BIAS ANALYSIS OF SPATIAL COHERENCE-BASED RTF VECTOR ESTIMATION FOR ACOUSTIC SENSOR NETWORKS IN A DIFFUSE SOUND FIELD

Wiebke Middelberg and Simon Doclo

University of Oldenburg, Department of Medical Physics and Acoustics
and Cluster of Excellence Hearing4all, Oldenburg, Germany

ABSTRACT

In many multi-microphone algorithms, an estimate of the relative transfer functions (RTFs) of the desired speaker is required. Recently, a computationally efficient RTF vector estimation method was proposed for acoustic sensor networks, assuming that the spatial coherence (SC) of the noise component between a local microphone array and multiple external microphones is low. Aiming at optimizing the output signal-to-noise ratio (SNR), this method linearly combines multiple RTF vector estimates, where the complex-valued weights are computed using a generalized eigenvalue decomposition (GEVD). In this paper, we perform a theoretical bias analysis for the SC-based RTF vector estimation method with multiple external microphones. Assuming a certain model for the noise field, we derive an analytical expression for the weights, showing that the optimal model-based weights are real-valued and only depend on the input SNR in the external microphones. Simulations with real-world recordings show a good accordance of the GEVD-based and the model-based weights. Nevertheless, the results also indicate that in practice, estimation errors occur which the model-based weights cannot account for.

Index Terms— RTF vector estimation, MVDR beamforming, acoustic sensor networks, bias analysis, external microphones

1. INTRODUCTION

In many speech communication applications, such as hearing aids, conferencing and hands-free systems, speech intelligibility is often decreased by the presence of undesired noise sources. Therefore, noise reduction is essential to increase the speech intelligibility [1, 2]. A well-known multi-microphone noise reduction algorithm is the minimum variance distortionless response (MVDR) beamformer [3, 4]. It has been shown that the MVDR beamformer can be steered without requiring the microphone positions to be known by the relative transfer function (RTF) vector of the desired speaker [5, 6]. The RTFs relate the speech component in a reference microphone signal to the speech component in all other microphone signals.

Instead of using a compact microphone array with closely spaced microphones for noise reduction, there has been a recent trend towards using spatially distributed microphones, also referred to as acoustic sensor networks, where the positions of the microphones are typically not known [7–12]. For an acoustic sensor network consisting of a local microphone array (LMA), e.g., binaural hearing aids, and one additional external microphone, a computationally efficient RTF vector estimation method was proposed [13, 14], assuming that the spatial coherence (SC) between the noise component in the external microphone and the LMA is low. A bias analysis in [14] showed that the RTF vector entry corresponding to the external microphone has a real-valued bias (i.e., a systematic
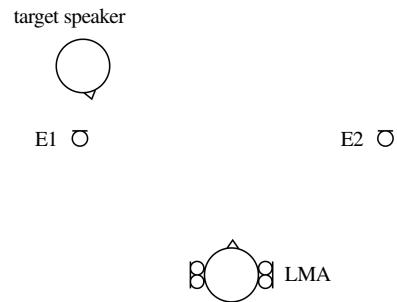
Fig. 1. Acoustic scene with a target speaker, binaural hearing aids (as the LMA) mounted on a dummy head and two external microphones (E1 and E2).

estimation error), which increases with decreasing input signal-to-noise ratio (SNR) in the external microphone. In [15], the SC RTF vector estimation method was extended to exploit multiple external microphones. In the so-called mSNR approach, it was proposed to linearly combine different (biased) RTF vector estimates in order to maximize the output SNR of an MVDR beamformer. It was shown that the optimal combination weights can be computed via a generalized eigenvalue decomposition (GEVD).

As an extension of the bias analysis for the SC method with one external microphone in [14], in this paper we perform a bias analysis for the mSNR approach with multiple external microphones in a diffuse noise field. Using a model of the diffuse sound field, we derive an analytic expression for the optimal combination weights. The model-based weights are found to be real-valued and only depend on the input SNRs in the external microphones. In addition, we show that the biases of the external microphone entries in the RTF vector estimate obtained using the mSNR approach are smaller than the corresponding biases in the RTF vector estimates obtained using the SC method. An experimental evaluation with real-world recordings shows that the real-valued model-based weights and the real value of the generally complex-valued GEVD-based weights are similar. Nevertheless, although the model-based weights are useful for analyzing the bias of the RTF vector entries, simulation results show that in terms of SNR improvement the GEVD-based weights outperform the model-based weights when used in an MVDR beamformer.

2. SIGNAL MODEL AND NOTATION

We consider an acoustic sensor network consisting of an LMA with M_a microphones and M_e external microphones, depicted in Fig. 1, giving a total of $M = M_a + M_e$ microphones. In the short-time Fourier transform (STFT)-domain, the noisy m-th microphone signal is defined as

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

where $X_m(k,l)$ and $N_m(k,l)$ denote the speech and noise component, respectively. For brevity, the frequency bin index k and the frame index

l are omitted in the remainder of this paper. Assuming a multiplicative transfer function model, the speech component can be written as $X_m = H_m X_1$, where H_m denotes the RTF of the target speaker between the m-th microphone and the reference microphone (chosen as the first microphone without loss of generality). The M-dimensional signal vector $\mathbf{y} = [Y_{a,1}, Y_{a,2}, ..., Y_{a,M_a}, Y_{e,1}, Y_{e,2}, ..., Y_{e,M_e}]^T$ contains both the LMA as well as the external microphone signals, where $\{\cdot\}^T$ denotes the transpose operator. The signal vector can be written as

$$\mathbf{y} = \mathbf{x} + \mathbf{n} = \mathbf{h}X_1 + \mathbf{n}, \tag{2}$$

where the speech and noise vector $\mathbf{x}$ and $\mathbf{n}$ are defined similarly to $\mathbf{y}$ and $\mathbf{h}$ denotes the M-dimensional RTF vector of the target speaker, with $H_1 = 1$.

Assuming speech and noise to be uncorrelated, the noisy covariance matrix $\mathbf{R}_y$ is equal to

$$\mathbf{R}_y = \mathcal{E}\{\mathbf{y}\mathbf{y}^H\} = \mathcal{E}\{\mathbf{x}\mathbf{x}^H\} + \mathcal{E}\{\mathbf{n}\mathbf{n}^H\} = \mathbf{R}_x + \mathbf{R}_n, \tag{3}$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator, $\{\cdot\}^H$ denotes the Hermitian transpose operator and $\mathbf{R}_x$ and $\mathbf{R}_n$ are the speech and noise covariance matrices, respectively. Based on (2), the speech covariance matrix is a rank-1 matrix, spanned by the RTF vector, i.e.,

$$\mathbf{R}_x = \phi_{x,1}\mathbf{h}\mathbf{h}^H, \tag{4}$$

where $\phi_{x,m} = \mathcal{E}\{|X_m|^2\}$ denotes the speech power spectral density (PSD) in the m-th microphone. In the remainder of this paper, we will assume that the noise component in any external microphones is uncorrelated with the noise component in any other microphone, i.e.,

$$R_{n,i,j} = 0, \quad \text{if} (i \in \{M_a+1, ..., M\} \vee j \in \{M_a+1, ..., M\}) \wedge i \neq j. \tag{5}$$

Although this assumption obviously does not hold for every acoustic scenario (especially not for coherent interfering sources), this assumption holds well for a diffuse noise field (e.g., babble noise) where all external microphones are placed at a large distance to all other microphones.

The MVDR beamformer minimizing the output noise PSD while preserving the speech component in the reference microphone signal is given by [2, 6]

$$\mathbf{w} = \frac{\mathbf{R}_n^{-1}\mathbf{h}}{\mathbf{h}^H\mathbf{R}_n^{-1}\mathbf{h}}, \tag{6}$$

where the beamformer output signal is equal to $Z = \mathbf{w}^H\mathbf{y}$. The input SNR of the m_e-th external microphone signal and the biased SNR of the beamformer output signal are given by

$$\mathrm{SNR}_{e,m_e} = \frac{\phi_{x,e,m_e}}{\phi_{n,e,m_e}}, \tag{7}$$

$$\mathrm{SNR}_{out}^b = \frac{\mathbf{w}^H\mathbf{R}_y\mathbf{w}}{\mathbf{w}^H\mathbf{R}_n\mathbf{w}} = \frac{\mathbf{w}^H\mathbf{R}_x\mathbf{w}}{\mathbf{w}^H\mathbf{R}_n\mathbf{w}} + 1 = \mathrm{SNR}_{out} + 1, \tag{8}$$

where ϕ_{x,e,m_e} and ϕ_{n,e,m_e} denote the speech and noise PSD in the m_e-th external microphone, respectively, and SNR_{out} denotes the unbiased output SNR.

3. RTF ESTIMATION USING EXTERNAL MICROPHONES

In this section, we present the SC-based RTF vector estimation method for one external microphone proposed in [14] and its extension for multiple external microphones [15]. Exploiting the assumption about spatial correlation of the noise field in (5), M_e different estimates of the RTF vector $\mathbf{h}$ can be efficiently obtained by selecting and normalizing

the column of the noisy covariance matrix corresponding to the external microphone [14], i.e.,

$$\tilde{\mathbf{h}}_{m_e}^{SC} = \frac{\mathbf{R}_y\mathbf{e}_{e,m_e}}{\mathbf{e}_1^T\mathbf{R}_y\mathbf{e}_{e,m_e}}, \quad m_e \in \{1, ..., M_e\}, \tag{9}$$

where $\mathbf{e}_{e,m_e}$ is an M-dimensional selection vector for the m_e-th external microphone. Using the model for $\mathbf{R}_x$ and $\mathbf{R}_n$ in (4) and (5), it can be easily shown that all entries of the RTF vector estimate in (9) are unbiased, except for the entry corresponding to the m_e-th external microphone, which is systematically biased and equal to [14]

$$\boxed{\tilde{H}_{e,m_e}^{SC} = \mathbf{e}_{e,m_e}^T\tilde{\mathbf{h}}_{m_e}^{SC} = \left(1 + \frac{1}{\mathrm{SNR}_{e,m_e}}\right)H_{e,m_e}} \tag{10}$$

This means that the estimated RTF is equal to the true RTF multiplied with a bias term which directly depends on the input SNR in the used m_e-th external microphone.

To decrease the biases of the estimated external RTFs, it was proposed in [15] to linearly combine the M_e RTF vector estimates in (9) i.e.,

$$\tilde{\mathbf{h}}^{mSNR} = \tilde{\mathbf{H}}\boldsymbol{\alpha}, \tag{11}$$

$$\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1^{SC}, \tilde{\mathbf{h}}_2^{SC}, ..., \tilde{\mathbf{h}}_{M_e}^{SC}], \tag{12}$$

where the elements of the M_e-dimensional (complex-valued) weights vector $\boldsymbol{\alpha}$ are constrained to sum up to 1, such that $\tilde{H}_1 = 1$. In the mSNR approach in [15], the optimal weight vector is calculated by maximizing the (biased) output SNR of the RTF-steered MVDR beamformer, i.e.,

$$\max_{\boldsymbol{\alpha}} \mathrm{SNR}_{out}^b(\boldsymbol{\alpha}), \quad \text{s.t.} \quad \mathbf{1}^T\boldsymbol{\alpha} = 1, \tag{13}$$

where $\mathbf{1}$ is an M_e-dimensional vector of ones. By using the MVDR beamformer in (6) with the RTF vector estimate in (11) in (8), the cost function $J(\boldsymbol{\alpha})$ corresponding to the biased output SNR can be written as

$$J(\boldsymbol{\alpha}) = \mathrm{SNR}_{out}^b(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^H\mathbf{A}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^H\mathbf{B}\boldsymbol{\alpha}}, \tag{14}$$

where the $M_e \times M_e$-dimensional matrices $\mathbf{A}$ and $\mathbf{B}$ are defined as $\mathbf{A} = \tilde{\mathbf{H}}^H\mathbf{R}_n^{-1}\mathbf{R}_y\mathbf{R}_n^{-1}\tilde{\mathbf{H}}$ and $\mathbf{B} = \tilde{\mathbf{H}}^H\mathbf{R}_n^{-1}\tilde{\mathbf{H}}$ in [15]. The optimal weighting vector maximizing (14) can hence be calculated using the GEVD of $\mathbf{A}$ and $\mathbf{B}$, where the normalization constraint in (13) is applied subsequently, i.e.,

$$\boldsymbol{\alpha}^{GEVD} = \frac{\mathcal{P}\{\mathbf{B}^{-1}\mathbf{A}\}}{\mathbf{1}^T\mathcal{P}\{\mathbf{B}^{-1}\mathbf{A}\}}, \tag{15}$$

where $\mathcal{P}\{\cdot\}$ denotes the principal eigenvector operator.

4. THEORETICAL BIAS ANALYSIS

In this section, we investigate the biases of the RTF vector estimate $\tilde{\mathbf{h}}^{mSNR}$ in (11) obtained by combining the RTF vector estimates $\tilde{\mathbf{h}}_{m_e}^{SC}$, with $m_e \in \{1, ..., M_e\}$, in (9) using the mSNR approach, i.e., using the weight vector in (15). To that end, we write the matrix of estimated RTF vector estimates $\tilde{\mathbf{H}}$ in (12) in terms of the true RTF vector $\mathbf{h}$ and a bias matrix $\mathbf{E}$, i.e.,

$$\tilde{\mathbf{H}} = \mathbf{H} + \mathbf{E}, \tag{16}$$

where, using (9), $\mathbf{H} = \mathbf{h}\mathbf{1}^T$ and

$$\mathbf{E} = \begin{bmatrix} \mathbf{0}_{M_a \times 1} & \mathbf{0}_{M_a \times 1} & ... & \mathbf{0}_{M_a \times 1} \\ \frac{H_{e,1}}{\mathrm{SNR}_{e,1}} & 0 & ... & 0 \\ 0 & \frac{H_{e,2}}{\mathrm{SNR}_{e,2}} & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & ... & ... & \frac{H_{e,M_e}}{\mathrm{SNR}_{e,M_e}} \end{bmatrix}. \tag{17}$$

By substituting $\tilde{\mathbf{H}}$ in (16) into the cost function $J(\boldsymbol{\alpha})$ in (13), we obtain

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^H (\mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_2^H + \mathbf{A}_3) \boldsymbol{\alpha}}{\boldsymbol{\alpha}^H (\mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_2^H + \mathbf{B}_3) \boldsymbol{\alpha}}, \tag{18}$$

with

$$\begin{aligned}
\mathbf{A}_1 &= \mathbf{1}\,(\mathbf{h}^H \mathbf{R}_n^{-1} \mathbf{R}_y \mathbf{R}_n^{-1} \mathbf{h})\,\mathbf{1}^T, & \mathbf{B}_1 &= \mathbf{1}\,(\mathbf{h}^H \mathbf{R}_n^{-1} \mathbf{h})\,\mathbf{1}^T, \\
\mathbf{A}_2 &= \mathbf{1} \mathbf{h}^H \mathbf{R}_n^{-1} \mathbf{R}_y \mathbf{R}_n^{-1} \mathbf{E}, & \mathbf{B}_2 &= \mathbf{1} \mathbf{h}^H \mathbf{R}_n^{-1} \mathbf{E}, \\
\mathbf{A}_3 &= \mathbf{E}^H \mathbf{R}_n^{-1} \mathbf{R}_y \mathbf{R}_n^{-1} \mathbf{E}, & \mathbf{B}_3 &= \mathbf{E}^H \mathbf{R}_n^{-1} \mathbf{E}.
\end{aligned} \tag{19}$$

We now aim at simplifying (18) by carefully investigating all matrices in (19). The matrices $\mathbf{A}_1$ and $\mathbf{B}_1$ can be written as a scaled matrix of ones, i.e., $\mathbf{A}_1 = (\mathbf{h}^H \mathbf{R}_n^{-1} \mathbf{R}_y \mathbf{R}_n^{-1} \mathbf{h})\,\mathbf{1}\mathbf{1}^T = a_1 \mathbf{1}\mathbf{1}^T$ and $\mathbf{B}_1 = (\mathbf{h}^H \mathbf{R}_n^{-1} \mathbf{h})\,\mathbf{1}\mathbf{1}^T = b_1 \mathbf{1}\mathbf{1}^T$, where a_1 and b_1 are real-valued positive scalars as they can be regarded as the quadratic form of a positive definite matrix.

Since the matrices $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{B}_2$ and $\mathbf{B}_3$ all contain $\mathbf{R}_n^{-1} \mathbf{E}$, we consider this term in more detail. Using (5) and (17), $\mathbf{R}_n^{-1} \mathbf{E}$ is equal to

$$\mathbf{R}_n^{-1} \mathbf{E} = \begin{bmatrix}
\mathbf{0}_{M_a \times 1} & \cdots & \mathbf{0}_{M_a \times 1} \\
\frac{H_{e,1}}{\mathrm{SNR}_{e,1}} \frac{1}{\phi_{n,e,1}} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \frac{H_{e,M_e}}{\mathrm{SNR}_{e,M_e}} \frac{1}{\phi_{n,e,M_e}}
\end{bmatrix}. \tag{20}$$

By realizing that $|H_{e,m_e}|^2 = \phi_{x,e,m_e}/\phi_{x,1}$, the left-hand multiplication of (20) with $\mathbf{h}^H$ yields

$$\mathbf{h}^H \mathbf{R}_n^{-1} \mathbf{E} = (1/\phi_{x,1})\,\mathbf{1}^T. \tag{21}$$

Hence, the matrix $\mathbf{B}_2$ can also be written as a scaled matrix of ones, i.e., $\mathbf{B}_2 = b_2 \mathbf{1}\mathbf{1}^T$ with $b_2 = 1/\phi_{x,1}$ a real-valued positive scalar. Using (3), (4) and the simplification in (21), the matrix $\mathbf{A}_2$ is equal to

$$\begin{aligned}
\mathbf{A}_2 &= \mathbf{1} \mathbf{h}^H \mathbf{R}_n^{-1} \phi_{x,1} \mathbf{h}\mathbf{h}^H \mathbf{R}_n^{-1} \mathbf{E} + \mathbf{1} \mathbf{h}^H \mathbf{R}_n^{-1} \mathbf{E} \\
&= (\mathbf{h}^H \mathbf{R}_n^{-1} \mathbf{h})\,\mathbf{1}\mathbf{1}^T + (1/\phi_{x,1})\,\mathbf{1}\mathbf{1}^T \\
&= (b_1 + b_2)\,\mathbf{1}\mathbf{1}^T,
\end{aligned} \tag{22}$$

i.e., $\mathbf{A}_2$ can also be written as a scaled matrix of ones. Finally, using (17) and (20), the matrix $\mathbf{B}_3$ can be written as

$$\mathbf{B}_3 = \mathbf{E}^H \mathbf{R}_n^{-1} \mathbf{E} = \frac{1}{\phi_{x,1}} \underbrace{\begin{bmatrix}
\frac{1}{\mathrm{SNR}_{e,1}} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \frac{1}{\mathrm{SNR}_{e,M_e}}
\end{bmatrix}}_{\mathbf{S}}, \tag{23}$$

where $\mathbf{S}$ is a full-rank matrix containing the inverse input SNR in the external microphones on its diagonal. Using (3), (4) and (21), the matrix $\mathbf{A}_3$ can be written as

$$\begin{aligned}
\mathbf{A}_3 &= \mathbf{E}^H \mathbf{R}_n^{-1} \phi_{x,1} \mathbf{h}\mathbf{h}^H \mathbf{R}_n^{-1} \mathbf{E} + \mathbf{E}^H \mathbf{R}_n^{-1} \mathbf{E} \\
&= (1/\phi_{x,1})\,\mathbf{1}\mathbf{1}^T + (1/\phi_{x,1})\,\mathbf{S} \\
&= b_2 \mathbf{1}\mathbf{1}^T + b_2 \mathbf{S}.
\end{aligned} \tag{24}$$

By now considering the normalization constraint in (13), it becomes evident that $\boldsymbol{\alpha}^H \mathbf{1}\mathbf{1}^T \boldsymbol{\alpha} = 1$. Together with the derived matrix expressions in the previous paragraph, the cost function in (18) then reduces to

$$J(\boldsymbol{\alpha}) = \frac{a + c(\boldsymbol{\alpha})}{b + c(\boldsymbol{\alpha})}, \tag{25}$$

where $a = a_1 + 2(b_1 + b_2) + b_2$, $b = b_1 + 2b_2$ and

$$c(\boldsymbol{\alpha}) = b_2\,\boldsymbol{\alpha}^H \mathbf{S} \boldsymbol{\alpha}, \tag{26}$$

with $\mathbf{S}$ defined in (23). Since $a \geq b$ and $c(\boldsymbol{\alpha})$ is real-valued and positive, maximizing $J(\boldsymbol{\alpha})$ in (25) corresponds to minimizing $c(\boldsymbol{\alpha})$. Hence, the optimal weights are the solution of the constraint optimization problem

$$\min_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^H \mathbf{S} \boldsymbol{\alpha}, \quad \text{s.t.} \quad \mathbf{1}^T \boldsymbol{\alpha} = 1, \tag{27}$$

i.e.,

$$\boldsymbol{\alpha}^{\mathrm{model}} = \frac{\mathbf{S}^{-1} \mathbf{1}}{\mathbf{1}^T \mathbf{S}^{-1} \mathbf{1}}. \tag{28}$$

By using (23), these weights can be written as

$$\boxed{\boldsymbol{\alpha}^{\mathrm{model}} = \frac{1}{\sum_{m_e'=1}^{M_e} \mathrm{SNR}_{e,m_e'}} \begin{bmatrix} \mathrm{SNR}_{e,1} \\ \vdots \\ \mathrm{SNR}_{e,M_e} \end{bmatrix}} \tag{29}$$

Hence, if the assumed model for the speech and noise covariance matrix $\mathbf{R}_x$ and $\mathbf{R}_n$ in (4) and (5) holds, the optimal weights for the mSNR approach are real-valued and equal to the normalized input SNRs in the external microphones. This means that in the linear combination for computing $\tilde{\mathbf{h}}^{\mathrm{mSNR}}$ in (11), SC RTF vector estimates corresponding to external microphones with a higher input SNR are assigned a larger weight than those with a low input SNR, which seems very intuitive.

When analyzing the bias of the RTF vector estimates $\tilde{\mathbf{h}}^{\mathrm{mSNR}}$ in (11) using the model weights in (29), the first M_a entries (corresponding to the LMA) are obviously still unbiased while the bias of the last M_e entries (corresponding to the external microphones) is equal to, using (10),

$$\boxed{\tilde{H}_{e,m_e}^{\mathrm{mSNR}} = \left(1 + \frac{1}{\sum_{m_e'=1}^{M_e} \mathrm{SNR}_{e,m_e'}}\right) H_{e,m_e}} \tag{30}$$

In contrast to the individual RTF vector estimates in (10), it can be observed that the bias of the optimal linear combination is the same for all entries corresponding to the external microphones. In addition, the bias in (30) for the mSNR approach (exploiting all external microphones) is always smaller than the bias in (10) for the individual RTF vector estimates.

5. EVALUATION

The following evaluation aims at demonstrating accordance and deviations between the theoretically found weights of the mSNR approach (referred to as "model") in (29) and the solution using a GEVD of the respective covariance matrices (referred to as "GEVD") in (15). It should be noted that in theory, when all assumptions are fulfilled, the GEVD- and model-based weights are identical. However, the assumptions used in the bias analysis in Section 4, such as the statistical independence of speech and noise in (3), the rank-1 model in (4) and the spatial coherence assumption in (5), do not perfectly hold in practice, where estimates of the covariance matrices are used.

5.1. Setup and Implementation

The evaluation was performed using real-world signals that were recorded in a laboratory room with a size of about $(7 \times 6 \times 2.7)\,\mathrm{m}^3$ and a reverberation time of approximately 400 ms. For the LMA, binaural hearing aids with 2 microphones per side ($M_a = 4$) with an inter-microphone distance of about 7 mm were mounted on a KEMAR dummy head. In addition, $M_e = 2$ external microphones were placed to the left and

the right front of the dummy head, both at a distance of about 2.3 m from the dummy head, as depicted in Fig. 1. The desired speech source was a male German speaker, walking from the left to the right side of the dummy head, i.e., from the first external microphone (E1) to the second external microphone (E2). Pseudo-diffuse background noise was generated using four loudspeakers facing the corners of the laboratory, playing back different multi-talker recordings. The speech and noise components were recorded separately and were subsequently mixed. The broadband input SNR in the LMA microphones varied from about 0 to 6 dB due to the movement of the speech source, while the input SNR in the external microphone signals varied as shown in the upper panel of Fig. 2. The sampling rate for all signals was 16 kHz.

All processing was performed in the STFT-domain with a frame length of 32 ms, a square-root-Hann window for analysis and synthesis and an overlap of 50%. To allow for a good visualization of the theoretical findings of the bias analysis, oracle estimates of the covariance matrices $\mathbf{R}_y$, $\mathbf{R}_x$ and $\mathbf{R}_n$ were computed on the separate signal components $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{n}$. To account for the dynamic acoustic scenario, the covariance matrices were updated by recursive smoothing with time constants of 250 ms for $\mathbf{R}_y$ and $\mathbf{R}_x$ and 1 s for $\mathbf{R}_n$, respectively, where $\mathbf{R}_y$ and $\mathbf{R}_x$ were only updated if speech was active (determined via an oracle broadband voice activity detection).

As a performance measure, we considered the broadband SNR improvement ($\Delta\mathrm{SNR} = \mathrm{SNR}_{\mathrm{out}} - \mathrm{SNR}_{\mathrm{in}}$) of the RTF-steered MVDR beamformer using

1. the SC RTF vector estimates $\tilde{\mathbf{h}}_{m_e}^{\mathrm{SC}}$ in (9) from the individual external microphones (SC-1 and SC-2), computed using the estimated covariance matrix $\hat{\mathbf{R}}_y$.

2. the mSNR combination $\tilde{\mathbf{h}}^{\mathrm{mSNR}}$ in (11) using the (complex-valued) GEVD-based weights $\boldsymbol{\alpha}^{\mathrm{GEVD}}$ in (15), computed using the estimated covariance matrices $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{R}}_n$.

3. the mSNR combination $\tilde{\mathbf{h}}^{\mathrm{mSNR}}$ in (11) using the (real-valued) model-based weights in (29), computed using the estimated covariance matrices $\hat{\mathbf{R}}_x$ and $\hat{\mathbf{R}}_n$.

The SNR improvement was averaged over the left and right side of the hearing aids to reduce effects caused by the moving speaker. Input and output SNR were computed in the time-domain using the shadow filter approach, only when the speaker was active.

5.2. Results

The upper panel of Fig. 2 shows the input SNR in the two external microphones (averaged over frequency) plotted over time. As expected from the movement of the source from E1 to E2, the input SNR decreases in E1 while increasing in E2 over time. The GEVD- and model-based weights to combine the RTF vector estimated obtained are shown in the lower panel of Fig. 2. For the model-based weights $\boldsymbol{\alpha}^{\mathrm{model}}$ it can be observed that the weight of the RTF vector estimate obtained from E1 $\alpha_1^{\mathrm{model}}$ (thick blue line) initially is almost equal to 1, meaning that this estimate dominates in the linear combination. Over time, with decreasing input SNR in E1, the weight $\alpha_1^{\mathrm{model}}$ also decreases, whereas the weight of the RTF vector estimate obtained from E2 $\alpha_2^{\mathrm{model}}$ (thick red line) increases. For the GEVD-based weights α_1^{GEVD} and α_2^{GEVD} a similar behavior can be observed (thin lines), showing a good accordance between the theoretical model-based and the practical GEVD-based weights. Nevertheless, since the GEVD-based weights are complex-valued and we only investigate the real part, there exists a larger deviation between practice and theory than can be observed from this figure.

Fig. 3 shows the SNR improvement (averaged over time) of the considered RTF-steered MVDR beamformers. It can be seen that
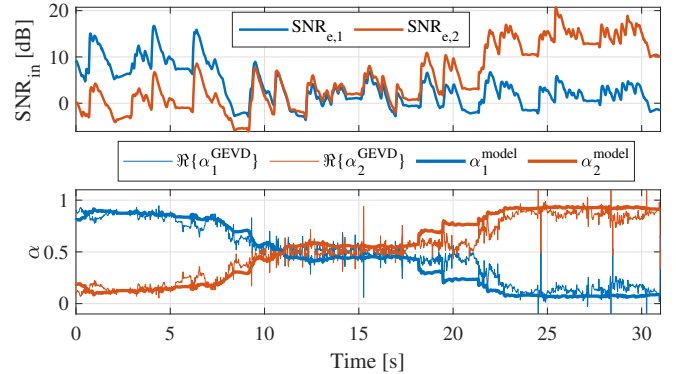

Fig. 2. Input SNR in both external microphones (upper panel) and weights of RTF vector estimates obtained by the GEVD-based and the model-based mSNR combination (lower panel).
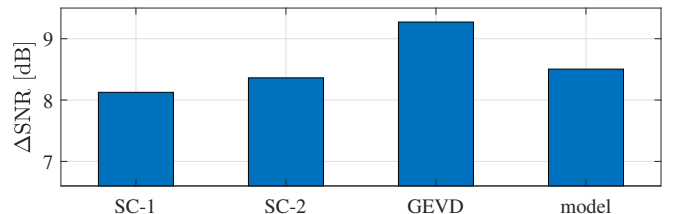

Fig. 3. Time-average broadband output SNR for the single SC RTF vector estimates (SC-1 and SC-2) and combinations of estimates using the GEVD- and the model-based weights used in an MVDR beamformer.

both SC RTF vector estimates (SC-1 and SC-2) perform very similarly, yielding an SNR improvement of about 8 dB. The model-based mSNR combination leads to a slightly better SNR improvement than the individual RTF vector estimates, SC-1 and SC-2. The GEVD-based mSNR combination clearly yields the best performance, leading to an SNR improvement of more than 9 dB. Hence, despite (the real parts of) the GEVD- and the model-based weights being rather similar, the complex-valued GEVD-based weights lead to better results than the real-valued model-based weights. The deviations of the theoretically identical GEVD- and model-based weights can be explained by the model assumptions, such as the rank-1 model in (4) and the spatial coherence assumption about the noise field in (5), not perfectly holding in practice. This leads to generally complex-valued GEVD-based weights, which can - to some extent - compensate for these deviations and estimation errors.

6. CONCLUSIONS

In this paper, we performed a theoretical analysis of the mSNR approach, which linearly combines (biased) RTF vector estimates obtained using the SC method. We derived an analytic expression for the optimal weights of the linear combination (usually obtained by a complex-valued GEVD), which is real-valued and only depends on the input SNRs in the external microphones. Furthermore, we showed that the biases of the optimal combination are smaller than the bias of every individual RTF vector estimate. In a simulation with real-world recordings, we showed that the theoretically equivalent GEVD- and model-based weights strongly resemble each other but some deviations can be observed in practice, leading to a better performance of the complex-valued GEVD-based weights in terms of SNR improvement. The observed differences between the model- and GEVD-based weights can be explained by the model assumptions made in the analysis not perfectly holding in practice.

7. 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*, chapter 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, no. 2, pp. 18–30, Mar. 2015.

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

[4] 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.

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

[6] 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.

[7] D. Yee, H. Kamkar-Parsi, R. Martin, and H. Puder, "A noise reduction post-filter for binaurally-linked single-microphone hearing aids utilizing a nearby external microphone," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 5–18, Jul. 2017.

[8] 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.

[9] 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, Oct. 2018, pp. 106–110.

[10] R. M. Corey and A. C. Singer, "Adaptive binaural filtering for a multiple-talker listening system using remote and on-ear microphones," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, Oct. 2021, pp. 1–5.

[11] W. Middelberg and S. Doclo, "Comparison of Generalized Sidelobe Canceller Structures Incorporating External Microphones for Joint Noise and Interferer Reduction," in *Proc. ITG Conference on Speech Communication*, Kiel, Germany, Oct. 2021, pp. 104–108.

[12] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, Nov. 2011, pp. 1–6.

[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, Sep. 2018, pp. 146–150.

[14] N. Gößling and S. Doclo, "RTF-steered binaural MVDR beamforming incorporating an external microphone for dynamic acoustic scenarios," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, May 2019, pp. 416–420.

[15] 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, Oct. 2019, pp. 368–372.