RTF-STEERED BINAURAL MVDR BEAMFORMING INCORPORATING AN EXTERNAL MICROPHONE FOR DYNAMIC ACOUSTIC SCENARIOS

Nico Gößling, Simon Doclo

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

ABSTRACT

A well-known binaural noise reduction algorithm is the binaural minimum variance distortionless response beamformer, which can be steered using the relative transfer function (RTF) vectors of the desired source. In this paper, we consider the recently proposed spatial coherence (SC) method to estimate the RTF vectors, requiring an additional external microphone that is spatially separated from the head-mounted microphones. Although the SC method provides a biased estimate of the RTF between the head-mounted microphones and the external microphone, we show that this bias is real-valued and only depends on the SNR in the external microphone. We propose to use the SC method to estimate the extended RTF vectors that also incorporate the external microphone, enabling to filter the external microphone signal in conjunction with the head-mounted microphones. Evaluation results using recorded signals of a moving speaker in diffuse noise show that the SC method yields a slightly better performance than the widely used covariance whitening method at a much lower computational complexity.

***Index Terms*—** noise reduction, binaural cues, external microphone, relative transfer function, MVDR beamformer

1. INTRODUCTION

Noise reduction algorithms for head-mounted assistive listening devices (e.g., hearing aids) are crucial to improve speech intelligibility and speech quality in noisy environments. Binaural noise reduction algorithms are able to use the spatial information captured by all microphones on both sides of the head [1, 2]. Besides suppressing undesired sound sources, binaural noise reduction algorithms also aim at preserving the listener's spatial perception of the acoustic scene, to reduce confusions due to a possible mismatch between acoustical and visual information and to enable the listener to exploit the binaural hearing advantage [3].

As shown in [1, 2, 4], the binaural minimum variance distortionless response beamformer (BMVDR) beamformer is able to preserve the binaural cues, i.e., the interaural level difference (ILD) and the interaural time difference (ITD), of the desired source. The BMVDR beamformer can either be implemented using the acoustic transfer functions (ATFs) between the desired source and all microphones or using the relative transfer functions (RTFs), relating the ATFs to a reference microphone [5].

Aiming at improving the performance of (binaural) noise reduction and source localisation algorithms, recently the use of an external microphone in combination with the head-mounted microphones has been explored [6–13]. In this paper, we consider a recently proposed computationally efficient RTF estimator exploiting the external microphone [11, 12]. This RTF estimator is based on a spatial coherence (SC) assumption about the noise field, namely that the noise component in the external microphone signal is uncorrelated with the noise component in the head-mounted microphone signals. Since the SC method provides a biased estimate of the RTF between the head-mounted microphones and the external microphone, in [11, 12] the SC method was only used to estimate the RTF vector of the head-mounted microphones and not the complete RTF vector including the external microphone.

In this paper, we show that this bias is real-valued (hence not affecting the phase of the RTF estimate) and only depends on the SNR in the external microphone. Furthermore, we show that the filter coefficients of the BMVDR beamformer are only scaled by real-valued factors. Therefore, in this paper we explore the usage of the SC method to estimate the complete RTF vector and compare its performance to the widely used covariance whitening (CW) method [14, 15], which is based on the eigenvalue decomposition of the pre-whitened noisy covariance matrix and hence has a much larger computational complexity than the SC method. Contrary to the simulations in [11, 12] with a spatially stationary speaker, in this paper we consider a highly dynamic scenario with a moving speaker in a reverberant environment with diffuse noise. Simulation results show that the performance of the SC method is similar (even slightly better) than the CW method at a much lower computational complexity. Moreover, simulation results show that the RTF-steered BMVDR beamformers filtering all microphone signals outperform the RTF-steered BMVDR beamformers filtering only the head-mounted microphone signals, a fixed BMVDR beamformer steered towards the frontal direction as well as the external microphone signal.

2. CONFIGURATION AND NOTATION

We consider an acoustic scenario with one desired source $S(\omega)$ and diffuse background noise in a reverberant environment. Moreover, we consider a binaural configuration, consisting of a left and a right device (each containing M microphones), and an external microphone that is spatially separated from the head-mounted microphones, cf. Figure 1. The m-th microphone signal of the left device $Y_{\mathrm{L},m}(\omega)$ can be written in the frequency-domain as

$$Y_{\mathrm{L},m}(\omega) = X_{\mathrm{L},m}(\omega) + N_{\mathrm{L},m}(\omega)\,, \quad m \in \{1,\ldots,M\}\,, \qquad (1)$$

where $X_{\mathrm{L},m}(\omega)$ denotes the desired speech component, $N_{\mathrm{L},m}(\omega)$ denotes the noise component and ω denotes the angular frequency.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 352015383 – SFB 1330 B2 and Cluster of Excellence 1077 Hearing4all, and by the joint Lower Saxony-Israeli Project ATHENA.

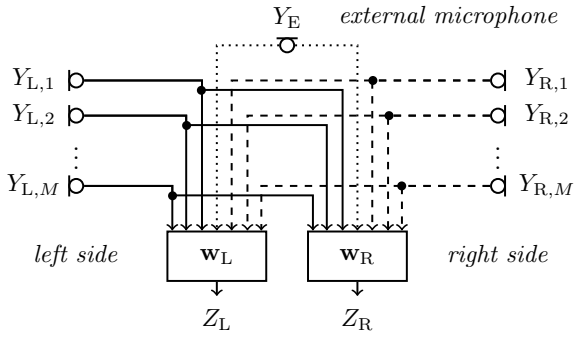

Fig. 1. Binaural hearing device configuration with a spatially separated external microphone.

For conciseness we will omit ω in the remainder of the paper. The m-th microphone signal of the right device $Y_{\mathrm{R},m}$ and the external microphone Y_{E} are defined similarly as in (1). The stacked vector of all head-mounted microphone signals is defined as

$$\mathbf{y} = [Y_{\mathrm{L},1},\ \ldots,\ Y_{\mathrm{L},M},\ Y_{\mathrm{R},1},\ \ldots,\ Y_{\mathrm{R},M}]^T \in \mathbb{C}^{2M}\,, \quad (2)$$

with $(\cdot)^T$ denoting transpose, which can be written as

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

where $\mathbf{x}$ and $\mathbf{n}$ are defined similarly as in (2). For a single desired source, the speech vector $\mathbf{x}$ is equal to

$$\mathbf{x} = \mathbf{a}S\,, \quad (4)$$

where the vector $\mathbf{a} = [A_{\mathrm{L},1},\ \ldots,\ A_{\mathrm{L},M},\ A_{\mathrm{R},1},\ \ldots,\ A_{\mathrm{R},M}]^T$ contains the ATFs between the desired source S and all microphones. Without loss of generality, we choose the first microphone on each device as reference microphone. The RTF vectors $\mathbf{a}_{\mathrm{L}}$ and $\mathbf{a}_{\mathrm{R}}$ of the desired source are defined by relating the ATF vector $\mathbf{a}$ to both reference microphones, i.e.,

$$\mathbf{a}_{\mathrm{L}} = \frac{\mathbf{a}}{\mathbf{e}_{\mathrm{L}}^T\mathbf{a}} = \frac{\mathbf{a}}{A_{\mathrm{L}}}\,, \quad \mathbf{a}_{\mathrm{R}} = \frac{\mathbf{a}}{\mathbf{e}_{\mathrm{R}}^T\mathbf{a}} = \frac{\mathbf{a}}{A_{\mathrm{R}}}\,, \quad (5)$$

where $\mathbf{e}_{\mathrm{L}}$ and $\mathbf{e}_{\mathrm{R}}$ are selection vectors consisting of zeros and one element equal to 1, i.e., $\mathbf{e}_{\mathrm{L}}(1) = 1$ and $\mathbf{e}_{\mathrm{R}}(M+1) = 1$. The speech covariance matrix $\mathbf{R}_{\mathrm{x}} \in \mathbb{C}^{2M\times 2M}$ and the noise covariance matrix $\mathbf{R}_{\mathrm{n}} \in \mathbb{C}^{2M\times 2M}$ are defined as

$$\mathbf{R}_{\mathrm{x}} = \mathcal{E}\{\mathbf{x}\mathbf{x}^H\} = P_{\mathrm{s}}\mathbf{a}\mathbf{a}^H\,, \quad \mathbf{R}_{\mathrm{n}} = \mathcal{E}\{\mathbf{n}\mathbf{n}^H\}\,, \quad (6)$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator, $(\cdot)^H$ denotes the conjugate transpose, and $P_{\mathrm{s}} = \mathcal{E}\{|S|^2\}$ denotes the power spectral density (PSD) of the desired source. Assuming statistical independence between the desired speech and noise components, the noisy covariance matrix is equal to

$$\mathbf{R}_{\mathrm{y}} = \mathcal{E}\{\mathbf{y}\mathbf{y}^H\} = \mathbf{R}_{\mathrm{x}} + \mathbf{R}_{\mathrm{n}}\,. \quad (7)$$

The output signals at the left and the right hearing device are obtained by filtering and summing all microphone signals using the complex-valued filter vectors $\mathbf{w}_{\mathrm{L}}$ and $\mathbf{w}_{\mathrm{R}}$, respectively, i.e.,

$$Z_{\mathrm{L}} = \mathbf{w}_{\mathrm{L}}^H\mathbf{y}\,, \quad Z_{\mathrm{R}} = \mathbf{w}_{\mathrm{R}}^H\mathbf{y}\,. \quad (8)$$

All aforementioned quantities will be referenced to as *extended* when the external microphone signal is included and will be denoted by a bar above the respective variable, e.g., the *extended* RTF vectors

$$\bar{\mathbf{a}}_{\mathrm{L}} = \begin{bmatrix}\mathbf{a}_{\mathrm{L}}\\ A_{\mathrm{E}}/A_{\mathrm{L}}\end{bmatrix}\,, \quad \bar{\mathbf{a}}_{\mathrm{R}} = \begin{bmatrix}\mathbf{a}_{\mathrm{R}}\\ A_{\mathrm{E}}/A_{\mathrm{R}}\end{bmatrix}\,. \quad (9)$$

3. BINAURAL MVDR BEAMFORMER

The BMVDR beamformer [2, 16] minimizes the output noise PSD while preserving the desired speech component in the reference microphones, hence preserving the binaural cues of the desired source. The optimization problem for the left filter vector $\mathbf{w}_{\mathrm{L}}$ is given by

$$\min_{\mathbf{w}_{\mathrm{L}}}\ \mathcal{E}\{|\mathbf{w}_{\mathrm{L}}^H\mathbf{n}|^2\} \quad \text{subject to} \quad \mathbf{w}_{\mathrm{L}}^H\mathbf{a}_{\mathrm{L}} = 1\,. \quad (10)$$

The optimization problem for the right filter vector $\mathbf{w}_{\mathrm{R}}$ is defined similarly. The filter vectors are equal to [1, 2, 5]

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

hence, requiring an estimate of the noise covariance matrix $\mathbf{R}_{\mathrm{n}}$ and the RTF vectors $\mathbf{a}_{\mathrm{L}}$ and $\mathbf{a}_{\mathrm{R}}$. Usually, the noise covariance matrix $\mathbf{R}_{\mathrm{n}}$ is either estimated during speech pauses or approximated using an appropriate model. Similarly, the RTF vectors $\mathbf{a}_{\mathrm{L}}$ and $\mathbf{a}_{\mathrm{R}}$ are either estimated from the microphone signals or approximated using – simulated or measured – anechoic RTFs corresponding to the assumed position of the desired source.
Please note that the *extended* BMVDR beamformer ($\bar{\mathbf{w}}_{\mathrm{L}}$ and $\bar{\mathbf{w}}_{\mathrm{R}}$), incorporating the external microphone, needs an estimate of the *extended* noise covariance matrix $\bar{\mathbf{R}}_{\mathrm{n}}$ and the *extended* RTF vectors $\bar{\mathbf{a}}_{\mathrm{L}}$ and $\bar{\mathbf{a}}_{\mathrm{R}}$ [10, 13].

4. RTF ESTIMATION APPROACHES

In this section, we consider two different methods to estimate the (extended) RTF vectors of the desired source. The well-known CW method [15, 17] requires an estimate of the (extended) noisy and noise covariance matrices to estimate the (extended) RTF vectors. The recently proposed SC method [11, 12] assumes that the spatial coherence between the noise components in the head-mounted microphone signals and the external microphone signal is zero. Although the SC method only provides an unbiased estimate of the RTF vectors $\mathbf{a}_{\mathrm{L}}$ and $\mathbf{a}_{\mathrm{R}}$, in this paper we explore the usage of the biased estimate of the extended RTF vectors $\bar{\mathbf{a}}_{\mathrm{L}}$ and $\bar{\mathbf{a}}_{\mathrm{R}}$. In Section 4.2 the bias on the RTF estimate and the extended filter vectors $\bar{\mathbf{w}}_{\mathrm{L}}$ and $\bar{\mathbf{w}}_{\mathrm{R}}$ is analyzed, showing that the bias is real-valued and only depends on the SNR in the external microphone signal.

4.1. Covariance Whitening (CW) method

Using a square-root decomposition (e.g., Cholesky decomposition) of the noise covariance matrix $\mathbf{R}_{\mathrm{n}}$, i.e.,

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

the pre-whitened noisy covariance matrix is equal to

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

The RTF vectors $\mathbf{a}_{\mathrm{L}}$ and $\mathbf{a}_{\mathrm{R}}$ can then be estimated as [15]

$$\mathbf{a}_{\mathrm{L}}^{\mathrm{CW}} = \frac{\mathbf{R}_{\mathrm{n}}^{1/2}\mathbf{v}}{\mathbf{e}_{\mathrm{L}}^T\mathbf{R}_{\mathrm{n}}^{1/2}\mathbf{v}}\,, \quad \mathbf{a}_{\mathrm{R}}^{\mathrm{CW}} = \frac{\mathbf{R}_{\mathrm{n}}^{1/2}\mathbf{v}}{\mathbf{e}_{\mathrm{R}}^T\mathbf{R}_{\mathrm{n}}^{1/2}\mathbf{v}}\,, \quad (14)$$

with $\mathbf{v}$ the principal eigenvector of $\mathbf{R}_{\mathrm{y}}^{\mathrm{w}}$, i.e., the eigenvector corresponding to the largest eigenvalue. Due to the eigenvalue decomposition, this method is computationally rather complex. Additionally, an estimate of the noise covariance matrix $\mathbf{R}_{\mathrm{n}}$ is required, although

this estimate is required anyway for the BMVDR beamformer.
The extended RTF vectors $\bar{\mathbf{a}}_{\mathrm{L}}^{\mathrm{CWE}}$ and $\bar{\mathbf{a}}_{\mathrm{R}}^{\mathrm{CWE}}$ can be estimated by simply applying the CW method on the extended microphone signal, i.e., based on the principal eigenvector of the pre-whitened extended noisy covariance matrix $\bar{\mathbf{R}}_{\mathrm{y}}^{\mathrm{w}}$. This method will be denoted as CWE.

4.2. Spatial Coherence (SC) method

The SC method [11, 12] assumes that the noise components in the head-mounted microphone signals are uncorrelated with the noise component in the external microphone signal, i.e.,

$$\mathcal{E}\{\mathbf{n}N_{\mathrm{E}}^{*}\} = \mathbf{0}_{2M\times 1}\,. \tag{15}$$

This can be assumed, e.g., for a diffuse noise field when the spatial separation between the external microphone and the head-mounted microphones is large enough. Using (7) and (15), the extended noisy covariance matrix can be written as

$$\bar{\mathbf{R}}_{\mathrm{y}} = \bar{\mathbf{R}}_{\mathrm{x}} + \bar{\mathbf{R}}_{\mathrm{n}} = \left[\begin{array}{c|c} \mathbf{R}_{\mathrm{y}} & \mathcal{E}\{\mathbf{x}X_{\mathrm{E}}^{*}\} \\ \hline \mathcal{E}\{\mathbf{x}^{H}X_{\mathrm{E}}\} & P_{\mathrm{y,E}} \end{array}\right], \tag{16}$$

with the PSD of the external microphone signal $P_{\mathrm{y,E}} = P_{\mathrm{s}}|A_{\mathrm{E}}|^2 + P_{\mathrm{n,E}}$, where $P_{\mathrm{n,E}}$ denotes the noise PSD in the external microphone. Using (4) and (5), an unbiased estimate of the RTF vectors can then be obtained as the normalized first M entries of the last column of $\bar{\mathbf{R}}_{\mathrm{y}}$ [11, 12], i.e.,

$$\mathbf{a}_{\mathrm{L}}^{\mathrm{SC}} = [\mathbf{I}_{2M},\, \mathbf{0}_{2M\times 1}]\,\frac{\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}{\mathbf{e}_{\mathrm{L}}^{T}\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}, \quad \mathbf{a}_{\mathrm{R}}^{\mathrm{SC}} = [\mathbf{I}_{2M},\, \mathbf{0}_{2M\times 1}]\,\frac{\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}{\mathbf{e}_{\mathrm{R}}^{T}\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}. \tag{17}$$

Here, we propose to use the SC method to estimate the extended RTF vectors, including the external microphone as

$$\boxed{\bar{\mathbf{a}}_{\mathrm{L}}^{\mathrm{SCE}} = \frac{\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}{\mathbf{e}_{\mathrm{L}}^{T}\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}},\ \bar{\mathbf{a}}_{\mathrm{R}}^{\mathrm{SCE}} = \frac{\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}{\mathbf{e}_{\mathrm{R}}^{T}\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}}} \tag{18}$$

This method will be denoted as SCE. Compared to the CW and CWE methods, note that the SC and SCE methods do not need an estimate of the (extended) noise covariance matrix and have a lower computational complexity.
By using (16) in (18), it can be shown that the last element of $\bar{\mathbf{a}}_{\mathrm{L}}^{\mathrm{SCE}}$ is biased, i.e.,

$$\mathbf{e}_{\mathrm{E}}^{T}\bar{\mathbf{a}}_{\mathrm{L}}^{\mathrm{SCE}} = \frac{P_{\mathrm{s}}|A_{\mathrm{E}}|^2 + P_{\mathrm{n,E}}}{P_{\mathrm{s}}A_{\mathrm{L}}A_{\mathrm{E}}^{*}} = \mathbf{e}_{\mathrm{E}}^{T}\bar{\mathbf{a}}_{\mathrm{L}}\,(1+\beta)\,, \tag{19}$$

with

$$\boxed{\beta = \frac{P_{\mathrm{n,E}}}{P_{\mathrm{s}}|A_{\mathrm{E}}|^2}} \tag{20}$$

The same holds for the last element of $\bar{\mathbf{a}}_{\mathrm{R}}^{\mathrm{SCE}}$. It can be observed that the bias factor β corresponds to the inverse of the SNR of the external microphone signal. Since this factor is real-valued only the amplitude but not the phase of the RTF estimate between the reference microphones and the external microphone is affected.
Using (11), (15) and (18), the extended filter vector of the left device using the SCE method can be written as

$$\bar{\mathbf{w}}_{\mathrm{L}}^{\mathrm{SCE}} = \frac{\left[\begin{array}{c} \mathbf{R}_{\mathrm{n}}^{-1}\mathbf{a}_{\mathrm{L}} \\ P_{\mathrm{n,E}}\mathbf{e}_{\mathrm{E}}^{T}\bar{\mathbf{a}}_{\mathrm{L}}(1+\beta) \end{array}\right]}{\mathbf{a}_{\mathrm{L}}^{H}\mathbf{R}_{\mathrm{n}}^{-1}\mathbf{a}_{\mathrm{L}} + P_{\mathrm{n,E}}|\mathbf{e}_{\mathrm{E}}^{T}\bar{\mathbf{a}}_{\mathrm{L}}|^2(1+\beta)^2}\,. \tag{21}$$

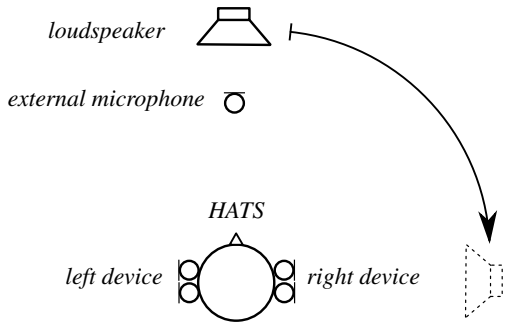

Fig. 2. Experimental setup. The loudspeaker was moved from its initial position in front of the listener to the right side.

It can be shown that the coefficients of the extended filter vector in (21) are equal to the coefficients of the extended BMVDR beamformer $\bar{\mathbf{w}}_{\mathrm{L}}$, scaled by a real-valued scaling factor α, i.e.,

$$\boxed{\bar{\mathbf{w}}_{\mathrm{L}}^{\mathrm{SCE}} = \left[\begin{array}{c} \alpha\cdot[\mathbf{I}_{2M},\, \mathbf{0}_{2M\times 1}]\,\bar{\mathbf{w}}_{\mathrm{L}} \\ \alpha(1+\beta)\cdot\mathbf{e}_{\mathrm{E}}^{T}\bar{\mathbf{w}}_{\mathrm{L}} \end{array}\right]} \tag{22}$$

with

$$\alpha = \frac{\mathbf{a}_{\mathrm{L}}^{H}\mathbf{R}_{\mathrm{n}}^{-1}\mathbf{a}_{\mathrm{L}}^{H} + P_{\mathrm{n,E}}|\mathbf{e}_{\mathrm{E}}^{T}\bar{\mathbf{a}}_{\mathrm{L}}|^2}{\mathbf{a}_{\mathrm{L}}^{H}\mathbf{R}_{\mathrm{n}}^{-1}\mathbf{a}_{\mathrm{L}}^{H} + P_{\mathrm{n,E}}|\mathbf{e}_{\mathrm{E}}^{T}\bar{\mathbf{a}}_{\mathrm{L}}|^2(1+\beta)^2}\,. \tag{23}$$

The same can be shown for the extended filter vector of the right device $\bar{\mathbf{w}}_{\mathrm{R}}$.

5. EXPERIMENTAL RESULTS

In this section, we present an experimental evaluation for a moving speaker of the (extended) BMVDR beamformer using the RTF estimators discussed in Section 4. In Section 5.1 the recording setup is described, while detailed information about the implementation is provided in Section 5.2 and the results are presented in Section 5.3.

5.1. Recording setup

All signals were recorded in a laboratory located at the University of Oldenburg, 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 $350\ \mathrm{ms}$. The experimental setup is depicted in Fig. 2. A KEMAR head-and-torso simulator (HATS) was placed approximately in the centre of the laboratory. Two behind-the-ear hearing aid dummies with two microphones each, i.e., $M = 2$, with inter-microphone distance of about $7\,\mathrm{mm}$, were placed on the ears of the HATS. The external microphone was placed at about $1.5\,\mathrm{m}$ in front of the HATS, which corresponds to, e.g., a table microphone or a smartphone that is connected to the binaural hearing device.
The desired source was a male German speech signal played back by a loudspeaker placed at about $2\,\mathrm{m}$ from the HATS at same height. Initially, the loudspeaker was placed at an angle of $0°$, i.e., in front of the HATS (at a distance of about $0.5\,\mathrm{m}$ to the external microphone). During the first $10\,\mathrm{s}$ the loudspeaker was moved (by hand) to an angle of about $75°$ to the right side of the HATS (at a distance of about $1.5\,\mathrm{m}$ to the external microphone), where it remained for another $5\,\mathrm{s}$. To generate pseudo-diffuse background noise, we placed four loudspeakers facing the corners of the laboratory, playing back different multi-talker recordings. The desired source and the background noise were recorded separately and mixed afterwards to an

average intelligibility-weighted SNR (iSNR) [18] of 0 dB in the reference microphone on the right hearing aid. The average iSNR in the external microphone was equal to about 14 dB. The complete signal had a length of 15 s with 0.5 s of noise-only at the beginning. All signals, i.e., the head-mounted microphone signals and the external microphone signal, were recorded synchronously, thereby neglecting synchronization and latency aspects.

5.2. Implementation and performance measures

We considered five different versions of the BMVDR beamformer in (11), either using the RTF vectors $\mathbf{a}_{\mathrm{L}}$ and $\mathbf{a}_{\mathrm{R}}$ (i.e., filtering only the head-mounted microphone signals but not the external microphone signal) or the extended RTF vectors $\bar{\mathbf{a}}_{\mathrm{L}}$ and $\bar{\mathbf{a}}_{\mathrm{R}}$ (i.e., filtering all available microphone signals), i.e.,

- FIX: Fixed BMVDR beamformer using *anechoic* RTFs $\mathbf{a}_{\mathrm{L}}$ and $\mathbf{a}_{\mathrm{R}}$ calculated from measured impulse responses [19] corresponding to a position in front of the listener.
- CW and CWE: RTF-steered BMVDR beamformer using the RTF estimation method in (14), without and with incorporating the external microphone.
- SC and SCE: RTF-steered BMVDR beamformer using RTF estimation methods in (17) and (18).

In addition, we considered the external microphone (EM) signal without applying any noise reduction.

All signals were processed at a sampling frequency of 16 kHz and transformed to 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 bins, we thresholded a speech presence probability estimate in every time-frequency bin [20]. The noisy covariance matrix $\hat{\mathbf{R}}_{\mathrm{y}}(k,l)$ and the noise covariance matrix $\hat{\mathbf{R}}_{\mathrm{n}}(k,l)$ were then recursively estimated as

$$\hat{\mathbf{R}}_{\mathrm{y}}(k,l) = \alpha_{\mathrm{y}}\hat{\mathbf{R}}_{\mathrm{y}}(k,l-1) + (1-\alpha_{\mathrm{y}})\mathbf{y}(k,l)\mathbf{y}^{H}(k,l)\,, \quad (24)$$

$$\hat{\mathbf{R}}_{\mathrm{n}}(k,l) = \alpha_{\mathrm{n}}\hat{\mathbf{R}}_{\mathrm{n}}(k,l-1) + (1-\alpha_{\mathrm{n}})\mathbf{y}(k,l)\mathbf{y}^{H}(k,l)\,, \quad (25)$$

during detected speech-plus-noise bins and noise-only bins, respectively. The forgetting factors were chosen as $\alpha_{\mathrm{y}} = 0.852$ and $\alpha_{\mathrm{n}} = 0.984$, corresponding to time constants of 100 ms and 1 s, respectively. The extended covariance matrices where estimated using the same procedure. The (time-varying) estimates of the covariance matrices were then used in the different RTF estimators and in the calculation of the (time-varying) BMVDR beamformers.

As performance measure for noise reduction we used the iSNR improvement (ΔiSNR) in blocks of 1 s between the reference microphone signal on the right hearing aid and the output signal on the right hearing aid or the external microphone signal. As performance measure for binaural cue preservation we used the reliable binaural cue errors of the desired speech component, i.e., ΔILD and ΔITD, based on an auditory model [21] and averaged over time and frequency.

5.3. iSNR improvement and binaural cues

Figure 3 depicts the iSNR improvement for all considered BMVDR beamformers and the EM. As expected, FIX leads to the worst performance (even negative ΔiSNR) of all beamformers since it does not track the movement of the desired source. The RTF-steered CW and SC beamformers both, not filtering the external microphone signal, show a similar iSNR improvement between 2 and 9 dB. Averaging the iSNR improvements over time shows that SC outperforms

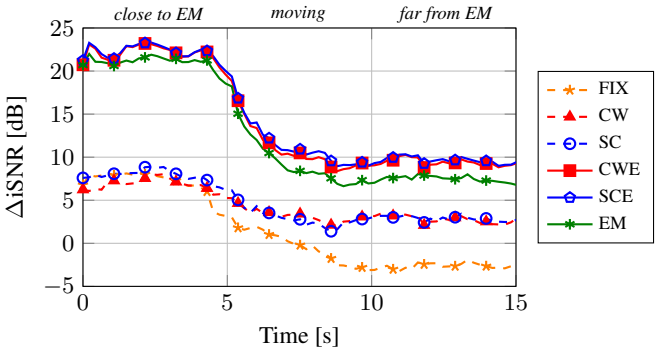

Fig. 3. Intelligibility-weighted SNR improvement (plotted over time) for all considered BMVDR beamformers and the external microphone.

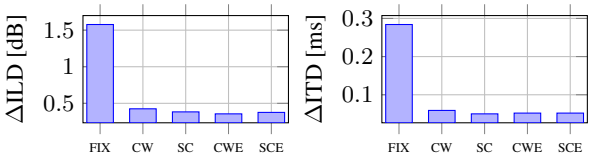

Fig. 4. Reliable binaural cue errors (averaged over time) for all considered BMVDR beamformers.

CW by about 0.31 dB. The RTF-steered CWE and SCE beamformers (both filtering all available microphone signals) outperform all other considered beamformers and the EM. At the initial position of the desired source (0-5 s, close to EM), the CWE and SCE beamformers outperform the CW and SC beamformers by about 14 dB and the EM by about 2 dB. At the final position of the desired source (10-15 s, far from EM), the CWE and SCE beamformers outperform the CW and SC beamformers by about 6 dB and the EM by about 3 dB. Averaging the iSNR improvement over time shows that SCE outperforms CWE by about 0.30 dB.

Figure 4 depicts the ILD and ITD errors for all considered beamformers. It should be stressed that directly using the EM signal does not provide any binaural cues to the user, hence leading to in-head localization. As expected, FIX shows worst performance, i.e., highest binaural cue errors. All RTF-steered BMVDR beamformers show small binaural cue errors, indicating that the desired source is perceived as coming from the correct direction. In conclusion, these results show that the biased SCE estimator can be used for different positions of the desired source and yields a similar (even slightly better) iSNR improvement and similar binaural cues as the CWE estimator at much lower complexity.

6. CONCLUSIONS

In this paper, we proposed to use the SC method to estimate the extended RTF vector including all microphones of a binaural hearing device and an external microphone. We showed, that the last element of the estimate is biased, but also that this bias is real-valued and hence not affecting the phase of the RTF estimate. Experimental evaluation in a highly dynamic scenario with a moving speaker showed that the SC estimator can be used for different positions of the desired source despite this bias and yields similar noise reduction and cue preservation performance as the widely-used CW method at much lower complexity.

7. 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*, chapter 18. Wiley, 2018.

[3] A. W. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1508–1516, 1988.

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

[5] S. Gannot, D. Burshtein, and E. Weinstein, "Signal Enhancement Using Beamforming and Non-Stationarity with Applications to Speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[6] 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, pp. 14 pages, Jan. 2009.

[7] 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.

[8] 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.

[9] 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, 2017.

[10] N. Gößling, D. Marquardt, and S. Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation in a homogeneous noise field," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, USA, Mar. 2017, pp. 1–5.

[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] 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.

[13] R. Ali, T. van Watershoot, and M. Moonen, "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.

[14] 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.

[15] 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.

[16] 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.

[17] 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 Transactions on Audio, Speech and Language Processing*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[18] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.

[19] 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, pp. 10 pages, 2009.

[20] 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.

[21] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, vol. 53, pp. 592–605, 2011.