RTF-STEERED BINAURAL MVDR BEAMFORMING INCORPORATING MULTIPLE EXTERNAL MICROPHONES

Nico Gößling, Wiebke Middelberg, Simon Doclo

University of Oldenburg, Department of Medical Physics and Acoustics
and Cluster of Excellence Hearing4all, Oldenburg, Germany
`nico.goessling@uni-oldenburg.de`

ABSTRACT

The binaural minimum-variance distortionless-response (BMVDR) beamformer is a well-known noise reduction algorithm that can be steered using the relative transfer function (RTF) vector of the desired speech source. Exploiting the availability of an external microphone that is spatially separated from the head-mounted microphones, an efficient method has been recently proposed to estimate the RTF vector in a diffuse noise field. When multiple external microphones are available, different RTF vector estimates can be obtained by using this method for each external microphone. In this paper, we propose several procedures to combine these RTF vector estimates, either by selecting the estimate corresponding to the highest input SNR, by averaging the estimates or by combining the estimates in order to maximize the output SNR of the BMVDR beamformer. Experimental results for a moving speaker and diffuse noise in a reverberant environment show that the output SNR-maximizing combination yields the largest binaural SNR improvement and also outperforms the state-of-the art covariance whitening method.

Index Terms— binaural noise reduction, relative transfer function, external microphones, hearing devices

1. INTRODUCTION

Noise reduction algorithms for head-mounted assistive listening devices (e.g., hearing aids, earbuds, headsets) are crucial to improve speech intelligibility and speech quality in noisy environments. Binaural noise reduction algorithms, which exploit the information captured by all microphones on both sides of the head [1, 2], do not only allow to reduce unwanted sound sources but also allow to preserve the listener's spatial impression of the acoustic scene. As a well-known example, the binaural minimum-variance distortionless-response (BMVDR) beamformer is able to preserve the binaural cues (i.e. the interaural time and level differences) of a desired speech source [1–3]. For a moving speech source in a reverberant environment, the BMVDR can be steered using the relative transfer functions (RTFs) [4], which relate the acoustic transfer functions between the desired speech source and all microphones to the so-called reference microphones.

To improve the performance of (binaural) algorithms in terms of noise reduction and source localization accuracy, it has been proposed to use an external microphone in conjunction with the head-mounted microphones [5–12]. For a diffuse noise field, an efficient RTF vector estimation method has been proposed in [11],

which exploits the spatial coherence (SC) properties of the noise field. More specifically, the SC method assumes that the noise component in the external microphone signal is uncorrelated with the noise components in the head-mounted microphone signals.

In this paper, we consider the more general scenario with multiple external microphones. Using the SC method, each external microphone yields a (different) RTF vector estimate, such that the question arises how to combine these RTF vector estimates. In the first procedure, we propose to select the RTF vector estimate corresponding to the external microphone with the highest narrowband signal-to-noise ratio (SNR). In the second procedure, we propose to simply average the different RTF vector estimates. In the third procedure, we propose to linearly combine the different RTF vector estimates such that the narrowband output SNR of the BMVDR is maximized. Experimental results of an on-line implementation of the BMVDR using recorded signals of a moving speaker and diffuse noise in a reverberant environment are provided. The results show that the output SNR-maximizing combination of the SC-based RTF vector estimates leads to the largest binaural SNR improvement compared to the other procedures and the state-of-the-art covariance whitening method [13, 14].

2. CONFIGURATION AND NOTATION

Consider the binaural hearing device configuration depicted in Figure 1, consisting of a left and a right hearing device (each equipped with M_{D} microphones), and M_{E} external microphones that are spatially separated from the head-mounted microphones, i.e. $M = 2M_{\mathrm{D}} + M_{\mathrm{E}}$ microphones in total. In the frequency-domain, the m-th microphone signal of the left device can be written as

$$y_{\mathrm{L},m}(\omega) = x_{\mathrm{L},m}(\omega) + n_{\mathrm{L},m}(\omega), \quad m \in \{1, \ldots, M_{\mathrm{D}}\}, \quad (1)$$

with $x_{\mathrm{L},m}(\omega)$ the desired speech component and $n_{\mathrm{L},m}(\omega)$ the noise component. For the sake of conciseness, the frequency ω will be omitted in the remainder of the paper. The m-th microphone signal of the right device $y_{\mathrm{R},m}$ and the i-th external microphone signal $y_{\mathrm{E},i}$ are defined similarly as in (1). The M-dimensional microphone signal vector, containing *all* microphone signals, is defined as

$$\boldsymbol{y} = [y_{\mathrm{L},1}, \ldots, y_{\mathrm{L},M_{\mathrm{D}}}, y_{\mathrm{R},1}, \ldots, y_{\mathrm{R},M_{\mathrm{D}}}, y_{\mathrm{E},1}, \ldots, y_{\mathrm{E},M_{\mathrm{E}}}]^T, \quad (2)$$

with $(\cdot)^T$ denoting the transpose operator. Using (1), the vector $\boldsymbol{y}$ can be written as

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}, \quad (3)$$

where the speech vector $\boldsymbol{x}$ and the noise vector $\boldsymbol{n}$ are defined similarly as in (2). Without loss of generality, the first microphone

on each device is chosen as the reference microphone, i.e.

$$y_{\mathrm{L}} = y_{\mathrm{L},1} = \boldsymbol{e}_{\mathrm{L}}^{T} \boldsymbol{y}, \quad y_{\mathrm{R}} = y_{\mathrm{R},1} = \boldsymbol{e}_{\mathrm{R}}^{T} \boldsymbol{y}, \qquad (4)$$

where $\boldsymbol{e}_{\mathrm{L}}$ and $\boldsymbol{e}_{\mathrm{R}}$ denote selection vectors consisting of zeros and one element equal to 1. Assuming a single desired speech source, the vector $\boldsymbol{x}$ can be written as

$$\boldsymbol{x} = \boldsymbol{a}_{\mathrm{L}} x_{\mathrm{L}} = \boldsymbol{a}_{\mathrm{R}} x_{\mathrm{R}}, \qquad (5)$$

where $\boldsymbol{a}_{\mathrm{L}}$ and $\boldsymbol{a}_{\mathrm{R}}$ denote the M-dimensional RTF vectors of the desired speech source with respect to the reference microphones on the left and the right device, respectively. It should be noted that one of the elements of the RTF vectors (corresponding to the reference microphone) is equal to 1 and that the RTF vectors are related as $\boldsymbol{a}_{\mathrm{R}} = \boldsymbol{a}_{\mathrm{L}}/\boldsymbol{e}_{\mathrm{R}}^{T}\boldsymbol{a}_{\mathrm{L}}$. The noisy input covariance matrix $\boldsymbol{R}_{\mathrm{y}}$, the speech covariance matrix $\boldsymbol{R}_{\mathrm{x}}$ and the noise covariance matrix $\boldsymbol{R}_{\mathrm{n}}$ are defined as

$$\boldsymbol{R}_{\mathrm{y}} = \mathcal{E}\{\boldsymbol{y}\boldsymbol{y}^{H}\}, \ \boldsymbol{R}_{\mathrm{x}} = \mathcal{E}\{\boldsymbol{x}\boldsymbol{x}^{H}\}, \ \boldsymbol{R}_{\mathrm{n}} = \mathcal{E}\{\boldsymbol{n}\boldsymbol{n}^{H}\}, \quad (6)$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator and $(\cdot)^{H}$ denotes the conjugate transpose. Assuming statistical independence between the desired speech component and the noise component, the noisy input covariance matrix is equal to $\boldsymbol{R}_{\mathrm{y}} = \boldsymbol{R}_{\mathrm{x}} + \boldsymbol{R}_{\mathrm{n}}$.

The output signals of the left and right devices are calculated by filtering and summing *all* microphone signals, i.e. the head-mounted microphone signals as well as the external microphone signals, using the complex-valued filter vectors $\boldsymbol{w}_{\mathrm{L}}$ and $\boldsymbol{w}_{\mathrm{R}}$ (see Figure 1), i.e.

$$z_{\mathrm{L}} = \boldsymbol{w}_{\mathrm{L}}^{H} \boldsymbol{y}, \quad z_{\mathrm{R}} = \boldsymbol{w}_{\mathrm{R}}^{H} \boldsymbol{y}. \qquad (7)$$

The input SNR of the m-th microphone signal is given by the ratio of the input power spectral density (PSD) of the desired speech component and the input PSD of the noise component, i.e.

$$\mathrm{SNR}_{m}^{\mathrm{in}} = \frac{\boldsymbol{e}_{m}^{T} \boldsymbol{R}_{\mathrm{x}} \boldsymbol{e}_{m}}{\boldsymbol{e}_{m}^{T} \boldsymbol{R}_{\mathrm{n}} \boldsymbol{e}_{m}} = \frac{\boldsymbol{e}_{m}^{T} \boldsymbol{R}_{\mathrm{y}} \boldsymbol{e}_{m}}{\boldsymbol{e}_{m}^{T} \boldsymbol{R}_{\mathrm{n}} \boldsymbol{e}_{m}} - 1, \qquad (8)$$

with $\boldsymbol{e}_{m}$ an M-dimensional vector selecting the element corresponding to the m-th microphone. Similarly, the output SNR of the left and the right output signals is given by the ratio of the output PSD of the desired speech component and the output PSD of the noise component, i.e.

$$\mathrm{SNR}_{\mathrm{L}}^{\mathrm{out}} = \frac{\boldsymbol{w}_{\mathrm{L}}^{H} \boldsymbol{R}_{\mathrm{x}} \boldsymbol{w}_{\mathrm{L}}}{\boldsymbol{w}_{\mathrm{L}}^{H} \boldsymbol{R}_{\mathrm{n}} \boldsymbol{w}_{\mathrm{L}}}, \quad \mathrm{SNR}_{\mathrm{R}}^{\mathrm{out}} = \frac{\boldsymbol{w}_{\mathrm{R}}^{H} \boldsymbol{R}_{\mathrm{x}} \boldsymbol{w}_{\mathrm{R}}}{\boldsymbol{w}_{\mathrm{R}}^{H} \boldsymbol{R}_{\mathrm{n}} \boldsymbol{w}_{\mathrm{R}}}. \qquad (9)$$

3. BINAURAL MVDR BEAMFORMER

The BMVDR [2, 15] aims at minimizing the output noise PSD while preserving the desired speech component in the reference microphone signals (x_{L} and x_{R}), hence preserving the binaural cues of the desired speech source. The optimization problem for the left filter vector $\boldsymbol{w}_{\mathrm{L}}$ is given by

$$\min_{\boldsymbol{w}_{\mathrm{L}}} \boldsymbol{w}_{\mathrm{L}}^{H} \boldsymbol{R}_{\mathrm{n}} \boldsymbol{w}_{\mathrm{L}} \quad \text{subject to} \quad \boldsymbol{w}_{\mathrm{L}}^{H} \boldsymbol{a}_{\mathrm{L}} = 1. \qquad (10)$$

The optimization problem for the right filter vector $\boldsymbol{w}_{\mathrm{R}}$ is defined similarly. The filter vectors solving the optimization problems are equal to [1, 2, 15]

$$\boldsymbol{w}_{\mathrm{L}} = \frac{\boldsymbol{R}_{\mathrm{n}}^{-1} \boldsymbol{a}_{\mathrm{L}}}{\boldsymbol{a}_{\mathrm{L}}^{H} \boldsymbol{R}_{\mathrm{n}}^{-1} \boldsymbol{a}_{\mathrm{L}}}, \quad \boldsymbol{w}_{\mathrm{R}} = \frac{\boldsymbol{R}_{\mathrm{n}}^{-1} \boldsymbol{a}_{\mathrm{R}}}{\boldsymbol{a}_{\mathrm{R}}^{H} \boldsymbol{R}_{\mathrm{n}}^{-1} \boldsymbol{a}_{\mathrm{R}}}. \qquad (11)$$
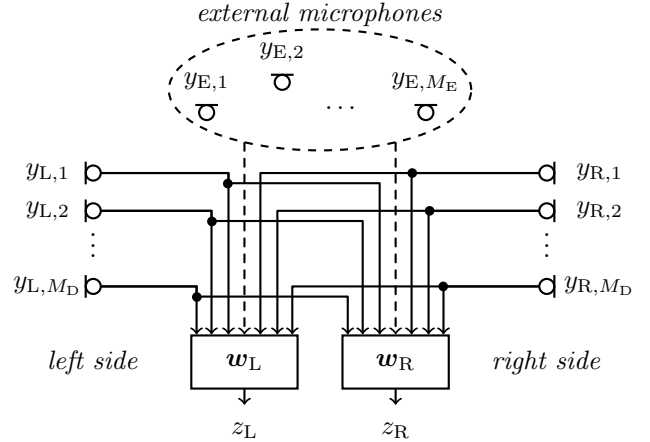

Figure 1: Binaural hearing device configuration incorporating multiple external microphones.

Hence, estimates of the noise covariance matrix $\boldsymbol{R}_{\mathrm{n}}$ and the RTF vectors $\boldsymbol{a}_{\mathrm{L}}$ and $\boldsymbol{a}_{\mathrm{R}}$ are required to compute the BMVDR filter vectors in practice. Typically, the noise covariance matrix $\boldsymbol{R}_{\mathrm{n}}$ is recursively estimated from the microphone signals during speech pauses, e.g., based on a voice activity detector or speech presence probability [16].

The following sections describe different methods to estimate the RTF vectors $\boldsymbol{a}_{\mathrm{L}}$ and $\boldsymbol{a}_{\mathrm{R}}$. Section 4 describes the covariance whitening method, which is a state-of-the-art RTF vector estimation method for a general noise field. In Section 5 we propose RTF vector estimation methods that assume that the noise component in each external microphone signal is uncorrelated with the noise components in all other microphone signals.

4. COVARIANCE WHITENING METHOD

The covariance whitening (CW) method [13, 14] is based on the generalized eigenvalue decomposition of the noisy input covariance matrix $\boldsymbol{R}_{\mathrm{y}}$ and the noise covariance matrix $\boldsymbol{R}_{\mathrm{n}}$. Using the Cholesky decomposition of the noise covariance matrix, i.e.

$$\boldsymbol{R}_{\mathrm{n}} = \boldsymbol{R}_{\mathrm{n}}^{H/2} \boldsymbol{R}_{\mathrm{n}}^{1/2}, \qquad (12)$$

the pre-whitened noisy input covariance matrix is defined as

$$\boldsymbol{R}_{\mathrm{y}}^{\mathrm{w}} = \boldsymbol{R}_{\mathrm{n}}^{-H/2} \boldsymbol{R}_{\mathrm{y}} \boldsymbol{R}_{\mathrm{n}}^{-1/2}. \qquad (13)$$

Using (12) and (13), the left RTF vector can be estimated as [14]

$$\boldsymbol{a}_{\mathrm{L}}^{\mathrm{CW}} = \frac{\boldsymbol{R}_{\mathrm{n}}^{1/2} \boldsymbol{p}}{\boldsymbol{e}_{\mathrm{L}}^{T} \boldsymbol{R}_{\mathrm{n}}^{1/2} \boldsymbol{p}}, \qquad (14)$$

with $\boldsymbol{p} = \mathcal{P}\{\boldsymbol{R}_{\mathrm{y}}^{\mathrm{w}}\}$ the principal eigenvector (corresponding to the largest eigenvalue) of the pre-whitened noisy input covariance matrix $\boldsymbol{R}_{\mathrm{y}}^{\mathrm{w}}$. Due to the Cholesky decomposition and the $M \times M$-dimensional eigenvalue decomposition (EVD) the CW method typically has a rather large computational complexity, especially for large M.

5. SPATIAL COHERENCE METHOD

In this section, we propose RTF vector estimation methods that assume that the noise component in each external microphone signal

is uncorrelated with the noise components in all other microphone signals. This can, e.g., be assumed for a diffuse noise field when the external microphones are spatially separated from each other and from the head-mounted microphones. In Section 5.1, we review the SC method as presented in [11] for one external microphone. In Section 5.2, we propose three different procedures to linearly combine the RTF vector estimates obtained by using the SC method for each external microphone.

5.1. SC method per external microphone

If the noise component in the i-th external microphone signal is uncorrelated with the noise components in all other microphone signals, it has been shown in [11, 12] that the left RTF vector can be efficiently estimated from the noisy input covariance matrix $\boldsymbol{R}_\mathrm{y}$ using the SC method as

$$\boldsymbol{a}_\mathrm{L}^{\mathrm{SC}-i} = \frac{\boldsymbol{R}_\mathrm{y}\boldsymbol{e}_{\mathrm{E},i}}{\boldsymbol{e}_\mathrm{L}^T\boldsymbol{R}_\mathrm{y}\boldsymbol{e}_{\mathrm{E},i}}, \quad i \in \{1,\ldots,M_\mathrm{E}\}, \qquad (15)$$

with $\boldsymbol{e}_{\mathrm{E},i}$ an M-dimensional vector, selecting the element corresponding to the i-th external microphone. The estimator in (15) yields an unbiased RTF vector estimate, except for a biased estimate of the RTF corresponding to the i-th external microphone. However, in [12] it has been shown that this bias is real-valued (hence not affecting the phase of the RTF vector estimate), depends on the input SNR in the i-th external microphone and typically can be neglected in practice.

5.2. Combination of SC-based RTF vector estimates

Since in practice an estimate of the noisy input covariance matrix $\hat{\boldsymbol{R}}_\mathrm{y}$ is used in (15), typically M_E different SC-based RTF vector estimates are obtained, such that the question arises how to use these estimates. In this paper, we propose to linearly combine the different RTF vector estimates (per frequency) and to use the resulting RTF vector in the BMVDR. The (normalized) combined RTF vector estimate is given by

$$\boxed{\boldsymbol{a}_\mathrm{L}^{\mathrm{SC}-\mathrm{C}} = \frac{\boldsymbol{A}_\mathrm{L}^{\mathrm{SC}}\boldsymbol{c}}{\boldsymbol{e}_\mathrm{L}^T\boldsymbol{A}_\mathrm{L}^{\mathrm{SC}}\boldsymbol{c}}} \qquad (16)$$

with $\boldsymbol{A}_\mathrm{L}^{\mathrm{SC}}$ an $M \times M_\mathrm{E}$-dimensional matrix, containing the M_E SC-based RTF vector estimates, i.e.

$$\boldsymbol{A}_\mathrm{L}^{\mathrm{SC}} = \left[\boldsymbol{a}_\mathrm{L}^{\mathrm{SC}-1},\ldots,\boldsymbol{a}_\mathrm{L}^{\mathrm{SC}-M_\mathrm{E}}\right], \qquad (17)$$

and $\boldsymbol{c}$ an M_E-dimensional (complex-valued) combination vector. Please note that the combination closest to the true RTF vector $\boldsymbol{a}_\mathrm{L}$ could be obtained by orthogonally projecting $\boldsymbol{a}_\mathrm{L}$ on the column space of $\boldsymbol{A}_\mathrm{L}^{\mathrm{SC}}$, which is obviously not possible in practice. In the following, we hence propose three different procedures to determine the combination vector $\boldsymbol{c}$ in practice.

The first procedure, denoted as **iSNR**, is to select the RTF vector estimate (per frequency) corresponding to the external microphone with the highest narrowband input SNR, similarly to [17]. Due to (8), this only requires an estimate of $\boldsymbol{R}_\mathrm{y}$ (and not $\boldsymbol{R}_\mathrm{x}$), i.e.

$$\boldsymbol{c}^{\mathrm{iSNR}} = \boldsymbol{e}_{\mathrm{E},\hat{i}}, \quad \hat{i} = \arg\max_i \frac{\boldsymbol{e}_{\mathrm{E},i}^T\boldsymbol{R}_\mathrm{y}\boldsymbol{e}_{\mathrm{E},i}}{\boldsymbol{e}_{\mathrm{E},i}^T\boldsymbol{R}_\mathrm{n}\boldsymbol{e}_{\mathrm{E},i}}. \qquad (18)$$

Especially for a dynamic acoustic scenario with a moving speaker, the iSNR-based selection procedure is expected to outperform the SC method only using on one external microphone.

Assuming a uniform distribution of the estimation errors for the SC-based RTF vector estimates, in the second procedure, denoted as **AV**, we propose to simply average the estimates, i.e.

$$\boldsymbol{c}^{\mathrm{AV}} = \left[\frac{1}{M_\mathrm{E}},\ldots,\frac{1}{M_\mathrm{E}}\right]^T. \qquad (19)$$

Intuitively, this procedure is sub-optimal, especially when the estimation errors are very different.

As a more sophisticated procedure, denoted as **mSNR**, we propose to combine the SC-based RTF vector estimates (per frequency) such that the narrowband output SNR of the BMVDR is maximized. Using (16) in (11), the left output SNR in (9) can be written as the generalized Rayleigh quotient

$$\mathrm{SNR}_{\mathrm{BMVDR,L}}^{\mathrm{out}} = \frac{\boldsymbol{c}^H\boldsymbol{\Lambda}_1\boldsymbol{c}}{\boldsymbol{c}^H\boldsymbol{\Lambda}_2\boldsymbol{c}} - 1, \qquad (20)$$

with

$$\boldsymbol{\Lambda}_1 = (\boldsymbol{A}_\mathrm{L}^{\mathrm{SC}})^H\boldsymbol{R}_\mathrm{n}^{-1}\boldsymbol{R}_\mathrm{y}\boldsymbol{R}_\mathrm{n}^{-1}\boldsymbol{A}_\mathrm{L}^{\mathrm{SC}}, \qquad (21)$$

$$\boldsymbol{\Lambda}_2 = (\boldsymbol{A}_\mathrm{L}^{\mathrm{SC}})^H\boldsymbol{R}_\mathrm{n}^{-1}\boldsymbol{A}_\mathrm{L}^{\mathrm{SC}}. \qquad (22)$$

Aiming at maximizing the output SNR of the BMVDR, the SNR-maximizing combination vector $\boldsymbol{c}^{\mathrm{mSNR}}$ is equal to the principal eigenvector of the $M_\mathrm{E} \times M_\mathrm{E}$-dimensional matrix $\boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Lambda}_1$, i.e.

$$\boxed{\boldsymbol{c}^{\mathrm{mSNR}} = \arg\max_{\boldsymbol{c}} \mathrm{SNR}_{\mathrm{BMVDR,L}}^{\mathrm{out}} = \mathcal{P}\{\boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Lambda}_1\}} \qquad (23)$$

which hence also only requires an estimate of $\boldsymbol{R}_\mathrm{y}$ (and not $\boldsymbol{R}_\mathrm{x}$). Although constructing the matrices $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ comes with some computational complexity, the computational complexity of the $M_\mathrm{E} \times M_\mathrm{E}$-dimensional EVD is always smaller than the $M \times M$-dimensional EVD required for the CW method (cf. Section 4).

6. EXPERIMENTAL RESULTS

For a dynamic acoustic scenario with a moving speaker in a reverberant room, in this section we compare the performance of the BMVDR using the different RTF vector estimation methods described in Sections 4 and 5 for a binaural hearing device incorporating three external microphones.

6.1. Recording setup and implementation

All signals were recorded in a laboratory where the reverberation time can be varied using absorber panels mounted on the walls and the ceiling. The room dimensions are about $(7 \times 6 \times 2.7)\,\mathrm{m}^3$ and the reverberation time was set to approximately $400\,\mathrm{ms}$. A KEMAR dummy head was placed approximately in the center of the room with two behind-the-ear (BTE) hearing devices mounted to the ears. Two microphones per hearing device, i.e. $M_\mathrm{D} = 2$, with an inter-microphone distance of about $7\,\mathrm{mm}$ were used. In addition, $M_\mathrm{E} = 3$ external microphones were placed in front of the dummy head as depicted in Figure 2. Hence, in total $M = 7$ microphones were used for the BMVDR. The desired speech source was a male speaker, walking from the first external microphone (E1) to the third external microphone (E3) while speaking ten German sentences with pauses of about half a second between the sentences. Pseudo-diffuse
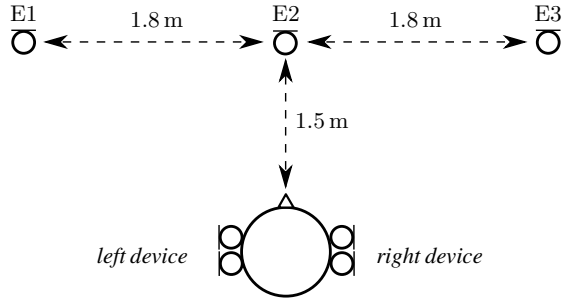
Figure 2: Experimental setup with BTE hearing devices mounted on a dummy head and three external microphones.

background noise was generated using four loudspeakers facing the corners of the laboratory, playing back different multi-talker recordings. The desired speech source and the background noise were recorded separately and mixed afterwards. Due to the moving speaker, the input SNR in the head-mounted reference microphone signals varied between approximately 0 and 6 dB, while the input SNR in the external microphone signals varied approximately between 0 and 11 dB. All signals were recorded synchronously, hence neglecting synchronization and latency aspects.

All signals were sampled at a sampling rate of 16 kHz and processed in the short-time Fourier transform domain using a 32 ms square-root Hann window with 50% overlap. To distinguish between speech-plus-noise and noise-only time-frequency bins, the estimated speech presence probabilities [16] in the three (noisy) external microphone signals were averaged and thresholded. The noisy input covariance matrix $\boldsymbol{R}_\text{y}$ and the noise covariance matrix $\boldsymbol{R}_\text{n}$ were then recursively estimated during detected speech-plus-noise and noise-only bins, respectively, using time constants of 250 ms ($\boldsymbol{R}_\text{y}$) and 1.5 s ($\boldsymbol{R}_\text{n}$).

As performance measure, we used the binaural SNR improvement (ΔBSNR), which is defined similarly as in (8) and (9) as

$$\Delta\text{BSNR} = 10\log_{10}\left(\frac{\boldsymbol{w}_\text{L}^H \boldsymbol{R}_\text{x} \boldsymbol{w}_\text{L} + \boldsymbol{w}_\text{R}^H \boldsymbol{R}_\text{x} \boldsymbol{w}_\text{R}}{\boldsymbol{w}_\text{L}^H \boldsymbol{R}_\text{n} \boldsymbol{w}_\text{L} + \boldsymbol{w}_\text{R}^H \boldsymbol{R}_\text{n} \boldsymbol{w}_\text{R}}\right) \quad (24)$$
$$- 10\log_{10}\left(\frac{\boldsymbol{e}_\text{L}^H \boldsymbol{R}_\text{x} \boldsymbol{e}_\text{L} + \boldsymbol{e}_\text{R}^H \boldsymbol{R}_\text{x} \boldsymbol{e}_\text{R}}{\boldsymbol{e}_\text{L}^H \boldsymbol{R}_\text{n} \boldsymbol{e}_\text{L} + \boldsymbol{e}_\text{R}^H \boldsymbol{R}_\text{n} \boldsymbol{e}_\text{R}}\right).$$

The binaural SNR improvement was computed in the time-domain using the shadow filter approach.

Seven different RTF vector estimates were considered for the BMVDR in (11):

- The state-of-the-art **CW** estimate in (14)

- The SC estimate in (15) using each external microphone separately, i.e. **SC-1**, **SC-2** and **SC-3**

- The proposed SC-C method using the combination vectors in (18), (19) and (23), i.e. **iSNR**, **AV** and **mSNR**

6.2. Results

Figure 3 depicts the ΔBSNR (averaged over time and frequency) for all considered RTF vector estimates. The CW method as a state-of-the-art benchmark yields an average ΔBSNR of 10.4 dB. The SC method using one external microphone, i.e. SC-1, SC-2 and SC-3, yields an average ΔBSNR of about 9 dB and hence could not reach the performance of the CW method. The input SNR-based
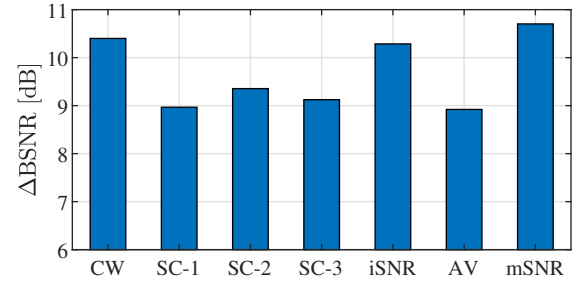

Figure 3: Binaural SNR improvement for all considered RTF vector estimation methods, averaged over time and frequency.
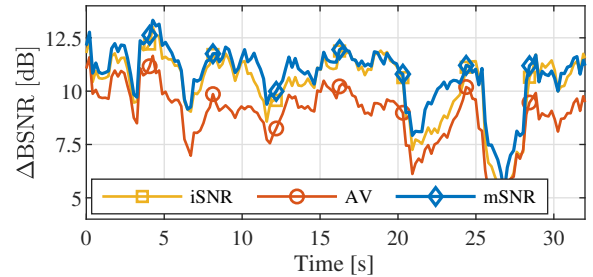

Figure 4: Binaural SNR improvement over time for the iSNR, AV and mSNR combination procedures, averaged over frequency.

combination (iSNR) yields an average ΔBSNR of 10.3 dB, which is similar to the CW method. The averaging combination (AV) yields an average ΔBSNR of only 8.9 dB, which is even worse than the SC method per external microphone. This can probably be explained by the rather different RTF vector estimation errors for the three external microphones. The SNR-maximizing combination (mSNR) yields an average ΔBSNR of 10.7 dB, hence outperforming all other combination procedures and RTF vector estimation methods. Comparing the computational complexity of the best three methods, the CW method has the largest complexity due to the 7-dimensional EVD, whereas the mSNR combination only requires a 3-dimensional EVD and the iSNR combination does not even require an EVD. Nevertheless, the mSNR combination enables to improve the ΔBSNR by about 0.5 dB compared to the iSNR combination. Figure 4 depicts the ΔBSNR over time (averaged over frequency) for the SC-C method using the proposed combination vectors in more detail. It can be observed that the mSNR combination outperforms the iSNR and AV combination for almost all time instances. The sound files of the input and output signals are available at [18].

7. CONCLUSIONS

In this paper, we proposed to use the SC-based RTF vector estimation method for a scenario where multiple external microphones are incorporated into the BMVDR processing of a binaural hearing device. Each external microphone was used to obtain an SC-based RTF vector estimate. We proposed to linearly combine the different RTF vector estimates using an input SNR-based selection, simple averaging and a combination that maximizes the narrowband output SNR of the BMVDR. Experimental evaluation in a dynamic scenario with a moving speaker in a reverberant environment showed that the SNR-maximizing combination yields the largest binaural SNR improvement and also outperforms the state-of-the art covariance whitening method.

8. REFERENCES

[1] 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.

[2] S. Doclo, S. Gannot, D. Marquardt, and E. Hadad, "Binaural speech processing with application to hearing devices," in *Audio Source Separation and Speech Enhancement*. Wiley, 2018, ch. 18, pp. 413–442.

[3] B. Cornelis, S. Doclo, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural multi-microphone noise reduction techniques," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

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

[5] 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, p. 14 pages, Jan. 2009.

[6] 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 Transactions on Audio, Speech and Language Processing*, vol. 24, no. 5, pp. 952–966, May 2016.

[7] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611–623, Mar. 2017.

[8] 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 Transactions on Audio Speech and Language Processing*, vol. 26, no. 1, pp. 5–18, Jan. 2018.

[9] R. Ali, T. van Waterschoot, and M. Moonen, "Generalised sidelobe canceller for noise reduction in hearing devices using an external microphone," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 521–525.

[10] ——, "Completing the RTF vector for an MVDR beamformer as applied to a local microphone array and an external microphone," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 211–215.

[11] 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 (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 146–150.

[12] ——, "RTF-steered binaural MVDR beamforming incorporating an external microphone for dynamic acoustic scenarios," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 416–420.

[13] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[14] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 544–548.

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

[16] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[17] T. C. Lawin-Ore and S. Doclo, "Reference microphone selection for MWF-based noise reduction using distributed microphone arrays," in *Proc. ITG Conference on Speech Communication*, Braunschweig, Germany, Sep. 2012, pp. 1–4.

[18] N. Gößling, W. Middelberg, and S. Doclo. (2019) RTF-steered binaural MVDR beamforming incorporating multiple external microphones. [Online]. Available: https://uol.de/en/sigproc/research/audio-demos/binaural-noise-reduction/