Comparison of Binaural RTF-Vector-Based Direction of Arrival Estimation Methods Exploiting an External Microphone

Daniel Fejgin and Simon Doclo
Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, University of Oldenburg, Germany
{daniel.fejgin,simon.doclo}@uol.de

***Abstract*—In this paper we consider a binaural hearing aid setup, where in addition to the head-mounted microphones an external microphone is available. For this setup, we investigate the performance of several relative transfer function (RTF) vector estimation methods to estimate the direction of arrival (DOA) of the target speaker in a noisy and reverberant acoustic environment. More in particular, we consider the state-of-the-art covariance whitening (CW) and covariance subtraction (CS) methods, either incorporating the external microphone or not, and the recently proposed spatial coherence (SC) method, requiring the external microphone. To estimate the DOA from the estimated RTF vector, we propose to minimize the frequency-averaged Hermitian angle between the estimated head-mounted RTF vector and a database of prototype head-mounted RTF vectors. Experimental results with stationary and moving speech sources in a reverberant environment with diffuse-like noise show that the SC method outperforms the CS method and yields a similar DOA estimation accuracy as the CW method at a lower computational complexity.**

***Index Terms*—direction of arrival estimation, relative transfer function, external microphone, binaural hearing aids**

I. Introduction

For binaural hearing aid (HA) applications, estimating the direction of arrival (DOA) of the target speaker in a noisy and reverberant acoustic environment is important, e.g., to steer a beamformer towards this speaker [1]. Several methods have been proposed for binaural DOA estimation, e.g., based on interaural time and level differences [2], [3], generalized cross-correlation (GCC) features [4]–[7], or relative transfer function (RTF) vectors [8]. For a binaural HA setup incorporating an external microphone, an RTF-vector-based DOA estimation method was proposed in [9], where it was however assumed that the external microphone was worn by the target speaker, such that the external microphone signal almost does not capture any noise or reverberation.

To estimate the RTF vector of the target speaker from noisy microphone signals, several methods have been proposed in the literature [10]–[14], where the most popular methods are based on covariance subtraction (CS) or covariance whitening (CW). These methods require an estimate of the covariance matrix of the noisy microphone signals (e.g., estimated during speech-plus-noise time-frequency (TF) bins) and the noise covariance matrix (e.g., estimated during noise-only TF bins). In should be realized that due to the involved eigenvalue decomposition the computational complexity of the CW method is in general high, which is especially relevant for an online implementation. Exploiting the availability of an external microphone, in [15], [16] the spatial coherence (SC) method was proposed to estimate the (head-mounted) RTF vectors. The SC method relies on the assumption that the coherence between the noise component in the external microphone signal and the noise components in the head-mounted microphone signals is low. This assumption holds quite well, for example, when the distance between the external microphone and the head-mounted microphones is large enough and the noise field is diffuse-like. In comparison to the CS and CW methods an additional advantage is the fact that no estimate of the noise covariance matrix is required.

For a binaural hearing aid setup with an external microphone that is not worn by the target speaker, in this paper we analyze the performance of several RTF-vector-based DOA estimation methods, more in particular, CS and CW (either incorporating the external microphone or not) and SC (incorporating the external microphone). Instead of using a statistical classifier or a neural network to estimate the DOA from the estimated RTF vectors [17], [18], we follow an approach similar to [5], [8], where the estimated head-mounted RTF vectors are compared to a database of anechoic prototype RTF vectors for several directions. However, instead of using a least-squares-based cost function, we propose to use a cost function based on the Hermitian angle. Experimental results using recorded signals in a reverberant environment with diffuse-like noise show that the SC method outperforms the CS method and yields a similar DOA estimation accuracy as the more computational complex CW method, both for a static as well as for a moving target speaker and for several positions of the external microphone.

II. Signal Model

We consider a binaural hearing aid setup consisting of M head-mounted microphones and one external microphone, which is spatially separated from the head-mounted microphones, thus, $M+1$ microphones in total. We consider a single speech source at DOA θ_{s} (in the azimuthal plane) in a noisy and reverberant acoustic environment, see Fig. 1. The m-th microphone signal can

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 352015383 (SFB 1330 B2) and Project ID 390895286 (EXC 2177/1).

be written in the short-time Fourier transform (STFT) domain as

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

where the speech and noise components at the k-th frequency bin ($k \in \{1,...,K\}$) and the l-th frame ($l \in \{1,...,L\}$) are denoted by $X_m(k,l)$ and $N_m(k,l)$, respectively. Since all frequency bins are assumed to be independent and are hence treated independently, we will omit the index k in the remainder of the paper where possible. Stacking the $M+1$ microphone signals in a vector $\mathbf{y}(l) = [Y_1(l),...,Y_{M+1}(l)]^T$, where $(\cdot)^T$ denotes transposition, and defining $\mathbf{x}(l)$ and $\mathbf{n}(l)$ similarly as $\mathbf{y}(l)$, the vector $\mathbf{y}(l)$ can be written as

$$\mathbf{y}(l) = \mathbf{x}(l) + \mathbf{n}(l) \in \mathbb{C}^{M+1}. \tag{2}$$

Assuming that the multiplicative transfer function approximation [19] holds, the speech vector $\mathbf{x}(l)$ can be written as

$$\mathbf{x}(l) = \mathbf{g}(l) X_1(l), \tag{3}$$

where the $(M+1)$-dimensional *extended* RTF vector

$$\mathbf{g}(l) = [1, G_2(l),...,G_{M+1}(l)]^T \tag{4}$$

contains the reverberant RTFs of the speech source between all microphones (including the external microphone) and the reference microphone, for which we have used the first microphone without loss of generality. The M-dimensional *head-mounted* RTF vector $\mathbf{g}_{\mathrm{h}}(l)$ corresponding to the head-mounted microphones can be extracted from $\mathbf{g}(l)$ in (4) as

$$\mathbf{g}_{\mathrm{h}}(l) = \mathbf{E}\mathbf{g}(l), \quad \mathbf{E} = [\mathbf{I}_{M\times M}, \mathbf{0}_M], \tag{5}$$

where $\mathbf{I}_{M\times M}$ is the $M \times M$-dimensional identity matrix and $\mathbf{0}_M$ is the M-dimensional zero vector. Since it can be assumed that the relative positions of the head-mounted microphones are fixed (ignoring small movements of the hearing aids due to head movements) whereas the external microphone can be located at an arbitrary position, it should be realized that although the extended RTF vector $\mathbf{g}(l)$ encodes the DOA θ_{s}, it depends on the (unknown) position of the external microphone, whereas the head-mounted RTF vector $\mathbf{g}_{\mathrm{h}}(l)$ encodes the DOA θ_{s} and obviously does not depend on the position of the external microphone. Hence, for DOA estimation, we will only consider the head-mounted RTF vector.

The $(M+1) \times (M+1)$-dimensional speech and noise covariance matrices are defined as

$$\mathbf{\Phi}_{\mathrm{x}}(l) = \mathcal{E}\{\mathbf{x}(l)\mathbf{x}^H(l)\} = \mathbf{g}(l)\mathbf{g}^H(l)\Phi_{X_1}(l), \tag{6}$$

$$\mathbf{\Phi}_{\mathrm{n}}(l) = \mathcal{E}\{\mathbf{n}(l)\mathbf{n}^H(l)\}, \tag{7}$$

where $\Phi_{X_1}(l) = \mathcal{E}\{|X_1(l)|^2\}$ denotes the power spectral density of the speech component in the reference microphone signal, and the operators $(\cdot)^H$ and $\mathcal{E}\{\cdot\}$ denote complex transposition and expectation, respectively. Assuming uncorrelated speech and noise components, the covariance matrix of the noisy microphone signals $\mathbf{\Phi}_{\mathrm{y}}(l)$ can be written as

$$\mathbf{\Phi}_{\mathrm{y}}(l) = \mathcal{E}\{\mathbf{y}(l)\mathbf{y}^H(l)\} = \mathbf{\Phi}_{\mathrm{x}}(l) + \mathbf{\Phi}_{\mathrm{n}}(l). \tag{8}$$

The $M \times M$-dimensional covariance matrices corresponding to the head-mounted microphones can be extracted from (6) - (8) as

$$\mathbf{\Phi}_{\mathrm{x,h}}(l) = \mathbf{E}\mathbf{\Phi}_{\mathrm{x}}(l)\mathbf{E}^T, \quad \mathbf{\Phi}_{\mathrm{n,h}}(l) = \mathbf{E}\mathbf{\Phi}_{\mathrm{n}}(l)\mathbf{E}^T, \tag{9}$$

$$\mathbf{\Phi}_{\mathrm{y,h}}(l) = \mathbf{E}\mathbf{\Phi}_{\mathrm{y}}(l)\mathbf{E}^T = \mathbf{\Phi}_{\mathrm{x,h}}(l) + \mathbf{\Phi}_{\mathrm{n,h}}(l). \tag{10}$$

III. RTF Vector Estimation

In this section we discuss several RTF vector estimation methods. In Sections III-A and III-B we review the state-of-the-art covariance subtraction (CS) and covariance whitening (CW) methods [10], [11], [14], which are general methods that can be used to estimate the extended RTF vector (using all microphones) or the head-mounted RTF vector (using only the head-mounted microphones). In Section III-C we discuss the recently proposed spatial coherence method [15], [16], which requires the availability of an external microphone to estimate the head-mounted RTF vector.

A. Covariance Subtraction (CS)

Using (6) and (8), the extended RTF vector $\mathbf{g}(l)$ can be obtained from any column of the rank-1 speech covariance matrix $\mathbf{\Phi}_{\mathrm{x}}(l)$ with appropriate normalization [10], [14], i.e.,

$$\mathbf{g}(l) = \frac{\mathbf{\Phi}_{\mathrm{x}}(l)\mathbf{e}_j}{\mathbf{e}_1^T\mathbf{\Phi}_{\mathrm{x}}(l)\mathbf{e}_j} = \frac{(\mathbf{\Phi}_{\mathrm{y}}(l) - \mathbf{\Phi}_{\mathrm{n}}(l))\mathbf{e}_j}{\mathbf{e}_1^T(\mathbf{\Phi}_{\mathrm{y}}(l) - \mathbf{\Phi}_{\mathrm{n}}(l))\mathbf{e}_j}, \tag{11}$$

where $\mathbf{e}_j = [0,...,1,0...,0]^T$ is an $(M+1)$-dimensional vector with zeros except the j-th element. In practice, estimates of the noisy covariance matrix $\hat{\mathbf{\Phi}}_{\mathrm{y}}(l)$ and the noise covariance matrix $\hat{\mathbf{\Phi}}_{\mathrm{n}}(l)$ are used (e.g., obtained via recursive smoothing during speech-plus-noise and noise-only TF bins), yielding the CS estimate of the extended RTF vector

$$\hat{\mathbf{g}}^{(\mathrm{CS})}(l) = \frac{\left(\hat{\mathbf{\Phi}}_{\mathrm{y}}(l) - \hat{\mathbf{\Phi}}_{\mathrm{n}}(l)\right)\mathbf{e}_j}{\mathbf{e}_1^T\left(\hat{\mathbf{\Phi}}_{\mathrm{y}}(l) - \hat{\mathbf{\Phi}}_{\mathrm{n}}(l)\right)\mathbf{e}_j}. \tag{12}$$

Similarly, when using the covariance matrices corresponding to the head-mounted microphones (i.e., not exploiting the external microphone), the CS estimate of the head-mounted RTF vector is given by

$$\boxed{\hat{\mathbf{g}}_{\mathrm{h}}^{(\mathrm{CS})}(l) = \frac{\left(\hat{\mathbf{\Phi}}_{\mathrm{y,h}}(l) - \hat{\mathbf{\Phi}}_{\mathrm{n,h}}(l)\right)\mathbf{e}_{\mathrm{h,j}}}{\mathbf{e}_{\mathrm{h},1}^T\left(\hat{\mathbf{\Phi}}_{\mathrm{y,h}}(l) - \hat{\mathbf{\Phi}}_{\mathrm{n,h}}(l)\right)\mathbf{e}_{\mathrm{h,j}}}} \tag{13}$$

where $\mathbf{e}_{\mathrm{h,j}} = [0,...,1,0...,0]^T$ is an M-dimensional vector with zeros except the j-th element. It can be easily shown that

$$\hat{\mathbf{g}}_{\mathrm{h}}^{(\mathrm{CS})}(l) = \mathbf{E}\hat{\mathbf{g}}^{(\mathrm{CS})}(l), \tag{14}$$

such that this estimate does not depend on the position of the external microphone. Hence, in the experiments in Section V we will only consider one version of the CS method (without the external microphone).

B. Covariance Whitening (CW)

Instead of subtracting $\hat{\mathbf{\Phi}}_{\mathrm{n}}(l)$ from $\hat{\mathbf{\Phi}}_{\mathrm{y}}(l)$, the CW method first prewhitens the estimated noisy covariance matrix with a square-root decomposition (e.g., Cholesky decomposition) of the estimated noise covariance matrix [11], [14], i.e.,

$$\hat{\mathbf{\Phi}}_{\mathrm{n}}(l)=\hat{\mathbf{L}}_{\mathrm{n}}(l)\hat{\mathbf{L}}_{\mathrm{n}}^{H}(l),\quad \hat{\mathbf{\Phi}}_{\mathrm{y}}^{(\mathrm{w})}(l)=\hat{\mathbf{L}}_{\mathrm{n}}^{-1}(l)\hat{\mathbf{\Phi}}_{\mathrm{y}}(l)\hat{\mathbf{L}}_{\mathrm{n}}^{-H}(l). \quad (15)$$

The CW estimate of the extended RTF vector is then obtained as the normalized de-whitened principal eigenvector of the pre-whitened noisy covariance matrix, i.e.,

$$\boxed{\hat{\mathbf{g}}^{(\mathrm{CW})}(l)=\frac{\hat{\mathbf{L}}_{\mathrm{n}}(l)\mathcal{P}\left\{\hat{\mathbf{\Phi}}_{\mathrm{y}}^{(\mathrm{w})}(l)\right\}}{\mathbf{e}_1^T\hat{\mathbf{L}}_{\mathrm{n}}(l)\mathcal{P}\left\{\hat{\mathbf{\Phi}}_{\mathrm{y}}^{(\mathrm{w})}(l)\right\}}} \quad (16)$$

where $\mathcal{P}\{\cdot\}$ denotes the principal eigenvector of a matrix.

Similarly, when using the covariance matrices corresponding to the head-mounted microphones (i.e., not exploiting the external microphone), the CW estimate of the head-mounted RTF vector is given by

$$\boxed{\hat{\mathbf{g}}_{\mathrm{h}}^{(\mathrm{CW})}(l)=\frac{\hat{\mathbf{L}}_{\mathrm{n,h}}(l)\mathcal{P}\left\{\hat{\mathbf{\Phi}}_{\mathrm{y,h}}^{(\mathrm{w})}(l)\right\}}{\mathbf{e}_{\mathrm{h},1}^T\hat{\mathbf{L}}_{\mathrm{n,h}}(l)\mathcal{P}\left\{\hat{\mathbf{\Phi}}_{\mathrm{y,h}}^{(\mathrm{w})}(l)\right\}}} \quad (17)$$

with

$$\hat{\mathbf{\Phi}}_{\mathrm{n,h}}(l)=\hat{\mathbf{L}}_{\mathrm{n,h}}(l)\hat{\mathbf{L}}_{\mathrm{n,h}}^{H}(l),\quad \hat{\mathbf{\Phi}}_{\mathrm{y,h}}^{(\mathrm{w})}(l)=\hat{\mathbf{L}}_{\mathrm{n,h}}^{-1}(l)\hat{\mathbf{\Phi}}_{\mathrm{y,h}}(l)\hat{\mathbf{L}}_{\mathrm{n,h}}^{-H}(l). \quad (18)$$

Since contrary to the CS method

$$\hat{\mathbf{g}}_{\mathrm{h}}^{(\mathrm{CW})}(l)\neq\mathbf{E}\hat{\mathbf{g}}^{(\mathrm{CW})}(l) \quad (19)$$

in the experiments in Section V we will consider two versions of the CW method, either exploiting the external microphone or not. Due to the required square-root decomposition in (15) or (18) and the eigenvalue decomposition in (16) or (17), the computational complexity for the CW method is larger than for the CS method.

C. Spatial Coherence (SC)

The SC method [15], [16] requires an external microphone and assumes a low coherence between the noise component in the external microphone signal and the noise components in the head-mounted microphone signals, i.e.,

$$\mathcal{E}\{N_i(l)N_{M+1}^*(l)\}\approx 0,\quad i\in\{1,...,M\}, \quad (20)$$

As shown in [15], [16], this assumption holds quite well for a diffuse-like noise field (e.g., multi-talker babble noise) when the distance between the external microphone and the head-mounted microphones is large enough. Using (20), it can be easily shown that

$$\mathcal{E}\{Y_i(l)Y_{M+1}^*(l)\}=\mathcal{E}\{X_i(l)X_{M+1}^*(l)\},\ i\in\{1,...,M\}, \quad (21)$$

such that, using (6), the head-mounted RTF vector can be estimated from the $(M+1)$-th column of the estimated noisy covariance matrix $\hat{\mathbf{\Phi}}_{\mathrm{y}}(l)$ as

$$\boxed{\hat{\mathbf{g}}_{\mathrm{h}}^{(\mathrm{SC})}(l)=\mathbf{E}\frac{\hat{\mathbf{\Phi}}_{\mathrm{y}}(l)\mathbf{e}_{M+1}}{\mathbf{e}_1^T\hat{\mathbf{\Phi}}_{\mathrm{y}}(l)\mathbf{e}_{M+1}}} \quad (22)$$

The SC method has a similar computational complexity as the CS method and a lower complexity as the CW method, but contrary to the CS and CW method does not require an estimate of the noise covariance matrix $\hat{\mathbf{\Phi}}_{\mathrm{n}}(l)$.

IV. DoA Estimation

To estimate the possibly time-varying DOA $\theta_{\mathrm{s}}(l)$ of the target speaker from the estimated head-mounted RTF vector $\hat{\mathbf{g}}_{\mathrm{h}}(k,l)$, different approaches have been proposed[1]. Instead of using a statistical classifier or a neural network as in [17], [18], in [5], [8] it has been proposed to simply compare the estimated head-mounted RTF vector with a database of anechoic prototype head-mounted RTF vectors $\bar{\mathbf{g}}_{\mathrm{h}}(k,\theta_i)$ for different discrete directions θ_i, $i=1,...,I$. These prototype head-mounted RTF vectors can either be obtained using, e.g., a spherical diffraction model [20], or measured using the same microphone array configuration as used during the actual source localization. Whereas the cost functions in [5], [8] use the (squared) norm between the (normalized) estimated and prototype head-mounted RTF vectors, in this paper we propose to use the so-called Hermitian angle [13] between the estimated and prototype head-mounted RTF vectors, i.e.,

$$d(k,l,\theta_i)=\arccos\left(\frac{|\bar{\mathbf{g}}_{\mathrm{h}}^H(k,\theta_i)\hat{\mathbf{g}}_{\mathrm{h}}(k,l)|}{\|\bar{\mathbf{g}}_{\mathrm{h}}(k,\theta_i)\|_2\|\hat{\mathbf{g}}_{\mathrm{h}}(k,l)\|_2}\right), \quad (23)$$

since this resulted in a better DOA estimation accuracy. The DOA of the target speaker is then estimated as the direction for which the frequency-averaged cost function in (23) is minimal, i.e.,

$$\hat{\theta}_{\mathrm{s}}(l)=\underset{\theta_i}{\operatorname{argmin}}\ J(l,\theta_i)=\underset{\theta_i}{\operatorname{argmin}}\ \frac{1}{K-1}\sum_{k=2}^{K}d(k,l,\theta_i). \quad (24)$$

V. Experimental Results

In this section we compare the DOA estimation accuracy using four different RTF vector estimates:

- The CS-based estimate $\hat{\mathbf{g}}_{\mathrm{h}}^{(\mathrm{CS})}(l)$ in (13) using only the head-mounted microphones. It should be noted that this is similar to the binaural DOA estimation method presented in [8].
- The CW-based estimates $\mathbf{E}\hat{\mathbf{g}}^{(\mathrm{CW})}(l)$ based on (16), using all microphones, and $\hat{\mathbf{g}}_{\mathrm{h}}^{(\mathrm{CW})}(l)$ in (17) using only the head-mounted microphones.
- The SC-based estimate $\hat{\mathbf{g}}_{\mathrm{h}}^{(\mathrm{SC})}(l)$ in (22) using all microphones.

The experimental setup and implementation details are described in Section V-A. Experimental results for a static and a moving speaker in a reverberant environment with diffuse-like noise are presented in Section V-B.

A. Experimental setup and implementation details

For the experiments we used recordings in a laboratory at the University of Oldenburg with dimensions about $(7\times6\times2.7)\mathrm{m}^3$, where the reverberation time can be easily changed by closing and opening absorber panels mounted to the walls and ceiling. Fig. 1 depicts the experimental setup, where a dummy head

[1] As already mentioned, since the estimated extended RTF vector $\hat{\mathbf{g}}(k,l)$ depends on the (unknown) position of the external microphone, it cannot be straightforwardly used for DOA estimation.

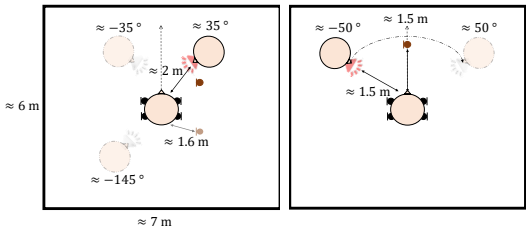

Fig. 1. Experimental setup for stationary speaker scenario (left) and moving speaker scenario (right). The external microphone is depicted in brown whereas the head-mounted microphones are depicted in black.

with binaural hearing aids ($M = 4$ microphones) is located approximately in the center of the laboratory. The external microphone is not restricted to be close to the target speaker. We consider two scenarios, either with a stationary speaker or with a moving speaker. For both scenarios, the speech and noise components were recorded separately. Diffuse-like noise was generated with four loudspeakers facing the corners of the laboratory, playing back different multi-talker recordings. The signal-to-noise ratio (SNR) was set as the ratio of the average broadband speech power to broadband noise power in the front microphones of both hearing aids.

For the stationary speaker scenario, three different positions of the speech source and two different positions of the external microphone are considered (see Fig. 1). The speech source is located at approximately 2m from the dummy head at either $-145°$, $-35°$, or $35°$. The external microphone is located at approximately 1.6m from the dummy head at either $45°$ or $130°$. The speech source is constantly active and comprises English sentences (duration: 30s).

For the moving speaker scenario, a male speaker moves from approximately $-50°$ to $50°$ at a distance of about 1.5m from the dummy head (see Fig. 1). The external microphone is located at approximately 1.5m in front of the dummy head. The speaker is constantly active (duration: 25s).

The microphone signals are recorded at a sampling frequency $f_\text{s} = 16\text{kHz}$ and processed in the STFT-domain using a 32ms square-root Hann window with 50% overlap. The noisy and noise covariance matrices are recursively estimated during detected speech-plus-noise and noise-only TF-bins, respectively, as in (25) and (26) using smoothing factors α_y and α_n corresponding to time constants of 250ms for $\hat{\mathbf{\Phi}}_\text{y}(k,l)$ and 500ms for $\hat{\mathbf{\Phi}}_\text{n}(k,l)$ for the stationary speaker scenario and using smoothing factors corresponding to time constants of 150ms for $\hat{\mathbf{\Phi}}_\text{y}(k,l)$ and 500ms for $\hat{\mathbf{\Phi}}_\text{n}(k,l)$ for the moving speaker scenario.

$$\hat{\mathbf{\Phi}}_\text{y}(k,l) = \alpha_\text{y}\hat{\mathbf{\Phi}}_\text{y}(k,l-1) + \mathbf{y}(k,l)\mathbf{y}^H(k,l) \tag{25}$$

$$\hat{\mathbf{\Phi}}_\text{n}(k,l) = \alpha_\text{n}\hat{\mathbf{\Phi}}_\text{y}(k,l-1) + \mathbf{y}(k,l)\mathbf{y}^H(k,l). \tag{26}$$

Speech-plus-noise and noise-only TF bins are distinguished based on the speech presence probabilities [21] in the head-mounted microphones, which are averaged and thresholded per TF bin. For the stationary speaker scenario initialization effects are mitigated by using the first half of the signal as initialization period and evaluating the performance on the second half only. The prototype head-mounted RTF vectors $\bar{\mathbf{g}}_\text{h}(k,\theta_i)$ were

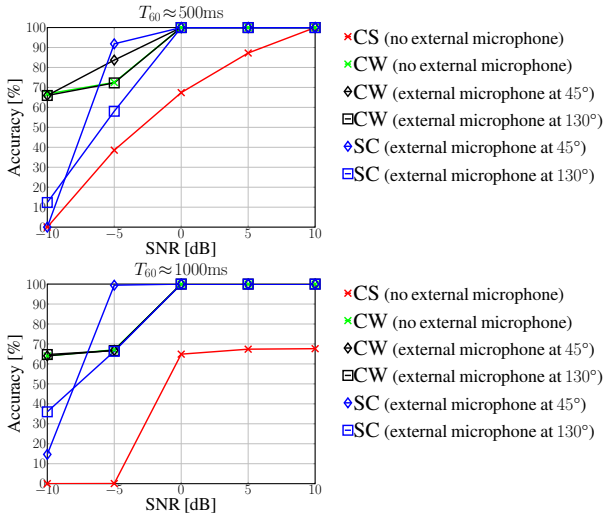

Fig. 2. Average localization accuracy for all considered RTF vector estimation methods for different SNRs. Top: $T_{60} \approx 500\text{ms}$, bottom: $T_{60} \approx 1000\text{ms}$.

generated using the database of binaural anechoic room impulse responses in [22] with an angular resolution of $5°$ ($I = 72$).

As performance measure we use the localization accuracy, i.e., the percentage of correctly localized frames, defined as

$$\text{ACC} = \frac{1}{L}\sum_{l=1}^{L} U\Big(\Delta\theta - f\Big(|\hat{\theta}_\text{s}(l) - \theta_\text{s}(l)|\Big)\Big) \times 100\%, \tag{27}$$

where U is the Heaviside step function and $f(\cdot)$ is a circular wrapping function to ensure an absolute error smaller than $180°$. As tolerance we used $\Delta\theta = 5°$, which corresponds to the resolution of the prototype RTF vectors.

B. DOA estimation accuracy

For the stationary speaker scenario, Fig. 2 depicts the localization accuracy (averaged over the three speaker positions) for all considered RTF vector estimation methods as a function of SNR for two reverberation times ($T_{60} \approx 500\text{ms}$, $T_{60} \approx 1000\text{ms}$). For the CW and SC methods exploiting the external microphone, the performance is shown for both considered positions of the external microphone ($45°$,$130°$). First, it can be observed that for both reverberation times and for all SNRs the CW and SC methods outperform the CS method. Second, it can be observed that for both reverberation times and for all SNRs except -10dB the SC method yields a similar localization accuracy as the CW methods. The performance of the SC method appears to depend more on the position of the external microphone than the performance of the CW method, which is especially noticeable at $\text{SNR} = -5\text{dB}$.

For the moving speaker scenario, we only consider the SC and CW methods incorporating the external microphone. Fig. 3 depicts for an SNR of 0dB and $T_{60} \approx 400\text{ms}$ the time-varying estimated DOA $\hat{\theta}_\text{s}(l)$ (solid red line), while the gray background encodes the cost function $J(l,\theta_i)$ in (24). Although no exact ground-truth DOA is available for the moving speaker scenario, it can be observed that the moving speaker can be localized well using both considered RTF vector estimation methods. In addition, it can be observed that a higher localization confidence is obtained

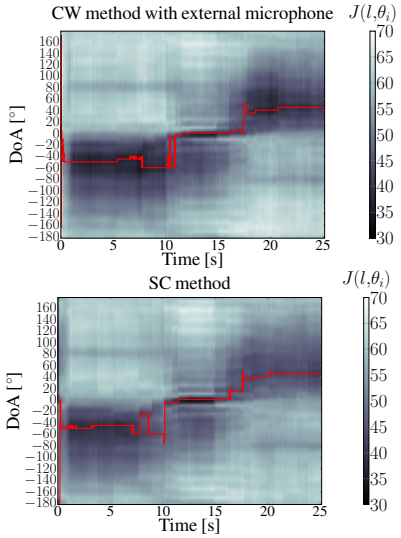

Fig. 3. Localization performance for the moving speaker scenario for SNR $= 0\,\mathrm{dB}$ and $T_{60} \approx 400\,\mathrm{ms}$. Top: CW method with external microphone, bottom: SC method

using the SC method than using the CW method, because the region of small Hermitian angles around the estimated DOA is more confined for the SC method than for the CW method.

The DOA estimation results for the stationary and moving speaker scenario show that the low-complexity SC method yields a comparable performance as the CW method, which is in line with the beamforming results reported in [15], [23].

VI. Conclusions

In this paper we analyzed the DOA estimation performance based on several RTF vector estimation methods for a binaural hearing aid setup with an external microphone that is not restricted to be close to the target speaker. More in particular, we compared the performance of the state-of-the-art CW and CS methods with the SC method. To estimate the DOA from the estimated head-mounted RTF vector, we proposed to minimize the frequency-averaged Hermitian angle between the estimated head-mounted RTF vector and anechoic prototype head-mounted RTF vectors for several directions. Experimental results with real-world data for stationary and moving speaker scenarios show that exploiting the external microphone using the SC method yields a similar DOA estimation accuracy as the CW method at a lower computational complexity.

References

[1] D. Marquardt and S. Doclo, "Performance comparison of bilateral and binaural MVDR-based noise reduction algorithms in the presence of DOA estimation errors," *Proc. ITG Symposium on Speech Communication*, Paderborn, Germany, Oct. 2016, pp. 1-5.

[2] M. Raspaud, H. Viste and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68-77, Jan. 2010.

[3] T. May, S. van de Par and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2016-2030, Sept. 2012.

[4] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," *Proc. International Workshop on Acoustic Signal Enhancement* (*IWAENC*), Juan-les-Pins, France, Sept. 2014, pp. 99-103.

[5] D. Marquardt and S. Doclo, "Noise power spectral density estimation for binaural noise reduction exploiting direction of arrival estimates," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), New Paltz, USA, Oct. 2017, pp. 234-238.

[6] N. Ma, J. A. Gonzalez and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122-2131, Nov. 2018.

[7] R. Varzandeh, K. Adiloğlu, S. Doclo and V. Hohmann, "Exploiting periodicity features for joint detection and DOA estimation of speech sources using convolutional neural networks," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Barcelona, Spain, May 2020, pp. 566-570.

[8] S. Braun, W. Zhou and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), New Paltz, USA, Oct. 2015, pp. 1-5.

[9] M. Farmani, M. S. Pedersen, Z. Tan and J. Jensen, "Bias-compensated informed sound source localization using relative transfer functions," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1275-1289, July 2018.

[10] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 5, pp. 451-459, Sept. 2004.

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

[12] A. Krueger, E. Warsitz and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 206-219, Jan. 2011.

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

[14] S. Markovich-Golan, S. Gannot and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," *Proc. European Signal Processing Conference* (*EUSIPCO*), Rome, Italy, Sept. 2018, pp. 2499-2503.

[15] N. Gößling and S. Doclo, "Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field," *Proc. International Workshop on Acoustic Signal Enhancement* (*IWAENC*), Tokyo, Japan, Sept. 2018, pp. 146–150.

[16] N. Gößling, "Binaural beamforming algorithms and parameter estimation methods exploiting external microphones", PhD Thesis, University of Oldenburg, Germany, Oct. 2020.

[17] X. Li, L. Girin, R. Horaud and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2171-2186, Nov. 2016.

[18] H. Hammer, S. E. Chazan, J. Goldberger, and S. Gannot, "FCN approach for dynamically locating multiple speakers," Aug. 2020, [Online], available: https://arxiv.org/abs/2008.11845.

[19] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337-340, May 2007.

[20] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, Nov. 1998.

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

[22] 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. 1–10, Jan. 2009.

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