\tilde{\mathbf{h}}_{\mathrm{a}} = 1$. The speech reference Y_{f} is obtained by applying $\mathbf{f}_{\mathrm{a}}$ to the local noisy input signals $\mathbf{y}_{\mathrm{a}}$, i.e.,

$$Y_{\mathrm{f}} = \mathbf{f}_{\mathrm{a}}^{H}\mathbf{y}_{\mathrm{a}}\,. \tag{5.2}$$

The lower branch of the structure in Fig. 5.1 aims at minimizing the total power at the output (similarly to the minimization problem in (3.2)). This is done by means of the $M_{\mathrm{a}} \times (M_{\mathrm{a}}\text{-}1)$-dimensional blocking matrix $\mathbf{C}_{\mathrm{a}}$, which is orthogonal to the a-priori RTF vector, that is $\mathbf{C}_{\mathrm{a}}^{H}\tilde{\mathbf{h}}_{\mathrm{a}} = \mathbf{0}_{(M_{\mathrm{a}}-1)\times 1}$. In the case of the first microphone being the reference microphone, $\mathbf{C}_{\mathrm{a}}$ can be constructed as in [8], i.e.,

$$\mathbf{C}_{\mathrm{a}} = \begin{bmatrix} -\tilde{H}_{\mathrm{a},2}^{*}\,, -\tilde{H}_{\mathrm{a},3}^{*}\,, \ldots\,, -\tilde{H}_{\mathrm{a},M_{\mathrm{a}}}^{*} \\ \mathbf{I}_{M_{\mathrm{a}}-1} \end{bmatrix}\,, \tag{5.3}$$

where $\tilde{H}_{\mathrm{a},m_{\mathrm{a}}}$ is the m_{a}-th entry of the a-priori RTF vector $\tilde{\mathbf{h}}_{\mathrm{a}}$. By applying $\mathbf{C}_{\mathrm{a}}$ to the local input signals $\mathbf{y}_{\mathrm{a}}$, M_{a}-1 so-called noise-and-interferer references, stacked in the vector $\mathbf{u}_{\mathrm{a}}$, are obtained, i.e.

$$\mathbf{u}_{\mathrm{a}} = \mathbf{C}_{\mathrm{a}}^{H}\mathbf{y}_{\mathrm{a}}\,. \tag{5.4}$$

In case of an ideal a-priori RTF vector, $\mathbf{u}_{\mathrm{a}}$ only contains noise and interferer components, but no target speech component. However, if $\tilde{\mathbf{h}}_{\mathrm{a}}$ is not identical to the true target RTF vector, speech leakage into $\mathbf{u}_{\mathrm{a}}$ will occur. The noise-and-interferer references are subsequently filtered by the filter vector $\mathbf{v}_{\mathrm{a}}$, giving the output signal

$$Z_{\mathrm{a}} = Y_{\mathrm{f}} - \mathbf{v}_{\mathrm{a}}^{H}\mathbf{u}_{\mathrm{a}}\,. \tag{5.5}$$

To minimize the output power by means of the filter vector $\mathbf{v}_{\mathrm{a}}$, the following unconstrained optimization problem is considered

$$\min_{\mathbf{v}_{\mathrm{a}}} \mathcal{E}\{|\mathbf{f}_{\mathrm{a}}^{H}\mathbf{y}_{\mathrm{a}} - \mathbf{v}_{\mathrm{a}}^{H}\mathbf{C}_{\mathrm{a}}^{H}\mathbf{y}_{\mathrm{a}}|^{2}\}\,. \tag{5.6}$$

In contrast to [17], where an MVDR implementation was chosen, i.e., aiming at only minimizing the noise component, in this thesis, an MPDR implementation of the GSC is chosen, since it can also suppress the interferer and not only the noise. The equivalence of MVDR and MPDR discussed in Section 3 (shown in Appendix A)) does not hold here, since the respective RTF vector is not blindly estimated using CW. The closed form of $\mathbf{v}_{\mathrm{a}}$, i.e.. Wiener filter solution to the optimization problem in (5.6), is given by

$$\mathbf{v}_{\mathrm{a}} = (\mathbf{C}_{\mathrm{a}}^{H}\hat{\mathbf{R}}_{y,\mathrm{a}}\mathbf{C}_{\mathrm{a}})^{-1}\mathbf{C}_{\mathrm{a}}^{H}\hat{\mathbf{R}}_{y,\mathrm{a}}\mathbf{f}_{\mathrm{a}}\,. \tag{5.7}$$

Using an MPDR implementation for $\mathbf{v}_{\mathrm{a}}$,on the one hand, allows to actively suppress the interferer, which would not be possible with an MVDR implementation, where only the noise is reduced actively. On the other hand, it comes with the risk of target speech cancellation if there is speech leakage into the noise-and-interferer references $\mathbf{u}_{\mathrm{a}}$. Note that often adaptive implementation are used for $\mathbf{v}_{\mathrm{a}}$, e.g. by using an adaptive algorithm such as normalized least-mean squares (NLMS) [31]. However, throughout this thesis, the closed form of $\mathbf{v}_{\mathrm{a}}$ in (5.7) is used due to its equivalence to the closed form MPDR beamformer in (3.4).

As already mentioned, the advantage of the GSC structure over the closed form MVDR beamformer in (3.3) is its larger flexibility: The structure allows to design the separate blocks, i.e., $\mathbf{f}_{\mathrm{a}}$, $\mathbf{C}_{\mathrm{a}}$ and $\mathbf{v}_{\mathrm{a}}$, independently. And even more importantly here, it has intermediate signal stages, that is, the spatially pre-filtered speech reference Y_{f} and the noise-and-interferer references $\mathbf{u}_{\mathrm{a}}$. These signals can be exploited for the RTF vector estimation of the eMics (see Section 6).

6 Extended GSC Structures

In this section, four different extended GSC structures are considered. All four structures exploit the pre-filtered local signals (i.e., obtained from the L-GSC) for an improved RTF vector estimation of the eMics. The obtained RTF vector estimate is then used in one of two ways:

1. An external blocking matrix is completed with the estimated external RTF vector, yielding additional external noise-and-interferer references. The external noise-and-interferer references are then used to minimize the noise and interferer components at the structures' output by either joint filter optimization with the local filters or by cascaded minimization of the residual noise and interferer components at the output of the L-GSC. The two structures using external noise-and-interferer references are called GSC with external noise references type 1 (GSC-ENR-1) and GSC with external noise references type 2 (GSC-ENR-2), respectively, and are both presented in Section 6.1.

2. The external RTF vector is used to steer a joint MPDR beamformer of the L-GSC output and the (pre-processed) eMic signals. The two structures which do so differ in one major point: In the GSC with external references (GSC-ER) (Section 6.2), no pre-processing of the eMic signals is done before the joint MPDR beamformer, while in the GSC with external speech references (GSC-ESR) (Section 6.3), the eMic signals are pre-processed by means of the filtered local noise-and-interferer references, which aims at interferer reduction in the eMic signals.

6.1 GSC with External Noise References

The two extended GSC structures with external noise-and-interferer references are introduced in this section. First, the GSC-ENR-1 is presented and subsequently the GSC-ENR-2 is presented. Both structures exploit the pre-filtered, local signals for a better RTF vector estimate compared to a completely blind estimation, which is used to complete a blocking matrix yielding additional external noise-and-interferer references.

6.1.1 GSC with External Noise References Type 1

The GSC-ENR-1 was proposed in [17], where it was used to incorporate eMics to achieve noise reduction in a scenario with several stationary, coherent noise sources. The formulation of this structure is changed here from an MVDR implementation

to an MPDR implementation, such that not only noise but also the interferer can be cancelled. The structure in itself, however, is the same as in [17] and is shown in Fig. 6.1.

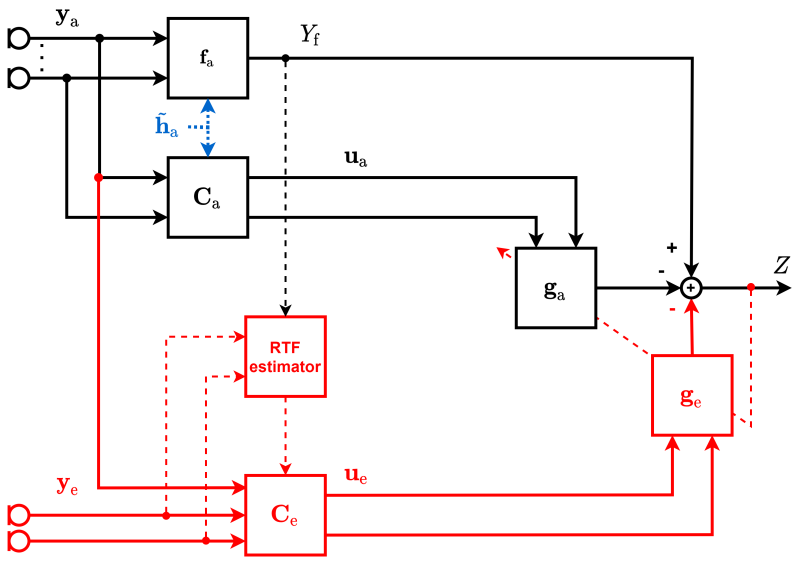

Fig. 6.1: Block diagram of the GSC-ENR-1, incorporating the eMics by creating external noise-and-interferer references. The RTF vector estimation is performed on Y_f and $\mathbf{y}_\mathrm{e}$ using CW.

In the GSC-ENR-1, the objective of incorporating the eMics is to create additional noise-and-interferer references from the eMic signals. This can be achieved by means of the blocking matrix $\mathbf{C}_\mathrm{e}$, which is defined as

$$\mathbf{C}_\mathrm{e} = \begin{bmatrix} -\hat{\mathbf{h}}_\mathrm{e}^H \\ \mathbf{I}_{(M_\mathrm{a}-1)\times(M_\mathrm{a}-1)} \end{bmatrix}, \tag{6.1}$$

which requires an estimate of the vector $\hat{\mathbf{h}}_\mathrm{e}$, containing the external RTFs. Applying $\mathbf{C}_\mathrm{e}$ to the signal of the first microphone of the LMA Y_1 and the eMic signals gives the M_e external noise-and-interferer references stacked in the vector $\mathbf{u}_\mathrm{e}$, i.e.,

$$\mathbf{u}_\mathrm{e} = \mathbf{C}_\mathrm{e}^H \begin{bmatrix} Y_1 \\ \mathbf{y}_\mathrm{e} \end{bmatrix}. \tag{6.2}$$

Since $\mathbf{C}_\mathrm{e}$ relies on the external RTF vector $\mathbf{h}_\mathrm{e}$, it has to be estimated. As discussed earlier, a blind RTF vector estimation using the CW method will lead to a biased RTF vector estimate due to the presence of the interferer, as shown for CW in Section 4.3. It has also been shown in Section 4.5 that a single enhanced channel (i.e., with a higher input SIR) can lead to a better RTF vector estimate by guiding the estimation. Therefore, a pre-filtered signal of the L-GSC can be exploited to obtain

a better estimate of $\mathbf{h}_{\mathrm{e}}$, where in principle, there are two available signals which can be used for an improved RTF vector estimation, namely the speech reference Y_{f} or the output signal Z_{a} of the L-GSC. In the GSC-ENR-1, a joint optimization of the filters $\mathbf{g}_{\mathrm{a}}$ and $\mathbf{g}_{\mathrm{e}}$ is considered, meaning that the signal Z_{a} will not be available. Note that the filter $\mathbf{v}_{\mathrm{a}}$ is renamed since it is not equivalent with the filter in the L-GSC. The joint optimization is depicted in Fig. 6.1, as well as the RTF vector estimation using the speech reference Y_{f}, i.e., using the $(M_{\mathrm{e}}+1)\times(M_{\mathrm{e}}+1)$-dimensional pre-processed noisy covariance matrices

$$\mathbf{R}_{y,z} = \mathcal{E}\{[Y_{\mathrm{f}},\, \mathbf{y}_{\mathrm{e}}^{T}]^{T}[Y_{\mathrm{f}}^{*},\, \mathbf{y}_{\mathrm{e}}^{H}]\}\,, \tag{6.3}$$

and the pre-processed noise covariance matrix $\mathbf{R}_{n,z}$ (defined similarly to (6.3)). For the matrices $\mathbf{R}_{y,z}$ and $\mathbf{R}_{n,z}$, the estimated quantities $\hat{\mathbf{R}}_{y,z}$ and $\hat{\mathbf{R}}_{n,z}$ are assumed to be available, as discussed in Section 4.1. The vector $\hat{\mathbf{h}}_{\mathrm{e}}$ is estimated by applying CW to the matrices $\hat{\mathbf{R}}_{y,z}$ and $\hat{\mathbf{R}}_{n,z}$, giving the $(M_{\mathrm{e}}+1)$-dimensional RTF vector $\hat{\mathbf{h}}_{z}$, which consists of a first entry equal to 1, corresponding to the speech reference Y_{f} or respectively the first channel of the LMA, respectively, and the vector $\hat{\mathbf{h}}_{\mathrm{e}}$, i.e.,

$$\hat{\mathbf{h}}_{z} = \begin{bmatrix} 1 \\ \hat{\mathbf{h}}_{\mathrm{e}} \end{bmatrix}. \tag{6.4}$$

Estimating the vector $\hat{\mathbf{h}}_{\mathrm{e}}$ by exploiting the pre-filtered local speech reference Y_{f} can directly be related to the third condition shown in the validation of the bias analysis (Section 4.5), where a particularly good channel (here Y_{f} as it is an enhanced version of Y_1) is used to estimate the RTF vector but is discarded afterwards.
It should be noted that in the case of a fixed beamformer $\mathbf{f}_{\mathrm{a}}$, when not designed to be distortionless as in (5.1), a compensation for the introduced distortion must be applied to (6.4), as done in [17]. When designing the fixed beamformer to introduce a known distortion, i.e., not distortionless, the target speech component in Y_{f} would not be the same as in the input signal Y_1. This would lead to an RTF vector estimate $\hat{\mathbf{h}}_{z}$ that does not match the speech component at the input, which, as described, requires for a compensation of this distortion. Here, however, $\mathbf{f}_{\mathrm{a}}$ is a matched filter and thus distortionless.
For the joint filter optimization of $\mathbf{g}_{\mathrm{a}}$ and $\mathbf{g}_{\mathrm{e}}$, the stacked filter vector

$$\mathbf{g} = \begin{bmatrix} \mathbf{g}_{\mathrm{a}} \\ \mathbf{g}_{\mathrm{e}} \end{bmatrix} \tag{6.5}$$

is considered. The respective optimization problem, aiming at minimizing the residual noise and interferer in the speech reference Y_{f}, reads

$$\min_{\mathbf{g}} \mathcal{E}\{|\mathbf{f}^H\mathbf{y} - \mathbf{g}^H\mathbf{C}^H\mathbf{y}|^2\}\,, \tag{6.6}$$

with the $M \times (M-1)$-dimensional blocking matrix $\mathbf{C}$, defined as

$$\mathbf{C} = \left[\begin{array}{c|c} \mathbf{C}_\mathrm{a} & \begin{array}{c} -\hat{\mathbf{h}}_\mathrm{e}^H \\ \mathbf{0}_{(M_\mathrm{a}-1)\times M_\mathrm{e}} \end{array} \\ \hline \mathbf{0}_{M_\mathrm{e}\times(M_\mathrm{a}-1)} & \mathbf{I}_{M_\mathrm{e}\times M_\mathrm{e}} \end{array}\right]. \tag{6.7}$$

The operation $\mathbf{f}^H\mathbf{y}$ yields the speech reference Y_f as in the L-GSC, for which the joint filter vector is defined as

$$\mathbf{f} = \left[\begin{array}{c} \mathbf{f}_\mathrm{a} \\ \mathbf{0}_{M_\mathrm{e}\times 1} \end{array}\right]. \tag{6.8}$$

The solution of (6.6) for the joint filter vector $\mathbf{g}$ is given by

$$\mathbf{g} = \left(\mathbf{C}^H\mathbf{R}_y\mathbf{C}\right)^{-1}\mathbf{C}^H\mathbf{R}_y\mathbf{f} \tag{6.9}$$

For the filtering, $\mathbf{g}_\mathrm{a}$ and $\mathbf{g}_\mathrm{e}$ are separated again as using (6.5), such that the output Z of the GSC-ENR-1 is given by

$$Z = Y_\mathrm{f} - \mathbf{g}_\mathrm{a}^H\mathbf{u}_\mathrm{a} - \mathbf{g}_\mathrm{e}^H\mathbf{u}_\mathrm{e}\,. \tag{6.10}$$

Note that, as for the L-GSC, the MPDR implementation of the GSC-ENR-1 can lead to target cancellation if there is speech leakage in the noise-and-interferer references $\mathbf{u}_\mathrm{a}$ and $\mathbf{u}_\mathrm{e}$, unlike an MVDR implementation which, however, would not allow to actively suppress the interferer.

6.1.2 GSC with External Noise References Type 2

A second structure which aims at creating external noise-and-interferer references, is the GSC-ENR-2. In contrast to the GSC-ENR-1, the output signal Z_a of the L-GSC is used for the RTF vector estimation of the eMics instead of the speech reference Y_f. Especially if there is no speech leakage into the local noise-and-interferer references $\mathbf{u}_\mathrm{a}$, the SIR in Z_a is expected to be higher than in Y_f, which allows for an even better RTF vector estimation. The processing scheme of the GSC-ENR-2 is shown in Fig. 6.2, where it can also be seen that RTF vector estimation using Z_a has the consequence that the filters $\mathbf{v}_\mathrm{a}$ and $\mathbf{v}_\mathrm{e}$ are optimized separately.
Similarly to the GSC-ENR-1, the blocking matrix $\mathbf{C}_\mathrm{e}$ is completed by means of the external RTFs as in (6.1). The vector $\hat{\mathbf{h}}_\mathrm{e}$ can be obtained by applying CW to the

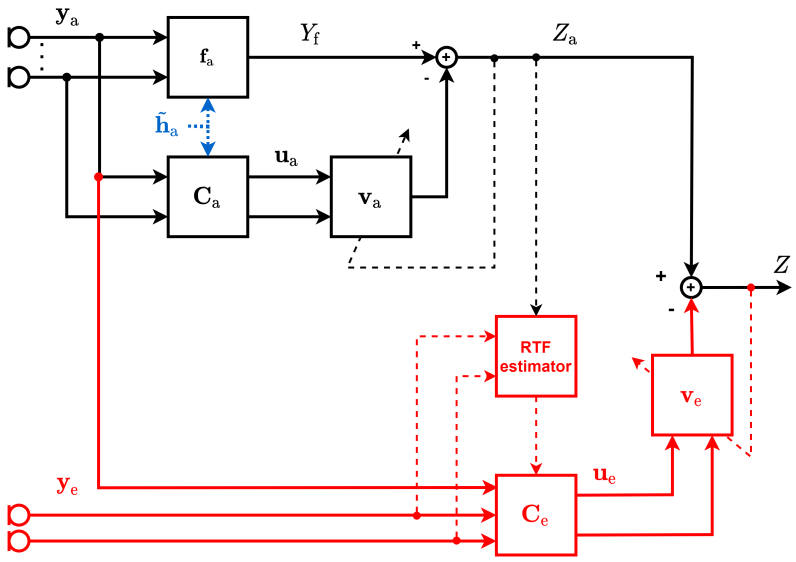

Fig. 6.2: Block diagram of the GSC-ENR-2, incorporating the eMics by creating external noise-and-interferer references. The RTF vector estimation is performed on Z_a and $\mathbf{y}_\mathrm{e}$ using CW.

estimates $\hat{\mathbf{R}}_{y,z}$ and $\hat{\mathbf{R}}_{n,z}$ of the pre-processed noisy covariance matrix

$$\mathbf{R}_{y,z} = \mathcal{E}\{[Z_\mathrm{a},\, \mathbf{y}_\mathrm{e}^T]^T [Z_\mathrm{a}^*,\, \mathbf{y}_\mathrm{e}^H]\} \tag{6.11}$$

and the pre-processed noise covariance matrix $\mathbf{R}_{n,z}$ (defined similarly to (6.11)). The obtained estimated RTF vector $\hat{\mathbf{h}}_z$ can be decomposed into the reference entry and the vector $\hat{\mathbf{h}}_\mathrm{e}$ as in (6.4). Since the fixed beamformer $\mathbf{f}_\mathrm{a}$ is designed to be distortionless, also the target speech component in the signal Z_a is assumed to be distortionless, which means that, as for the GSC-ENR-1, no distortions must be compensated.

The external noise-and-interferer references $\mathbf{u}_\mathrm{e}$ are obtained as in (6.2). $\mathbf{u}_\mathrm{e}$ is subsequently used to minimize the remaining noise and interferer in the output signal Z_a of the L-GSC. Hence, the filter $\mathbf{v}_\mathrm{e}$ is optimized by minimizing the cost function

$$\min_{\mathbf{v}_\mathrm{e}} \mathcal{E}\{|Z_\mathrm{a} - \mathbf{v}_\mathrm{e}^H \mathbf{u}_\mathrm{e}|^2\}\,. \tag{6.12}$$

The solution to (6.12) is given by

$$\mathbf{v}_\mathrm{e} = \hat{\mathbf{R}}_{\mathbf{u}_\mathrm{a}}^{-1} \hat{\mathbf{R}}_{\mathbf{u}_\mathrm{a},Z_\mathrm{a}}\,, \tag{6.13}$$

where $\hat{\mathbf{R}}_{\mathbf{u}_\mathrm{a},Z_\mathrm{a}}$ is an estimate of the cross-correlation between the external noise-and-interferer references $\mathbf{u}_\mathrm{e}$ and the L-GSC output Z_a, defined as

$$\mathbf{R}_{\mathbf{u}_\mathrm{a},Z_\mathrm{a}} = \mathcal{E}\{\mathbf{u}_\mathrm{e} Z_\mathrm{a}^*\}\,. \tag{6.14}$$

The matrix $\hat{\mathbf{R}}_{\mathbf{u}_\mathrm{a}}$ is an estimate of the covariance matrix of $\mathbf{u}_\mathrm{e}$, defined as

$$\begin{aligned}\mathbf{R}_{\mathbf{u}_\mathrm{a}} &= \mathcal{E}\{\mathbf{u}_\mathrm{e}\mathbf{u}_\mathrm{e}^H\} \\ &= \mathbf{C}_\mathrm{e}^H \mathbf{E}_{1,\mathrm{e}}^T \mathbf{R}_y \mathbf{E}_{1,\mathrm{e}} \mathbf{C}_\mathrm{e}\,,\end{aligned} \tag{6.15}$$

with $\mathbf{E}_{1,\mathrm{e}}$ the $M \times (M_\mathrm{e}+1)$ dimensional selection matrix for the first microphone and the eMics, defined as

$$\mathbf{E}_{1,\mathrm{e}} = \left[\begin{array}{c|c} 1 & \\ \mathbf{0}_{(M_\mathrm{a}-1)\times 1} & \mathbf{0}_{M_\mathrm{a}\times M_\mathrm{e}} \\ \hline \mathbf{0}_{M_\mathrm{e}\times 1} & \mathbf{I}_{M_\mathrm{e}\times M_\mathrm{e}} \end{array}\right]. \tag{6.16}$$

The output signal Z of the GSC-ENR-2 is given by

$$Z = Z_\mathrm{a} - \mathbf{v}_\mathrm{e}^H \mathbf{u}_\mathrm{e}\,. \tag{6.17}$$

As also mentioned for the GSC-ENR-1, an MPDR is required to suppress the interferer, but it comes with the risk of speech cancellation if there is speech leakage in the noise-and-interferer reference $\mathbf{u}_\mathrm{a}$ and $\mathbf{u}_\mathrm{e}$.

6.2 GSC with External References

The idea of the GSC-ER is to incorporate the eMics by means of a joint MPDR beamformer of the eMic signals $\mathbf{y}_\mathrm{e}$ and the L-GSC output signal Z_a. The respective processing scheme is shown in Fig. 6.3. When comparing the GSC-ER with the blind MPDR in Section 3, the incorporation of the L-GSC can be seen as a local pre-filtering operation giving one signal (namely Z_a) with a particularly high SIR (compared to the input signals $\mathbf{y}_\mathrm{a}$) which, as shown in Section 4.3, leads to an RTF vector estimate that might favor the target speaker over the interferer. Technically, the L-GSC in the GSC-ER could also be implemented as a closed form MPDR beamformer using the a-priori RTF vector $\tilde{\mathbf{h}}_\mathrm{a}$, since only its output signal Z_a, not its intermediate signal stages, is exploited. However, for a better (visual) comparability to the other extended GSC structures, the L-GSC is kept instead of condensing it into an MPDR block for the LMA.

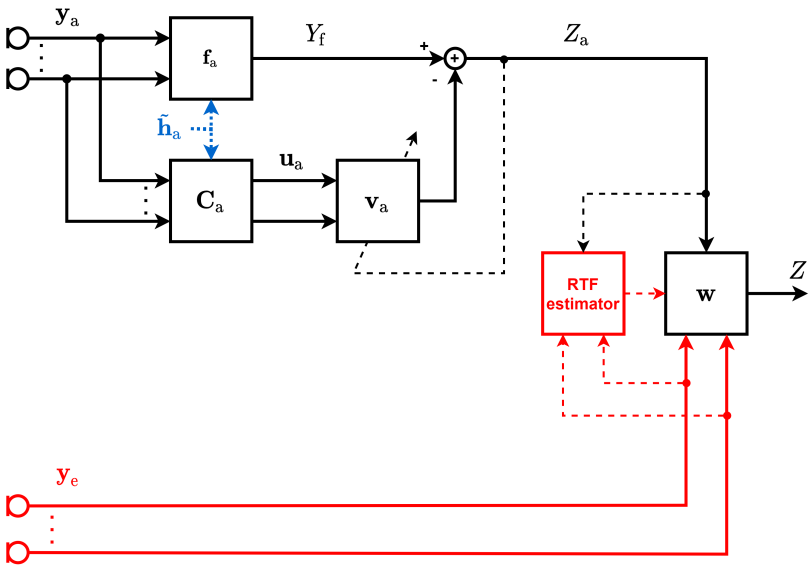

Fig. 6.3: Block diagram of the GSC-ER, incorporating the eMic in a joint beamformer $\mathbf{w}$ with Z_a. The RTF vector estimation is performed on Z_a and $\mathbf{y}_\mathrm{e}$ using CW.

The joint MPDR beamformer on output signal Z_a of the L-GSC and eMic signals $\mathbf{y}_\mathrm{e}$ can be written as

$$\mathbf{w} = \frac{\hat{\mathbf{R}}_{y,z}^{-1}\hat{\mathbf{h}}_z}{\hat{\mathbf{h}}_z^H\hat{\mathbf{R}}_{y,z}^{-1}\hat{\mathbf{h}}_z}, \tag{6.18}$$

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

$$\mathbf{R}_{y,z} = \mathcal{E}\{[Z_\mathrm{a},\, \mathbf{y}_\mathrm{e}^T]^T[Z_\mathrm{a}^*,\, \mathbf{y}_\mathrm{e}^H]\} \tag{6.19}$$

and the pre-processed noise covariance matrix $\mathbf{R}_{n,z}$ is defined similarly. As in the GSC-ENR-2, the RTF vector $\hat{\mathbf{h}}_z$ is estimated using the matrices $\hat{\mathbf{R}}_{y,z}$ and $\hat{\mathbf{R}}_{n,z}$ using the CW method. However, here, the entire vector $\hat{\mathbf{h}}_z$ is used and not only the entries corresponding to the eMics. Therefore, the same external RTFs as in the GSC-ENR-2 structure are obtained, but they are used in a joint beamformer instead of completing a blocking matrix. Note that, as discussed in Section 3, for the joint beamformer $\mathbf{w}$, it does not matter if it is implemented as an MVDR beamformer or as an MPDR beamformer, since CW is used for the RTF vector estimation. Only for consistency with the other filters, an MPDR beamformer is chosen in (6.18). The output signal Z of the GSC-ER is given by

$$Z = \mathbf{w}^H \begin{bmatrix} Z_\mathrm{a} \\ \mathbf{y}_\mathrm{e} \end{bmatrix} . \tag{6.20}$$

Even though it does not matter for the joint beamformer $\mathbf{w}$ whether an MVDR or an MPDR beamformer is chosen, the L-GSC's performance might suffer due to an RTF vector mismatch of the a-priori RTF vector $\tilde{\mathbf{h}}_\mathrm{a}$ and the true target RTF vector of the LMA, which subsequently also influences the RTF vector estimation for the joint beamformer. Therefore, it is expected that an RTF vector mismatch of the a-priori RTF vector also affects the performance of the GSC-ER.

6.3 GSC with External Speech References

Like the GSC-ENR-1, the GSC-ESR was proposed in [17], there referred to as the generalized eigenvalue decomposition (GEVD)-based method. The mathematical formulation of this structure is changed here from an MVDR implementation (as in [17]) to an MPDR implementation, such that not only noise but also interferer can potentially be suppressed. The structure in itself, shown in Fig. 6.4, however, is the same as in [17].

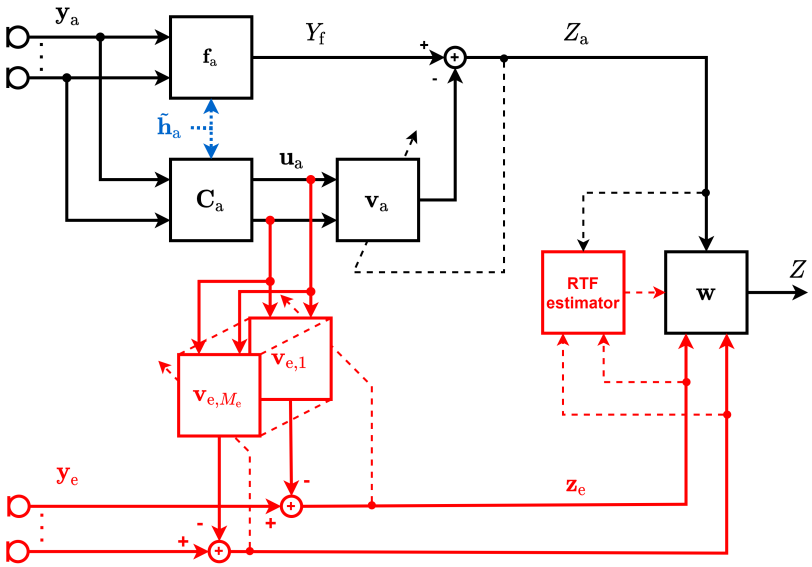

Fig. 6.4: Block diagram of the GSC-ESR, incorporating the eMics in a joint beamformer $\mathbf{w}$ with Z_a after pre-filtering with the local noise-and-interferer references. The RTF vector estimation is performed on Z_a and $\mathbf{z}_\mathrm{e}$ using CW.

In the processing scheme in Fig. 6.4, a clear similarity with the GSC-ER can be observed: The L-GSC can here also be seen as a pre-filtering of the local input signal vector $\mathbf{y}_\mathrm{a}$ which is used in a joint MPDR beamformer in conjunction with the - now

pre-processed - eMic signals, in contrast to the GSC-ER where the un-processed eMic signals are used.

The basic idea is to pre-process the noisy external microphone signals $\mathbf{y}_{\mathrm{e}}$ by means of the filters $\mathbf{v}_{\mathrm{e},m_{\mathrm{e}}}$, $m_{\mathrm{e}} \in \{1, \ldots, M_{\mathrm{e}}\}$, which aim at cancelling correlated components between the local noise-and-interferer $\mathbf{u}_{\mathrm{a}}$ and the eMic signals in $\mathbf{y}_{\mathrm{e}}$. The respective optimization problem to obtain the filters $\mathbf{v}_{\mathrm{e},m_{\mathrm{e}}}$ reads

$$\min_{\mathbf{v}_{\mathrm{e},m_{\mathrm{e}}}} \mathcal{E}\{|Y_{\mathrm{e},m_{\mathrm{e}}} - \mathbf{v}_{\mathrm{e},m_{\mathrm{e}}}^{H}\mathbf{u}_{\mathrm{a}}|^2\}\,. \tag{6.21}$$

The solution of (6.21) is given as

$$\mathbf{v}_{\mathrm{e},m_{\mathrm{e}}} = \left(\mathbf{C}_{\mathrm{a}}^{H}\hat{\mathbf{R}}_{y,\mathrm{a}}\mathbf{C}_{\mathrm{a}}\right)^{-1}\mathbf{C}_{\mathrm{a}}^{H}\mathbf{E}_{\mathrm{a}}\hat{\mathbf{R}}_{y}\mathbf{e}_{\mathrm{e},m_{\mathrm{e}}}\,, \tag{6.22}$$

where $\mathbf{e}_{\mathrm{e},m_{\mathrm{e}}}$ is a selection vector for the m_{e}-th eMic and $\mathbf{E}_{\mathrm{a}}$ is the selection matrix for the LMA defined in (2.14). The pre-processed eMic signals

$$Z_{\mathrm{e},m_{\mathrm{e}}} = Y_{\mathrm{e},m_{\mathrm{e}}} - \mathbf{v}_{\mathrm{e},m_{\mathrm{e}}}^{H}\mathbf{u}_{\mathrm{a}}\,, \tag{6.23}$$

are the elements of the vector $\mathbf{z}_{\mathrm{e}}$. When now assuming an ideal a-priori RTF vector, i.e., there is no speech leakage into the local noise-and-interferer references $\mathbf{u}_{\mathrm{a}}$, only the undesired components are reduced in the eMic signals, leading to a higher SIR in the pre-processed eMic signals $\mathbf{z}_{\mathrm{e}}$ than in the un-processed eMic signals $\mathbf{y}_{\mathrm{e}}$. The exact amount of interferer or noise that is cancelled by the pre-filtering operation depends on the correlation properties of the noise between the local noise-and-interferer references and the eMic signals and the INR in the noise-and-interferer references. However, generally, the interferer is partially suppressed in $\mathbf{z}_{\mathrm{e}}$ allowing for an improved RTF vector estimation in comparison to the GSC-ER. The downside of the pre-processing operation is the risk of target cancellation if there is speech leakage into the local noise-and-interferer references $\mathbf{u}_{\mathrm{a}}$. In that case, the GSC-ER might be favorable over the GSC-ESR, since the target cannot be cancelled in the eMic signals by the filters $\mathbf{v}_{\mathrm{e},m_{\mathrm{e}}}$.

The pre-processed eMic signals $\mathbf{z}_{\mathrm{e}}$ are then used in a joint MPDR beamformer with the L-GSC output signal Z_{a}, i.e.,

$$\mathbf{w} = \frac{\hat{\mathbf{R}}_{y,z}^{-1}\hat{\mathbf{h}}_{z}}{\hat{\mathbf{h}}_{z}^{H}\hat{\mathbf{R}}_{y,z}^{-1}\hat{\mathbf{h}}_{z}}\,, \tag{6.24}$$

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

$$\mathbf{R}_{y,z} = \mathcal{E}\{[Z_{\mathrm{a}},\, \mathbf{z}_{\mathrm{e}}^T]^T [Z_{\mathrm{a}}^*,\, \mathbf{z}_{\mathrm{e}}^H]\}\,. \tag{6.25}$$

The pre-processed noise covariance matrix $\mathbf{R}_{n,z}$ is defined similarly to (6.25). The $(M_{\mathrm{e}}+1)$-dimensional RTF vector estimate $\hat{\mathbf{h}}_z$ is obtained by applying CW to the estimates of the pre-processed covariance matrices $\hat{\mathbf{R}}_{y,z}$ and $\hat{\mathbf{R}}_{n,z}$.
The output of the GSC-ESR is given by

$$Z = \mathbf{w}^H \begin{bmatrix} Z_{\mathrm{a}} \\ \mathbf{z}_{\mathrm{e}} \end{bmatrix}. \tag{6.26}$$

6.4 Summary of Extended GSC Structures

To summarize and compare the considered extended GSC structures and state-of-the-art beamforming structures from Section 3 and Section 5, a general overview of the blind MPDR beamformer, L-GSC, GSC-ENR-1 and GSC-ENR-2, GSC-ER and GSC-ESR structures is given in this section.

The blind MPDR beamformer (completely blind algorithm)

- does not exploit the a-priori RTF vector $\tilde{\mathbf{h}}_{\mathrm{a}}$.
- includes an RTF vector estimation for **all** microphones using CW on the unprocessed noisy input signals.
- performs poorly in terms of interferer reduction if the interferer is the dominant speaker (shown in Section 4.3).
- performs well in terms of noise and interferer reduction if the target speaker is the dominant speaker (shown in Section 4.3).

The L-GSC (completely informed algorithm)

- exploits the a-priori RTF vector $\tilde{\mathbf{h}}_{\mathrm{a}}$ and may therefore suffer if $\tilde{\mathbf{h}}_{\mathrm{a}}$ does not correspond to the true target RTF vector.
- does not incorporate the eMics, which limits the spatial sampling of the sound field.
- does not incorporate any RTF vector estimation and is therefore not prone to an estimation bias at low input SIRs.

The GSC-ENR-1 and GSC-ENR-2 (semi-informed algorithms)

- exploit $\tilde{\mathbf{h}}_{\mathrm{a}}$ for pre-filtering local input signals.

- exploit enhanced local signals to guide the RTF vector estimation for eMics.
- create external noise references $\mathbf{u}_\mathrm{e}$ by complete an external blocking matrix $\mathbf{C}_\mathrm{e}$ using the estimated external RTF vector.
- differ in terms of their RTF vector estimation: GSC-ENR-1 uses the speech reference Y_f and the GSC-ENR-2 uses the output of the L-GSC Z_a to guide the RTF vector estimation for the eMics.
- differ in terms of filter optimization: Joint optimization of local and external filters in the GSC-ENR-1 and separate, cascaded optimization of filters in the GSC-ENR-2.
- might lead to strong speech distortions, since both structures 1) might suffer from a mismatch between $\tilde{\mathbf{h}}_\mathrm{a}$ and the true target RTF vector and 2) might misestimate the external RTF vector.
- make it difficult to predict which structure will perform better due to a complex interplay of different RTF vector estimation and different filter optimization.

The GSC-ER (semi-informed algorithm)

- exploits $\tilde{\mathbf{h}}_\mathrm{a}$ for pre-filtering local input signals.
- exploits output signal Z_a of the L-GSC for the RTF vector estimation of the eMics.
- uses Z_a and the eMic signals $\mathbf{y}_\mathrm{e}$ in a joint MPDR beamformer.
- is expected to obtain a better RTF vector estimate than the blind MPDR due to an improved SIR in one channel (i.e., using Z_a instead of the noisy input $\mathbf{y}_\mathrm{a}$).
- might not perform well at low SIRs since the interferer might be the dominant source in the eMics.

The GSC-ESR (semi-informed algorithm)

- exploits $\tilde{\mathbf{h}}_\mathrm{a}$ for pre-filtering local input signals.
- enhances eMic signals by pre-filtering the local noise-and-interferer reference, which improves the SIR in the pre-processed signals $\mathbf{z}_\mathrm{e}$ compared to the input signals $\mathbf{y}_\mathrm{e}$.
- exploits enhanced local signal Z_a and pre-processed eMic signals $\mathbf{z}_\mathrm{e}$ for the RTF vector estimation of the eMics.
- uses the output signal of the L-GSC and the pre-processed eMic signals in a joint MPDR beamformer.

- is expected to obtain a better RTF vector estimate than the GSC-ER due to an improved SIR in eMic signals.
- might suffer strongly from RTF mismatch between $\tilde{\mathbf{h}}_{\mathrm{a}}$ and the true target RTF vector due to pre-processing of eMic signals which might cancel the target speech in $\mathbf{z}_{\mathrm{e}}$.

7 Evaluation

In this section, the considered extended GSC structures are evaluated in terms of noise and interferer reduction and target speech distortion using real-world recordings. Due to the large number of requirements to the audio material for this evaluation, i.e., the microphone configuration with an LMA and eMics, two distinct talkers from different directions, stationary background noise, knowledge about the LMA's exact geometry to generate artificial steering vectors or RTF vectors with a certain (controllable) mismatch and oracle knowledge about all signal components, typical data sets that are often used for extensive evaluations are not suitable here. The evaluation is therefore only performed on one particular internal database (also used in [32]) where all these conditions are fulfilled. The obtained results are therefore not quite generalizable, since they only reflect the performance in one particular acoustic scenario but give an overall impression of the potential of the different algorithms.
In Section 7.1, the used audio material and the acoustic scenario are described. In Section 7.2, the implementation and framework parameters are described. In Section 7.3, the evaluation results are presented, where first, the case of an ideal a-priori RTF vector and second, the case of an anechoic a-priori RTF vector is considered.

7.1 Recording Setup and Conditions

The signals used in the evaluation were recorded in the *Variable Acoustics Laboratory* at the University of Oldenburg. The laboratory has a size of $(7 \times 6 \times 2.7)$ m and a changeable reverberation time, here set to approximately 350 ms. The microphone configuration is shown in Fig. 7.1. All sources and microphones were spatially stationary during the entire recording. The LMA consisted of binaural hearing aids with two microphones per ear, mounted on a KEMAR head-and-torso simulator (HATS). In addition to the $M_{\mathrm{a}} = 4$ LMA microphones, two eMics were placed approximately 1.5 m away from the HATS, as depicted in Fig. 7.1. The front microphone on the left side on the HATS was chosen as the reference channel.
The target speaker was a male English speaker, played back via a loudspeaker. The target speaker was placed approximately 2 m away from the HATS at an angle of about 35° to the right of the HATS with a distance of about 0.5 m to one of the eMics. The interferer was a female English speaker, played back via a loudspeaker. The interferer was placed approximately 2 m away from the HATS at an angle of about 35° to the left of the HATS with a distance of about 0.5 m to the other eMic. The noise was generated by four loudspeakers facing the corners of the room, playing

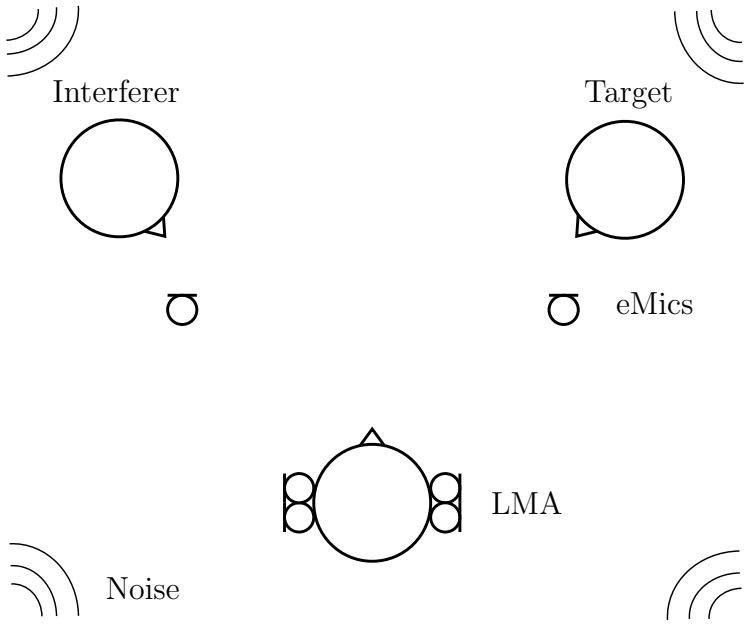

Fig. 7.1: Acoustic scenario and configuration for evaluation the evaluation of the considered algorithms. As the LMA binaural hearing aids on a HATS are used and two eMics in front of the target speaker and the interferer are included.

back different versions of multi-talker babble noise. The resulting noise field had the property to be quasi-diffuse and was therefore almost completely uncorrelated between the LMA microphones and the eMics. The magnitude-squared coherence (MSC) of the noise field for two microphone pairs is depicted in Fig. 7.2: Once between the front microphone on the left hearing aid and the rear microphone on the left hearing aid (Ch. 1 and Ch. 2) and once between the front microphone on the left hearing aid and the eMic on the right side (Ch. 1 and Ch. 5). In Fig. 7.2 the MSC for the two microphone pairs is plotted over frequency f, including a moving average over 10 frequency bins for a smoother representation. It can clearly be seen for the case of two nearby microphones that the MSC is rather high (almost 1) for low frequencies and then drops for higher frequencies. For the second microphone pair, which is much further apart, the MSC is constantly low meaning that it can be assumed that the noise between the LMA microphones and the eMics is almost completely uncorrelated for all frequencies. The observed curves match the expectation for a diffuse sound field, since small distances (Ch. 1 and Ch. 2) lead to highly correlated signals in the low frequency range, while large distances (Ch. 1 and Ch. 5) lead to almost completely uncorrelated signals.

The recordings were performed at a sampling frequency of 48 kHz and down-sampled to 16 kHz for the processing and evaluation. All signal components, i.e., target speech, interferer and noise, were recorded separately and mixed subsequently at

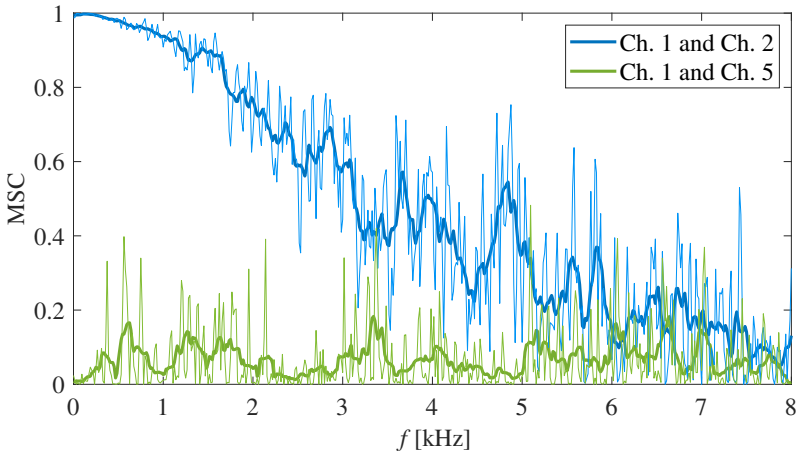

Fig. 7.2: MSC between first and second channel and first and fifth channel respectively with moving average of the MSC for two microphone pairs: The front microphone on the left hearing aid and the rear microphone on the left hearing aid (Ch. 1 and Ch. 2) and the front microphone on the left hearing aid and the eMic on the right side (Ch. 1 and Ch. 5).

different input SIRs (SIR_{in} = [-10, 0, 10] dB) and different input SNRs (SNR_{in} = [-10, 0, 10] dB).

Two different a-priori RTF vectors $\tilde{\mathbf{h}}_{\text{a}}$ are used in these simulations: One is an ideal RTF vector obtained from the measured RIR of the target speaker, which is used in all considered structures. The other a-priori RTF vector is obtained from an anechoic RIR database [33] which was recorded using the same devices as in the simulations. The anechoic a-priori RTF vector for and azimuth of 35° and elevation of 0° is used to simulate a realistic RTF vector mismatch.
The algorithms that are evaluated are the two baseline systems and the four extended GSC structures presented Section 6. The baseline systems are the L-GSC using the respective a-priori RTF vector and a blind MPDR beamformer which uses all microphones and the entire RTF vector is estimated blindly using CW. The external RTF vectors in the extended GSC structures are also estimated using CW on the pre-processed signals, while the a-priori RTF vector is used for the LMA.

7.2 Implementation and Parameters

For the evaluation of the considered algorithms, all filtering operations are done in the STFT domain. The settings of the STFT framework are a frame length of 1024 samples, corresponding to 64 ms, an overlap of 50%, i.e., 512 samples, and a square-root-Hann window for analysis and synthesis are chosen. All filters are implemented closed form (i.e., using the Wiener filter solution).

A batch implementation is chosen, where the covariance matrices are estimated as an average over the entire signal. From this, it follows that the entire signal is processed using the same filter vectors, which do not adapt over time. Furthermore, as already mentioned in Section 4.1, no VAD algorithm is included, since its performance would be strongly affected by different SNRs and SIRs. Estimation errors of covariance matrices might yield results that are difficult to interpret and from which no clear conclusions about the algorithms' performance could be drawn. Therefore, oracle knowledge about the noise is assumed, meaning that good estimates of $\mathbf{R}_y$ and $\mathbf{R}_n$ are available, which are used for the RTF estimation and filter optimization.
The blindly optimized filters are applied to all signal components separately to access the respective algorithm's performance (see shadow filtering in (2.16) - (2.18)).
The evaluation of the objective measures is done in the time domain, where broadband measures are obtained, as described in Section 2.1. All measures are computed in sequences where the target speaker and the interferer were active simultaneously.
Lastly, it should be mentioned that, even though the STFT framework described above leads to perfect reconstruction of the input signal, the clean input signal components are processed with the framework, to compensate for windowing effects in the beginning or the end of the processed signals. Put differently, the input signal components are once transformed to the STFT domain and then directly back transformed to the time domain without any other processing steps. The metrics like SD or SNR improvement, etc., are then computed using these forth and back transformed input signal components to assess the input powers of the different signal components.

7.3 Results

In this section, the results of the described simulations are presented. First, the ideal case is investigated where the a-priori RTF vector is obtained from the RIR of the target speaker. These results are shown for all considered algorithms. Second, the influence of an RTF vector mismatch is investigated, i.e., where the anechoic RTF vector is used in the L-GSC and the extended GSC structures. In this case, extended structures, which already perform worse than the baseline in the first, ideal condition, are not regarded anymore.

7.3.1 Using an Oracle A-Priori RTF Vector

In this first section of results, the case of an ideal a-priori RTF vector is considered. Fig. 7.3 - 7.5 show the simulation results for the three different input SIRs. The

first three panels shows the SIR improvement, the SNR improvement, and the SINR improvement, respectively. The last panel shows the SD of the target speech. On the x-axis in all panels, the input SNR can be found.

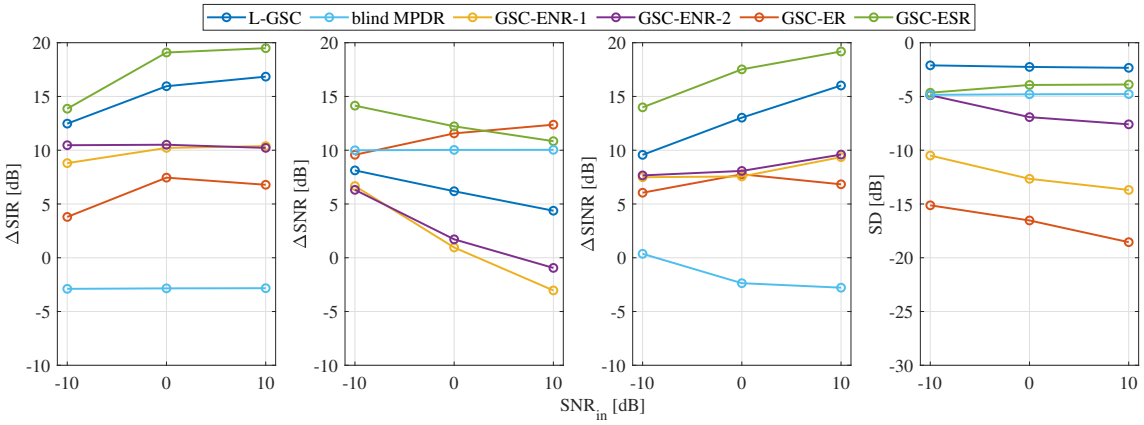

Fig. 7.3: Evaluation results for an ideal a-priori RTF vector obtained from the target RIR and an input SIR of -10 dB. The performance measures ΔSIR, ΔSNR, ΔSINR and SD are shows per panel, each for the two baseline systems (L-GSC and blind MPDR) and the four extended GSC structures.

At the input SIR of -10 dB, i.e., in Fig. 7.3, it can be seen that the blind MPDR performs the worst in terms of SIR improvement (about -3 dB), while performing well in terms of SNR improvement (about 10 dB) for all input SNRs. The overall SINR improvement, however, has partly negative scores of 0 dB to -3 dB. For all input SNRs, the SD score is at about -5 dB. The results go along with the theoretical analysis in Section 4.2, where it was shown that the dominant source (here the interferer) is preserved better, which in this case leads to a negative ΔSIR score.
The ΔSIR scores of the L-GSC range from 12 dB (at $\mathrm{SNR}_{\mathrm{in}}$ = -10 dB) to 17 dB (at $\mathrm{SNR}_{\mathrm{in}}$ = 10 dB) and the ΔSNR scores range from 8 dB to 4 dB, respectively. The SD score for the L-GSC is the best among all considered algorithms with an almost constant value of -2 dB. Note that despite using the ideal RTF vector (obtained from the target RIR) as the a-priori RTF vector $\tilde{\mathbf{h}}_{\mathrm{a}}$, speech distortions can be observed. This is caused by the SD score being a measure based on the power of in- and output signal, where also reverberation is included. Since the frames of the STFT framework are shorter than the RIR, to some extent dereverberation is performed by the beamformer, which induces speech distortions. When taking the theoretical results from [34,35] into account, the L-GSC performs as expected: Since the noise field is diffuse and therefore contains uncorrelated parts, the correlation between the local interferer-and-noise references $\mathbf{u}_{\mathrm{a}}$ and the speech reference Y_{f} is largest at high input SNR where only little noise is present. In literature, it has been shown that for a higher correlation of the undesired component, the performance of

the filter $\mathbf{v}_{\mathrm{a}}$ is larger and more undesired components can be cancelled. Vice versa, a low correlation between $\mathbf{u}_{\mathrm{a}}$ and Y_{f} (here in the presence of much noise) means that less undesired components can be cancelled.
Both the extended GSC structures, which create external noise references by means of completing a blocking matrix, i.e., the GSC-ENR-1 and the GSC-ENR-2, score lower values in all metrics in comparison to the L-GSC. The SD scores range from -5 dB to -8 dB for the GSC-ENR-2 and from -10 dB to -14 dB for the GSC-ENR-1. For both structures, the ΔSIR scores are almost constant at around 10 dB and the ΔSNR scores range from 6 dB down to negative values of -1 dB or respectively -3 dB at an input SNR of 10 dB. The resulting ΔSINR scores are also lower than for the L-GSC. These results already indicate that the GSC-ENR-1 and the GSC-ENR-2 are not suitable for joint noise and interferer reduction, since they appear constantly perform worse than the baseline system in adverse conditions. A possible explanation (which holds for both structures) is that the estimation of the external RTF vector is strongly biased due to the interferer being the dominant source at an SIR of -10 dB. This leads to a significant amount of target speech leakage into the external noise-and-interferer references $\mathbf{u}_{\mathrm{e}}$, which then causes target speech cancellation. However, note that in [17], the GSC-ENR-1 was proposed only for noise reduction and hence, an MVDR implementation was used. Though the GSC-ENR-1 and GSC-ENR-2 might not be suitable for joint noise and interferer reduction, they might yet lead to improvements in different acoustic scenarios where only noise is present.
Also the GSC-ER performs worse than the baseline system in terms of SIR improvement and SINR improvement, which can be related to the very low scores in terms of SD, where values of -15 dB to -18 dB are obtained. The SNR improvement is rather good, ranging between 10 dB and 12 dB. However, the GSC-ER does not seem to be suitable for joint noise and interferer reduction at low input SIRs, which can be explained by a strongly biased estimation of the external RTF vector due to the low SIR in the unprocessed eMic signals.
The GSC-ESR performs best in almost all conditions: The SD score is at -4 dB to -5 dB for all input SNRs and thus the best for all extended structures. In terms of ΔSIR and ΔSNR it outperforms the L-GSC by 4 dB and 5 dB, respectively, which leads to the overall best performance in terms of ΔSINR. These results already indicate a benefit of the pre-filtering operation of the GSC-ESR compared to the GSC-ER. Particularly at high input SNRs, the interferer can be cancelled well by the filters $\mathbf{v}_{\mathrm{e},m_{\mathrm{e}}}$. The explanation can be found when considering that the noise filed is diffuse and therefore uncorrelated between the local noise-and-interferer references

$\mathbf{u}_\mathrm{a}$ and the eMic signals $\mathbf{y}_\mathrm{e}$ due to the large distance between the LMA and the eMics (as shown in Fig. 7.2). A lower input SNR therefore leads to a lower correlation between these signals which prevents the filters $\mathbf{v}_{\mathrm{e},m_\mathrm{e}}$ from cancelling the interferer in the eMic signals. Generally, the results for this adverse scenario already indicate that the GSC-ESR is the only extended GSC structure that has the potential to increase the performance of the L-GSC.

In Fig. 7.4, the results for an input SIR of 0 dB are shown. The first overall trend that can be observed is that the differences in the performance among the different algorithms become smaller: While the ΔSINR scores at an input SIR of -10 dB ranged from -3 dB (for the blind MPDR beamformer) to 19 dB (for the GSC-ESR), at an input SIR of 0 dB, the range goes from 1 dB (for the blind MPDR beamformer) to 15 dB (for the GSC-ESR). Similarly, the maximal speech distortion is now obtained by the GSC-ENR-1 (-8 dB) instead of the GSC-ER (-18 dB) at an input SIR of -10 dB.

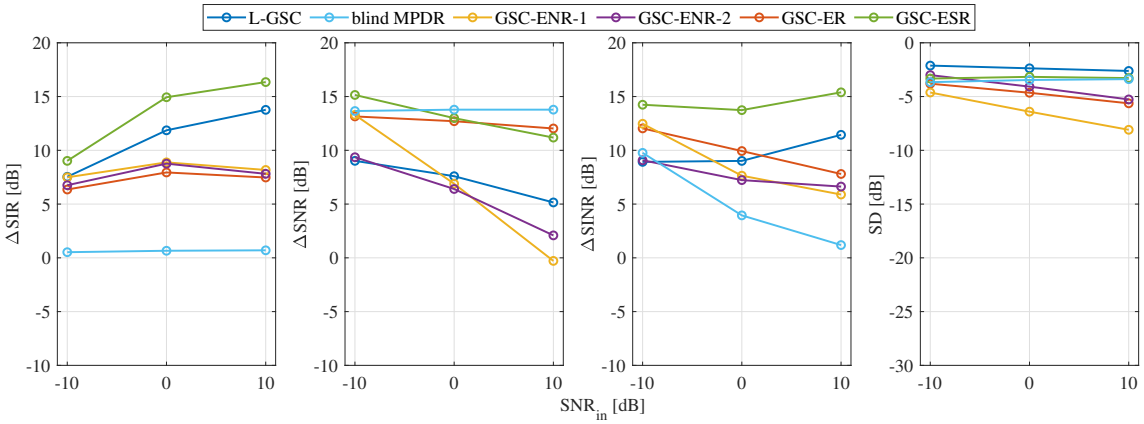

Fig. 7.4: Evaluation results for an ideal a-priori RTF vector obtained from the target RIR and an input SIR of 0 dB. The performance measures ΔSIR, ΔSNR, ΔSINR and SD are shows per panel, each for the two baseline systems (L-GSC and blind MPDR) and the four extended GSC structures.

Again, looking at the baseline systems first, it can be seen that the blind MPDR performs better than at a lower SIR. The SD score improves from -5 dB to about -3 dB. The ΔSIR score is relatively constant between 0 dB and 1 dB for all input SNRs. The results are in line with the theoretical findings in Section 4.3 that at an input SIR of 0 dB both speech sources, i.e., target speaker and interferer, influence the RTF vector estimation equally much and are therefore also cancelled equally much, leading to a ΔSIR score of 0 dB. The ΔSNR scores are constantly at about 14 dB for the blind MPDR beamformer, which is among the best of all algorithms for all input SNRs. The ΔSINR scores, however, are still lower than for all other considered algorithms.

For the L-GSC, the results are similar as for an input SIR of -10 dB, though the ΔSIR scores are generally lower (by 3 dB to 5 dB), but the overall tendencies of the curves for the different input SNRs remain the same. The SD scores are, as before, at about -2 dB for all input SNRs. The decreased in the SIR improvement (compared to an input SIR of -10 dB) can be explained by a lower INR in the local noise-and-interferer references $\mathbf{u}_\mathrm{a}$, which leads to a lower correlation between $\mathbf{u}_\mathrm{a}$ and the speech reference Y_f.
The GSC-ENR-1 and the GSC-ENR-2 perform better at an input SIR of 0 dB than at an input SIR of -10 dB. However, their performance in terms of ΔSIR and ΔSNR is still below the performance of the L-GSC for all input SNRs except for -10 dB, where the GSC-ENR-1 outperforms the L-GSC in terms of ΔSNR, while having the same ΔSIR score of almost 8 dB. In terms of SD, the GSC-ENR-1 and the GSC-ENR-2 introduce larger distortions than the L-GSC.
The GSC-ER, which performed rather poorly in terms of SIR improvement and SD for an input SIR of -10 dB, still scores rather low values in these metrics relative to other algorithms, but yet shows a much better performance than before. For ΔSIR, values of 6 dB to 8 dB are obtained and the SD scores lie between -4 dB and -6 dB. The overall performance in terms of ΔSINR is around 12 dB (at SNR_in = -10 dB) to 8 dB (at SNR_in = 10 dB), which is partly better than the L-GSC, which is caused by the good performance in terms of ΔSNR, where values of 12 dB to 13 dB are obtained. At an input SIR of 0 dB, it can therefore be concluded that the GSC-ER performs well in terms of noise reduction, but not in terms of interferer reduction.
The results obtained for the GSC-ESR show that this structure outperforms all other algorithms in terms of ΔSIR, even though the score drops by about 3 dB in comparison to the condition with an input SIR of -10 dB. In terms of ΔSNR, the GSC-ESR performs similarly to the blind MPDR beamformer and the GSC-ER. The ΔSNR scores are about 1 dB higher than at an input SIR of -10 dB. The SD score improves by about 1 dB in comparison to the lower input SIR, which is therefore the best among all structures which use the eMics. The results can be explained as follows: The 3 dB decrease in terms of SIR improvement can be explained similarly as for the L-GSC. The local noise-and-interferer references $\mathbf{u}_\mathrm{a}$ are less correlated with the eMic signals $\mathbf{y}_\mathrm{e}$ than at an input SIR of -10 dB, meaning that less interferer can be cancelled by the filters $\mathbf{v}_{\mathrm{e},m_\mathrm{e}}$, which leads to a lower SIR improvement in the pre-processed eMic signals $\mathbf{z}_\mathrm{e}$. The generally higher input SIR yet allows for a better RTF vector estimation of $\hat{\mathbf{h}}_z$, which subsequently leads to better SD scores that at an input SIR of -10 dB.

The results for an input SIR of 10 dB are shown in Fig. 7.5. The overall tendency of all algorithms to perform more similarly at higher input SIRs continues at the input SIR of 10 dB. The extended GSC structures all outperform the L-GSC in terms of ΔSIR and ΔSNR in almost all conditions. Only in terms of SD, the L-GSC still performs best, though the performance of all algorithms is very similar for this metric (ranging from -2 dB to -4 dB).

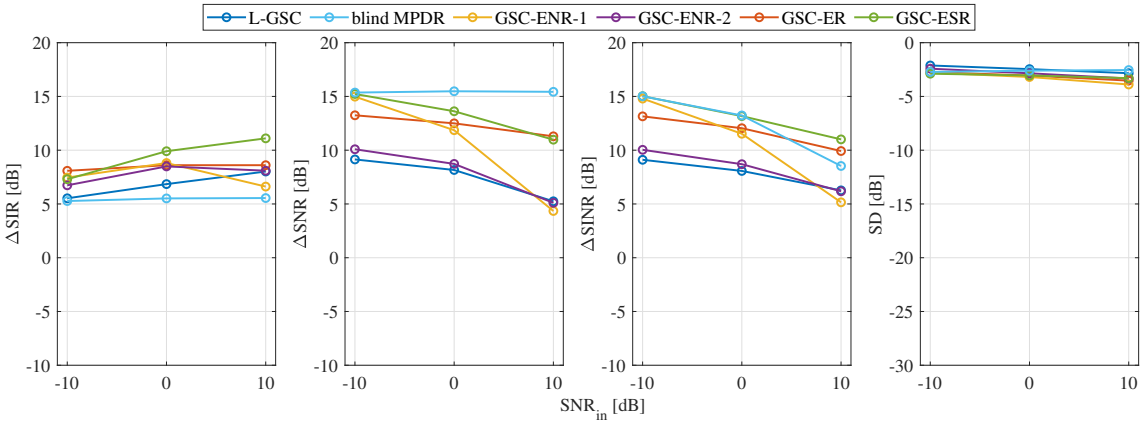

Fig. 7.5: Evaluation results for an ideal a-priori RTF vector obtained from the target RIR and an input SIR of 10 dB. The performance measures ΔSIR, ΔSNR, ΔSINR and SD are shows per panel, each for the two baseline systems (L-GSC and blind MPDR) and the four extended GSC structures.

The best performing algorithms are the blind MPDR in terms of ΔSNR and the GSC-ESR in term of ΔSIR and ΔSINR. It can also be seen that the GSC-ER performs more similarly to the GSC-ESR at high input SIRs. This can be explained by the decreasing INR in the local noise-and-interferer references $\mathbf{u}_\mathrm{a}$ for higher input SIRs, which leads to a lower correlation between $\mathbf{u}_\mathrm{a}$ and the eMic signals $\mathbf{y}_\mathrm{e}$, such that their interferer reduction performance by the filters $\mathbf{v}_{\mathrm{e},m_\mathrm{e}}$ is very limited.

Summarizing the first part of the evaluation, the simulation results for an ideal a-priori RTF vector $\tilde{\mathbf{h}}_\mathrm{a}$ showed that the GSC-ENR-1 and the GSC-ENR-2 are not suitable for joint noise and interferer reduction, as they perform worse than the L-GSC, which acts as the baseline system. These two algorithms are therefore not taken into consideration in the following part, where an RTF vector mismatch for the L-GSC is investigated. Also the GSC-ER does not perform well at low input SIRs, but performs well at high input SIRs, where the performance becomes similar to the performance of the GSC-ESR. The GSC-ESR is found to be the best performing among the extended GSC structures, as it outperforms the L-GSC in all conditions in terms of noise and interferer reduction. The GSC-ESR is therefore the only structure

which incorporates the eMics that seems promising for joint noise and interferer reduction. Though not performing well at low input SIRs, the GSC-ER is considered in the next part of the evaluation since it allows for a more detailed investigation of the pre-filtering operation of the GSC-ESR. This becomes particularly interesting as speech leakage into the local noise-and-interferer references $\mathbf{u}_\mathrm{a}$ might occur due to an RTF vector mismatch of the true target RTF vector and the anechoic a-priori RTF vector, considered in the next part of the evaluation. Furthermore, the blind MPDR beamformer is considered in the nect part of the evaluation, as it does not rely on the a-priori RTF vector and can therefore show limitations or disadvantages of the extended GSC structures.

7.3.2 Using an Anechoic A-Priori RTF Vector

In the second part of the evaluation, the case of an approximate, anechoic a-priori RTF vector $\tilde{\mathbf{h}}_\mathrm{a}$ is considered, which is more realistic since an a-priori RTF obtained from the RIR of the target speaker is not available in practice. As mentioned above, only the blind MPDR beamformer, the L-GSC, the GSC-ESR and the GSC-ER are considered. Although the results for the blind MPDR beamformer do not change between the first and the second part of the evaluation, the results are taken into account to stress the consequences of an RTF vector mismatch at high input SIRs. The results are structured as before, where each figure of Fig. 7.6 - 7.8 show the results for a different input SIR.

In Fig. 7.6, the results for an input SIR of -10 dB are shown. It can be seen that the performance of all algorithms using the anechoic a-priori RTF vector achieve a lower performance in terms of SIR and SNR improvement, which is caused by the larger amount of SD (2 dB worse for the L-GSC, 4 dB worse for the GSC-ESR and 5 dB to 10 dB worse for the GSC-ER) in comparison to the condition where an ideal a-priori RTF vector was used (see Fig. 7.3).
Similarly to the SD scores, the ΔSIR, ΔSNR and ΔSIR scores for the three structures exploiting the a-priori RTF vector decrease compared to the condition where an ideal a-priori RTF vector was used. As before, the blind MPDR beamformer and the GSC-ER perform worse than the L-GSC at an input SIR of -10 dB in terms of interferer reduction, while performing similarly or even better in terms of noise reduction. The overall SINR improvement for the blind MPDR beamformer and the GSC-ER is still at least 5 dB lower than for the L-GSC. The tendencies of more interferer being cancelled at high input SNRs, which was found for the case of an ideal a-priori RTF vector, can also be seen here for the L-GSC and the GSC-ESR.

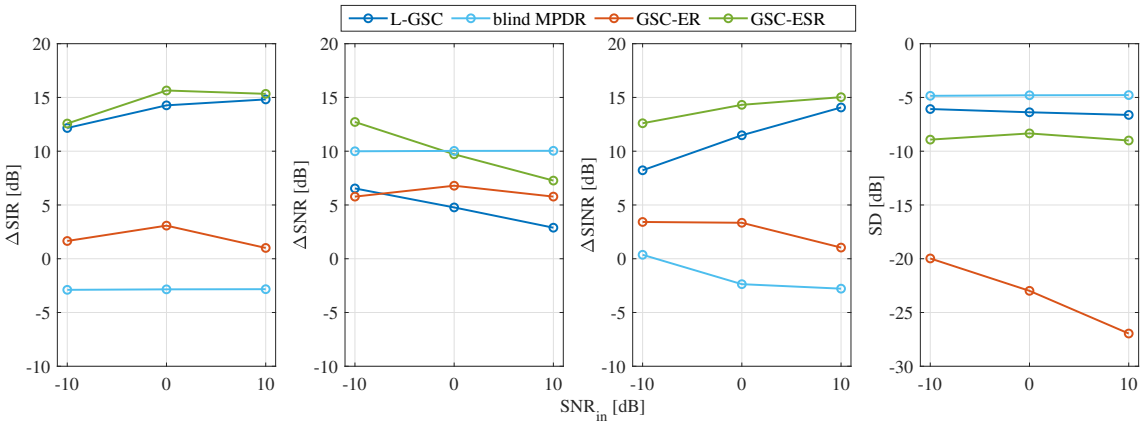

Fig. 7.6: Evaluation results for an approximate, anechoic a-priori RTF vector obtained from a database and an input SIR of -10 dB. The performance measures ΔSIR, ΔSNR, ΔSINR and SD are shows per panel, each for the two baseline systems (L-GSC and blind MPDR), the GSC-ER and the GSC-ESR.

Despite the decrease in performance, the GSC-ESR is still found to be the best performing algorithm in terms of ΔSIR (about 1 dB better than the L-GSC) and for most input SNRs also in terms of ΔSNR, except for an input SNR of 10 dB, where the blind MPDR beamformer is better by 2 dB. Therefore, even with an RTF vector mismatch, the L-GSC and the GSC-ESR perform well in adverse conditions. The incorporation of the eMics in the GSC-ESR leads to SNR improvements compared to the L-GSC.

The results for an input SIR of 0 dB and 10 dB are shown in Fig. 7.7 and Fig. 7.8, respectively. In contrast to the results for an ideal a-priori RTF vector in Fig. 7.4 and Fig. 7.5, the different algorithms do not perform more similarly to each other at higher input SIR, but rather more differently: While the GSC-ER and the blind MPDR improve their performance in terms of all metrics, the L-GSC and the GSC-ESR perform worse than (or maximally equally as) at an input SIR of -10 dB in terms of all metrics.

It can also be observed that at an input SIR of 0 dB (see Fig. 7.7), the SIR improvement stays almost constant for the L-GSC and the GSC-ESR over all input SNRs. At an input SIR of 10 dB, it even strongly decreases towards higher input SNRs. Similar observations can be made for the SD scores. This larger amount of SD which occurs at high input SNRs can be explained by speech leakage into the local noise-and-interferer references $\mathbf{u}_\mathrm{a}$. At high input SIRs, this has the consequence that the target 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. As the cancellation of coherent sources depends

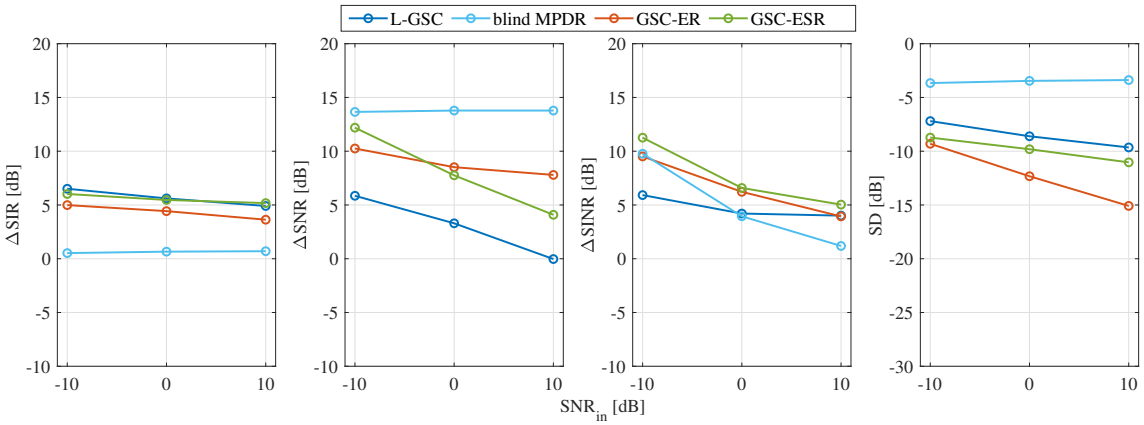

Fig. 7.7: Evaluation results for an approximate, anechoic a-priori RTF vector obtained from a database and an input SIR of 0 dB. The performance measures ΔSIR, ΔSNR, ΔSINR and SD are shows per panel, each for the two baseline systems (L-GSC and blind MPDR), the GSC-ER and the GSC-ESR.

on the overall correlation between $\mathbf{u}_\mathrm{a}$ and Y_f or $\mathbf{y}_\mathrm{e}$, respectively, which decreases at lower input SNRs, the target is cancelled the most if only little noise is present, i.e., at high input SNRs. This effect becomes particularly pronounced at an input SIR of 10 dB and an input SNR of 10 dB, where negative scores are obtained for ΔSIR by the L-GSC and the GSC-ESR. At high input SIRs, it seems beneficial to blindly estimate the entire RTF vector, i.e., as for the blind MPDR beamformer, which yields the best SD score and also outperforms all other algorithms in terms of SIR and SNR improvement at an input SIR of 10 dB.

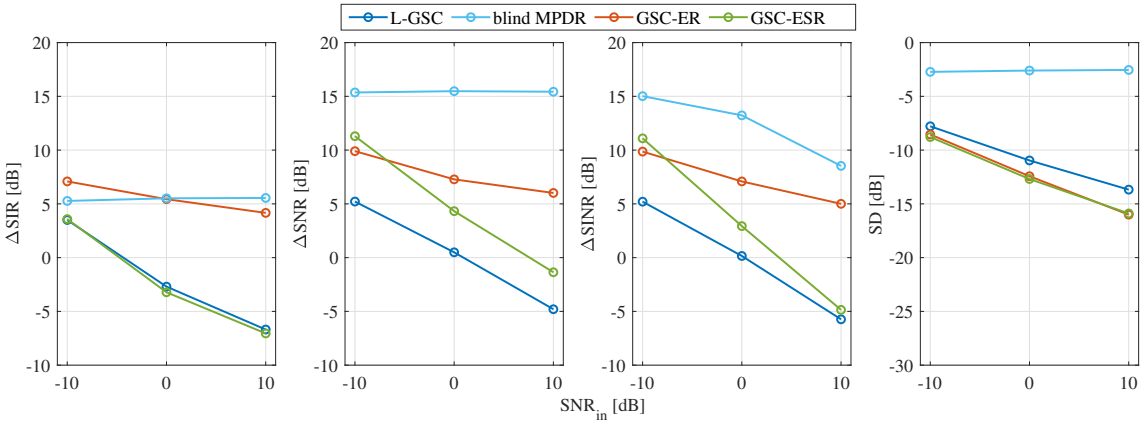

Fig. 7.8: Evaluation results for an approximate, anechoic a-priori RTF vector obtained from a database and an input SIR of 10 dB. The performance measures ΔSIR, ΔSNR, ΔSINR and SD are shows per panel, each for the two baseline systems (L-GSC and blind MPDR), the GSC-ER and the GSC-ESR.

Generally speaking, the L-GSC suffers from a mismatch of the a-priori RTF vector $\tilde{\mathbf{h}}_{\mathrm{a}}$ and the true target RTF vector, especially at high input SIRs due to speech leakage into the noise-and-interferer references $\mathbf{u}_{\mathrm{a}}$. At high input SIRs, the target is therefore the most coherent source between $\mathbf{u}_{\mathrm{a}}$ and the speech reference Y_{f} and is cancelled. Similarly, the GSC-ESR is affected by the pre-filtering operation using the pre-filters $\mathbf{v}_{\mathrm{e},m_{\mathrm{e}}}$ applied to $\mathbf{u}_{\mathrm{a}}$, which cancels the target speech in the eMic signals $\mathbf{y}_{\mathrm{e}}$, such that the SIR in the pre-processed eMic signals $\mathbf{z}_{\mathrm{e}}$ is lower than in $\mathbf{y}_{\mathrm{e}}$. This explanation is supported by the fact that the GSC-ER performs better at high input SIRs than the L-GSC and the GSC-ESR since it can exploit a high input SIR in the external microphones which are not pre-processed in the GSC-ER. In the condition of a high input SIR, it therefore seems beneficial to apply fewer pre-processing steps, since the blind MPDR beamformer performs even better than the GSC-ER. However, at low input SIR, the pre-processing of the eMic signals seems beneficial, despite the RTF vector mismatch. While the blind MPDR and the GSC-ER perform poorly at an input SIR of -10 dB, the L-GSC leads to good SIR and SNR improvements which are even outperformed by the GSC-ESR.

8 Conclusions and Outlook

In this thesis, an acoustic scenario with one target speaker, an interfering speaker and background noise was considered, where the goal was to perform joint noise and interferer reduction. To this end, hearing aids in conjunction with external microphones (eMics) were used. To show the influence of the interferer on the estimation of the target relative transfer function (RTF) vector, a theoretical bias analysis of a state-of-the-art RTF vector estimator, namely the covariance whitening (CW) method, was performed. It was shown that the estimated RTF vector is a linear combination of the target RTF vector and the interferer RTF vector, where the weighting of the two RTF vectors depends on the multi-channel signal-to-interferer ratio (SIR). A closed form expression was found for the ratio of the scaling factors for the target and the interferer RTF vector. In a validation of this theoretical bias analysis artificial data was used to investigate the influence of this biased RTF vector estimation on an minimum power distortionless response (MPDR) beamformer. In agreement with the theoretical findings, using the biased RTF vector estimate in an MPDR beamformer led to a low performance in term of noise and interferer reduction and introduced strong target speech distortions at low input SIRs. At high input SIRs using the blindly estimated RTF vector in an MPDR beamformer led to a good results for all metrics. These results strongly indicate that a blind target RTF vector estimation in the presence of an interferer is not suitable, when aiming at joint noise and interferer reduction. Motivated by this finding, it was assumed that the target RTF vector of the hearing aid microphones - or in practice rather an approximate of it - is known. Using this so-called a-priori RTF vector, a local generalized sidelobe canceller (L-GSC) was considered for joint noise and interferer reduction.

To improve the performance of the L-GSC, eMics were included in the processing. To this end, four different extended GSC structures were presented. All extended structures include an RTF vector estimation for the target RTFs corresponding to the eMics, for which the spatially pre-filtered signals of the L-GSC are exploited. In an experimental evaluation using real-world recordings, it was found that two of the considered structures which create external noise-and-interferer references from the eMic signals, namely the GSC with external noise references type 1 (GSC-ENR-1) and the GSC with external noise references type 2 (GSC-ENR-2) did not perform well, even in conditions where an ideal a-priori RTF vector for the L-GSC was considered. The GSC with external references (GSC-ER) which processes the eMic signals together with the L-GSC output in a joint MPDR beamformer also did not perform well in adverse conditions with a low input SIR. Since these three structures

did not lead to SIR and signal-to-noise ratio (SNR) improvements compared to the L-GSC in scenarios with a low input SIR, they are not suitable for joint noise and interferer reduction. The last extended GSC structure considered in this thesis was the GSC with external speech references (GSC-ESR), which was adopted from [17]. The major adjustment for this structure made in this thesis was to change its formulation such that not only noise but also interferer could be suppressed. This structure aims at an improved RTF vector estimation by pre-processing the eMic signals using the local noise-and-interferer references. In the evaluation, this structure outperformed the L-GSC and the other extended GSC structures in terms of SIR and SNR improvement, especially in adverse conditions with a low input SIR. Though not performing well if there was a mismatch of the a-priori RTF vector and the true target RTF vector, the GSC-ESR could still outperform the L-GSC in adverse conditions. At high input SIRs the L-GSC and the GSC-ESR both performed poorly due to speech target leakage into the local noise-and-interferer references. In these conditions the L-GSC and the GSC-ESR were outperformed by the blind MPDR beamformer and the GSC-ER. However, in conditions, where joint noise and interferer reduction is required the most, i.e., at low input SIRs, the GSC-ESR seemed to perform reliably and has great potential for being further developed for larger robustness against RTF vector mismatches.

To give an outlook about future work, the weakness of the GSC-ESR to introduce target speech cancellation caused by RTF vector mismatches at high input SIRs is addressed, which gives rise to several possible approaches:
1) A very general next step is a theoretical performance analysis of the GSC-ESR to gain understanding about the exact processes and limitations of this structure in different acoustic scenarios. Such an analysis subsequently allows to tackle different problems that might occur in this structure, such as target speech cancellation, and therefore helps designing even more sophisticated structures based on the GSC-ESR. The points below and their consequences or benefits could then directly be analyzed in theory.
2) Instead of a classical voice activity detection (VAD) algorithm, a neural network based approach can be considered to detect the activity of the target speaker. This would allow for accessing the covariance matrix of all undesired sources (i.e., the matrix $\mathbf{R}_v$), which does not contain the target speaker, preventing it from being cancelled by the filters $\mathbf{v}_\mathrm{a}$ and $\mathbf{v}_{\mathrm{e},m_\mathrm{e}}$ at high input SIRs.
3) A detection of the dominant speaker, i.e., distinguishing whether the target speaker or the interferer is the dominant source, could allow for adapting the pro-

cessing strategy depending on the acoustic scenario. This could, for example, mean that if the interferer is found to be the dominant source, the GSC-ESR is used for the processing, while in scenarios where the target is the dominant source, i.e., the interferer does not lead to a strong bias in the blind RTF vector estimation, the blind MPDR beamformer is used for the processing. A gradual transition between processing strategies could also be considered when both sources have approximately equal powers. However, an adaptation of the processing strategy depending on the dominant source would require an algorithm that identifies the dominant speech source.

Appendix A - Equivalence of MVDR and MPDR Beamformer

As motioned in Section 3, the MVDR and the MPDR beamformer can be distinguished by means of their objective function. In literature [6, 36], it has been discussed that the two beamformers are identical in the case of no estimation errors of the noisy covariance matrix $\mathbf{R}_y$ or the noise covariance matrix $\mathbf{R}_n$ and if the ideal RTF vector is known. In practice, however, this is usually not the case and often the MVDR is chosen, since it is considered to be more robust towards RTF vector mismatches [37].
In the following, a different, more practical, analysis is performed. Here, no ideally estimated covariance matrices or perfect RTF vectors are assumed, meaning that only the properties of $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{R}}_n$ to be Hermitian and positive definite must be fulfilled. This also implies that the following analysis holds for any quasi signal model (regardless the number or characteristics of present sources). With these assumptions, the equivalence of the MVDR and the MPDR is now analyzed for the case that CW (see Section 4.2) is used to obtain an RTF vector estimate to steer the beamformers. The matrices used to estimate the RTF vector and those used in the beamformer are completely identical (as also in practice, unless regularization is performed only in the beamformer).

In the following, a different formulation of the CW method than in Section 4.2 is chosen, namely, where the principal eigenvector $\mathbf{v}_{\max}$ is obtained as $\mathbf{v}_{\max} = \mathcal{P}\{\hat{\mathbf{R}}_n^{-1}\hat{\mathbf{R}}_y\}$. This means that the whitening operation applied to $\hat{\mathbf{R}}_y$ is not performed via the square-root decomposition of $\hat{\mathbf{R}}_n$, but by directly whitening with the entire matrix $\hat{\mathbf{R}}_n$, as in [11]. This means for the estimated RTF vector that

$$\hat{\mathbf{h}}^{\mathrm{CW}} = \frac{1}{N_{\mathrm{rtf}}}\hat{\mathbf{R}}_n\mathbf{v}_{\max}\,, \tag{A.1}$$

where the de-whitening of $\mathbf{v}_{\max}$ is performed with $\hat{\mathbf{R}}_n$ and $N_{\mathrm{rtf}} = \mathbf{e}_1^T\hat{\mathbf{R}}_n\mathbf{v}_{\max}$ is the normalization factor for the RTF vector, where $\mathbf{e}_1$ is a selection vector for the first channel. The associated generalized eigenvalue problem for $\mathbf{v}_{\max}$ reads

$$\hat{\mathbf{R}}_y\mathbf{v}_{\max} = \lambda_1\hat{\mathbf{R}}_n\mathbf{v}_{\max}\,, \tag{A.2}$$

where λ_1 is the largest eigenvalue of the matrix $\hat{\mathbf{R}}_n^{-1}\hat{\mathbf{R}}_y$, which has the properties to be real-valued and larger than 0, which results from the properties of $\hat{\mathbf{R}}_n$ and $\hat{\mathbf{R}}_y$ to be Hermitian and positive definite. To obtain $\mathbf{v}_{\max}$, (A.2) can be rearranged as

$$\mathbf{v}_{\max} = \frac{1}{\lambda_1}\hat{\mathbf{R}}_n^{-1}\hat{\mathbf{R}}_y\mathbf{v}_{\max}\,. \tag{A.3}$$

Using $\hat{\mathbf{h}}^{\mathrm{CW}}$ from (A.1) in the MPDR beamformer in (3.4) and in the MVDR beamformer in (3.3) yields

$$\mathbf{w}_{\mathrm{MPDR}} = N_{\mathrm{rtf}}^{*}\frac{\hat{\mathbf{R}}_y^{-1}\hat{\mathbf{R}}_n\mathbf{v}_{\max}}{\mathbf{v}_{\max}^{H}\hat{\mathbf{R}}_n\hat{\mathbf{R}}_y^{-1}\hat{\mathbf{R}}_n\mathbf{v}_{\max}}\,, \tag{A.4}$$

and

$$\begin{aligned}\mathbf{w}_{\mathrm{MVDR}} &= N_{\mathrm{rtf}}^{*}\frac{\hat{\mathbf{R}}_n^{-1}\hat{\mathbf{R}}_n\mathbf{v}_{\max}}{\mathbf{v}_{\max}^{H}\hat{\mathbf{R}}_n\hat{\mathbf{R}}_n^{-1}\hat{\mathbf{R}}_n\mathbf{v}_{\max}} \\ &= N_{\mathrm{rtf}}^{*}\frac{\mathbf{v}_{\max}}{\mathbf{v}_{\max}^{H}\hat{\mathbf{R}}_n\mathbf{v}_{\max}}\,,\end{aligned} \tag{A.5}$$

where in (A.5), the products $\hat{\mathbf{R}}_n^{-1}\hat{\mathbf{R}}_n = \mathbf{I}_{M\times M}$ cancel out. Now using (A.3) to substitute $\mathbf{v}_{\max}$ in (A.4) yields

$$\begin{aligned}\mathbf{w}_{\mathrm{MPDR}} &= N_{\mathrm{rtf}}^{*}\lambda_1\frac{\hat{\mathbf{R}}_y^{-1}\hat{\mathbf{R}}_n\hat{\mathbf{R}}_n^{-1}\hat{\mathbf{R}}_y\mathbf{v}_{\max}}{\mathbf{v}_{\max}^{H}\hat{\mathbf{R}}_y\hat{\mathbf{R}}_n^{-1}\hat{\mathbf{R}}_n\hat{\mathbf{R}}_y^{-1}\hat{\mathbf{R}}_n\hat{\mathbf{R}}_n^{-1}\hat{\mathbf{R}}_y\mathbf{v}_{\max}} \\ &= N_{\mathrm{rtf}}^{*}\lambda_1\frac{\mathbf{v}_{\max}}{\mathbf{v}_{\max}^{H}\hat{\mathbf{R}}_y\mathbf{v}_{\max}}\,.\end{aligned} \tag{A.6}$$

Considering the MVDR in (A.5) and using the definition of $\mathbf{v}_{\max}$ in (A.3) to substitute only the one $\mathbf{v}_{\max}$ in the denominator yields

$$\begin{aligned}\mathbf{w}_{\mathrm{MPDR}} &= N_{\mathrm{rtf}}^{*}\lambda_1\frac{\mathbf{v}_{\max}}{\mathbf{v}_{\max}^{H}\hat{\mathbf{R}}_n\hat{\mathbf{R}}_n^{-1}\hat{\mathbf{R}}_y\mathbf{v}_{\max}} \\ &= N_{\mathrm{rtf}}^{*}\lambda_1\frac{\mathbf{v}_{\max}}{\mathbf{v}_{\max}^{H}\hat{\mathbf{R}}_y\mathbf{v}_{\max}}\,.\end{aligned} \tag{A.7}$$

Finally, it can be stated that the MVDR and the MPDR beamformer are identical if CW is used for the RTF vector estimation, i.e.

$$\boxed{\mathbf{w}_{\mathrm{MPDR}} = N_{\mathrm{rtf}}^{*}\lambda_1\frac{\mathbf{v}_{\max}}{\mathbf{v}_{\max}^{H}\hat{\mathbf{R}}_y\mathbf{v}_{\max}} = \mathbf{w}_{\mathrm{MVDR}}\,.} \tag{A.8}$$

From this it can be seen that the MVDR and the MPDR beamformer are identical in the case that the RTF vector is estimated using CW. However, for any other RTF vector (e.g., the a-priori RTF vector of the LMA), this equivalence does not hold.

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] E. C. Cherry, "Some experiments on the recognition of speech, with one or two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, Sep. 1953.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] G. Reuven, S. Gannot, and I. Cohen, "Dual-source transfer-function generalized sidelobe canceller," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, pp. 711–727, May 2008.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] N. Gößling, *Binaural Beamforming Algorithms and Parameter Estimation Methods Exploiting External Microphones*. PhD thesis, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany, Oct. 2020.

[15] N. Yousefian and L. P, "A dual-microphone speech enhancement algorithm based on the coherence function," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, pp. 599–609, Feb. 2012.

[16] D. Yee, H. Kamkar-Parsi, R. Martin, and H. Puder, "A noise reduction postfilter 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, Jul. 2017.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, (San Francisco, USA), pp. 11–15, Mar. 2017.

[22] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, pp. 926–935, Aug. 1972.

[23] S. Doclo, *Multi-microphone noise reduction and dereverberation techniques for speech applications*. PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium, May 2003.

[24] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, pp. 1383–1393, May 2012.

[25] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 139, pp. 377–380, Aug. 1992.

[26] 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.

[27] 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.

[28] 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.

[29] E. A. P. Habets, "Room impulse response generator," tech. rep., Technische Universiteit Eindhoven, 2006. RIR generator by E. Habets.

[30] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating non-stationary multisensor signals under a spatial coherence constraint," *Journal of the Acoustical Society of America*, vol. 124, pp. 2911–2917, Nov. 2008.

[31] S. S. Haykin, *Adaptive filter theory*. Pearson Education, 2005.

[32] N. Gößling and S. Doclo, "RTF-based binaural MVDR beamformer exploiting an external microphone in a diffuse noise field," in *Proc. ITG Conference on Speech Communication*, (Oldenburg, Germany), pp. 106–110, Oct. 2018.

[33] 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.

[34] 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.

[35] 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.

[36] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of MVDR and MPDR beamformers," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, (Eilat, Israel), pp. 416–420, Dec. 2010.

[37] H. Cox, "Spatial correlation in arbitrary noise fields with application to ambient sea noise," *The Journal of the Acoustical Society of America*, vol. 54, pp. 1289–1301, Nov. 1973.

Selbstständigkeitserklärung

Hiermit erkläre ich an Eides statt, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutz habe. Außerdem versichere ich, dass ich die allgemeinen Prinzipien wissenschaftlicher Arbeit und Veröffentlichung, wie sie in den Leitlinien guter wissenschaftlicher Praxis der Carl von Ossietzky Universität Oldenburg festgelegt sind, befolgt habe.

Oldenburg, 15.06.2021

Ort, Datum	Unterschrift Wiebke Middelberg