DOI: 10.61782/fa.2023.1003

forum acusticum 2023

EXPLOITING AN EXTERNAL MICROPHONE FOR BINAURAL RTF-VECTOR-BASED DIRECTION OF ARRIVAL ESTIMATION FOR MULTIPLE SPEAKERS

Daniel Fejgin* **Simon Doclo**
University of Oldenburg, Department of Medical Physics and Acoustics
and Cluster of Excellence Hearing4all, Oldenburg, Germany

ABSTRACT

In hearing aid applications, an important objective is to accurately estimate the direction of arrival (DOA) of multiple speakers in noisy and reverberant environments. Recently, we proposed a binaural DOA estimation method, where the DOAs of the speakers are estimated by selecting the directions for which the so-called Hermitian angle spectrum between the estimated relative transfer function (RTF) vector and a database of prototype anechoic RTF vectors is maximized. The RTF vector is estimated using the covariance whitening (CW) method, which requires a computationally complex generalized eigenvalue decomposition. The spatial spectrum is obtained by only considering frequencies where it is likely that one speaker dominates over the other speakers, noise and reverberation. In this contribution, we exploit the availability of an external microphone that is spatially separated from the hearing aid microphones and consider a low-complexity RTF vector estimation method that assumes a low spatial coherence between the undesired components in the external microphone and the hearing aid microphones. Using recordings of two speakers and diffuse-like babble noise in acoustic environments with mild reverberation and low signal-to-noise ratio, simulation results show that the proposed method yields a comparable DOA estimation performance as the CW method at a lower computational complexity.

Keywords: *direction of arrival estimation, relative transfer function, external microphone, binaural hearing aids*

*__Corresponding author__: *daniel.fejgin@uol.de.*

1. INTRODUCTION

In speech communication applications such as hearing aids, methods for estimating the direction of arrival (DOA) of multiple speakers are often required. To solve this estimation task, (deep) learning-based and model-based methods are continuously developed and advanced [1, 2]. However, only few methods exploit the availability of external mobile devices equipped with microphones [3–6], although wirelessly linking hearing aids to these devices has become increasingly popular [7].

Recently, we proposed relative-transfer-function (RTF) vector-based DOA estimation methods for a single speaker in [5, 6], without relying on the external microphone to be close to the target speaker and capturing only little noise or reverberation as in [3, 4]. We estimated the DOA as the direction that maximized the similarity between the estimated RTF vector and a database of prototype anechoic RTF vectors for different directions in terms of a frequency-averaged distance function.

However, the methods in [3–6] considered only a single speaker. To address DOA estimation for multiple speakers, we introduced the so-called frequency-averaged Hermitian angle spectrum from which the DOAs were estimated as the directions corresponding to the peaks of this spatial spectrum (throughout the paper, we refer to a direction-dependent similarity score as a spatial spectrum) [8]. Opposed to [5, 6], the spatial spectrum was constructed from time-frequency (TF) bins where one speaker was assumed to be dominant over all other speakers, noise, and reverberation, solely.

Estimation of the RTF vector of a speaker from noisy microphone signals can be accomplished using, e.g., the state-of-the-art covariance whitening (CW) method [9] or the spatial coherence (SC) method [10]. Despite the effectiveness of the CW method and the possibility to apply the method using only

the head-mounted microphone signals or all available signals, such a computationally expensive method (due to the inherent generalized eigenvalue decomposition) is less desirable than methods with a lower computation complexity for resource-constrained applications like hearing aids. Opposed to the CW method, the SC method requires an external microphone but does not perform expensive matrix decompositions. The SC method relies on the assumption of a low spatial coherence between the undesired component in one of the microphone signals and the undesired components in the remaining microphone signals. As shown in [10], this assumption holds quite well, for example, when the distance between the external microphone and the head-mounted microphones is large enough and the undesired component is spatially diffuse-like.

In this paper, we propose to construct the frequency-averaged Hermitian angle spectrum for DOA estimation for multiple speakers using the computationally inexpensive SC method. We compare the DOA estimation accuracy when estimating the RTF vector using the SC method or the CW method in a reverberant acoustic scenario with diffuse-like babble noise. Experimental results show for multiple positions of the external microphone that estimating the RTF vector with the SC method yields a DOA estimation accuracy that is comparable to the CW method at a lower computational complexity.

2. SIGNAL MODEL AND NOTATION

We consider a binaural hearing aid setup with M microphones, i.e., $M/2$ microphones on each hearing aid, and one external microphone that is spatially separated from the head-mounted microphones and can be located at an arbitrary position, i.e., $M+1$ microphones in total. We consider an acoustic scenario with J simultaneously active speakers with DOAs $\theta_{1:J}$ (in the azimuthal plane) in a noisy and reverberant environment, where J is assumed to be known. In the short-time Fourier transform (STFT) domain, the m-th microphone signal can be written as

$$Y_m(k,l) = \sum_{j=1}^{J} X_{m,j}(k,l) + N_m(k,l), \tag{1}$$

where $m \in \{1,...,M+1\}$ denotes the microphone index, $k \in \{1,...,K\}$ and $l \in \{1,...,L\}$ denote the frequency bin index and the frame index, respectively, and $X_{m,j}(k,l)$ and $N_m(k,l)$ denote the j-th speech component and the noise component in the m-th microphone signal, respectively. For conciseness, we will omit the frequency bin index k and the frame index l in the remainder of this paper wherever possible. Assuming sparsity in the STFT domain and one dominant speaker (indexed by $j=d$) per TF bin [11], and stacking all microphone signals in an $(M+1)$-dimensional vector $\mathbf{y} = [Y_1, \dots, Y_{M+1}]^T$, where $(\cdot)^T$ denotes transposition, the vector $\mathbf{y}$ is given by

$$\mathbf{y} = \sum_{j=1}^{J} \mathbf{x}_j + \mathbf{n} \approx \mathbf{x}_d + \mathbf{n}, \tag{2}$$

with $\mathbf{x}_j$, $\mathbf{x}_d$, and $\mathbf{n}$ defined similarly as $\mathbf{y}$.

Choosing the first microphone as the reference microphone (without loss of generality) and assuming that the speech component for each (dominant) speaker can be decomposed into a direct-path component $\mathbf{x}_d^{\mathrm{DP}}$ and a reverberant component $\mathbf{x}_d^{\mathrm{R}}$, $\mathbf{x}_d$ can be written as

$$\mathbf{x}_d = \mathbf{x}_d^{\mathrm{DP}} + \mathbf{x}_d^{\mathrm{R}} = \mathbf{g}_d X_{1,d}^{\mathrm{DP}} + \mathbf{x}_d^{\mathrm{R}}, \tag{3}$$

where

$$\mathbf{g}_d = [1, G_2, \dots, G_{M+1}]^T \tag{4}$$

denotes the extended $(M+1)$-dimensional direct-path RTF vector and $X_{1,d}^{\mathrm{DP}}$ denotes the direct-path speech component of the dominant speaker in the reference microphone. The M-dimensional head-mounted direct-path RTF vector $\mathbf{g}_{\mathrm{H}_d}$ corresponding to the head-mounted microphone signals can be extracted from $\mathbf{g}_d$ as

$$\mathbf{g}_{\mathrm{H}_d} = \mathbf{E}_{\mathrm{H}} \mathbf{g}_d, \quad \mathbf{E}_{\mathrm{H}} = [\mathbf{I}_{M\times M}, \mathbf{0}_M], \tag{5}$$

where $\mathbf{E}_{\mathrm{H}}$ denotes the $(M \times M+1)$-dimensional selection matrix for the head-mounted microphone signals with $\mathbf{I}_{M\times M}$ denoting an $(M \times M)$-dimensional identity matrix and $\mathbf{0}_M$ denoting an M-dimensional vector of zeros. Both RTF vectors $\mathbf{g}_d$ and $\mathbf{g}_{\mathrm{H}_d}$ encode the DOA of the dominant speaker. However, the extended RTF vector $\mathbf{g}_d$ depends on the (unknown) position of the external microphone, whereas the head-mounted RTF vector $\mathbf{g}_{\mathrm{H}_d}$ with fixed relative positions of the head-mounted microphones (ignoring small movements of the hearing aids due to head movements) does not depend on the position of the external microphone. Hence, for DOA estimation, we will only consider the head-mounted RTF vector $\mathbf{g}_{\mathrm{H}_d}$.

The noise and reverberation components are condensed into the undesired component $\mathbf{u}_d = \mathbf{x}_d^{\mathrm{R}} + \mathbf{n}$ such that $\mathbf{y} \approx \mathbf{x}_d^{\mathrm{DP}} + \mathbf{u}_d$.

Assuming uncorrelated direct-path speech and undesired components, the covariance matrix of the noisy microphone signals can be written as

$$\mathbf{\Phi}_{\mathrm{y}} = \mathcal{E}\{\mathbf{y}\mathbf{y}^H\} = \mathbf{\Phi}_{\mathrm{x}_d}^{\mathrm{DP}} + \mathbf{\Phi}_{\mathrm{u}}, \tag{6}$$

with

$$\mathbf{\Phi}_{\mathrm{x}_d}^{\mathrm{DP}} = \mathbf{g}_d \mathbf{g}_d^H \Phi_{X_d}^{\mathrm{DP}}, \quad \mathbf{\Phi}_{\mathrm{u}} = \mathcal{E}\{\mathbf{u}_d \mathbf{u}_d^H\}, \tag{7}$$

where $(\cdot)^H$ and $\mathcal{E}\{\cdot\}$ denote the complex transposition and expectation operator, respectively. $\mathbf{\Phi}_{\mathrm{x}_d}^{\mathrm{DP}}$ and $\mathbf{\Phi}_{\mathrm{u}}$ denote the covariance matrices of the direct-path dominant speech component and undesired component, respectively, and $\Phi_{X_d}^{\mathrm{DP}} = \mathcal{E}\left\{|X_{1,d}^{\mathrm{DP}}|^2\right\}$ denotes the power spectral density of the direct-path dominant speech component in the reference microphone.

3. RTF-VECTOR-BASED DOA ESTIMATION

In this section, we review the RTF-vector-based DOA estimation method proposed in [8] that is based on finding the directions corresponding to the peaks of the spatial spectrum called frequency-averaged Hermitian angle spectrum.

To estimate the DOAs $\theta_{1:J}$ of the speakers from the estimated head-mounted [1] RTF vector $\hat{\mathbf{g}}_{\mathrm{H}_d}(k,l)$, the estimated head-mounted RTF vector $\hat{\mathbf{g}}_{\mathrm{H}_d}(k,l)$ is compared to a database of prototype anechoic RTF vectors $\bar{\mathbf{g}}(k,\theta_i)$ for several directions $\theta_i, i = 1,\ldots,I$ using the Hermitian angle [12] as a measure of dissimilarity, i.e.,

$$p(k,l,\theta_i) = h(\hat{\mathbf{g}}_{\mathrm{H}_d}(k,l), \bar{\mathbf{g}}(k,\theta_i)), \tag{8}$$

$$h(\hat{\mathbf{g}},\bar{\mathbf{g}}) = \arccos\left(\frac{|\bar{\mathbf{g}}^H \hat{\mathbf{g}}|}{\|\bar{\mathbf{g}}\|_2 \|\hat{\mathbf{g}}\|_2}\right). \tag{9}$$

These prototype anechoic head-mounted RTF vectors can be obtained, e.g., via measurements using the same microphone array configuration as used during the actual source localization or using spherical diffraction models [13].

Accounting for the disjoint activity of the speakers in the STFT domain and aiming at including only TF bins where the estimated head-mounted RTF vector $\hat{\mathbf{g}}_{\mathrm{H}_d}(k,l)$ is a good estimate for the direct-path RTF vector in (5) (of one of the speakers), the narrowband spatial spectrum (8) is integrated over a set $\mathcal{K}(l)$ of selected frequency bins, where it is likely that one speaker dominates over all other speakers, noise, and reverberation [8], i.e.,

$$P(l,\theta_i) = -\sum_{k\in\mathcal{K}(l)} p(k,l,\theta_i). \tag{10}$$

Based on the usage of the Hermitian angle for the construction of (8), the spatial spectrum in (10) is called the frequency-averaged Hermitian angle spectrum. The DOAs $\theta_{1:J}(l)$ are estimated by selecting the directions corresponding to the J peaks of this spatial spectrum (assuming J to be known).

In the context of DOA estimation, coherence-based quantities such as the coherent-to-diffuse ratio (CDR) are a common criterion for frequency subset selection [8, 14–17]. The usage of the CDR as a criterion for frequency subset selection can be motivated by the fact, that for higher values of the CDR at the respective TF bin it is more likely that a speaker dominates over all other speakers, noise, and reverberation at the respective TF bin. As in [8], the subset $\mathcal{K}(l)$ is obtained using the coherent-to-diffuse ratio (CDR) criterion (11), i.e.,

$$\mathcal{K}(l) = \left\{k : \widehat{\mathrm{CDR}}(k,l) \geq \mathrm{CDR}_{\mathrm{thresh}}\right\}, \tag{11}$$

where the CDR is estimated as

$$\widehat{\mathrm{CDR}}(k,l) = f\left(\widehat{\Gamma}_{y,\mathrm{eff}}(k,l), \widetilde{\Gamma}_u(k)\right), \tag{12}$$

with the CDR-functional f defined in (22) for a single microphone pair comprising the microphones $m = i$ and $m = j$ [18]. The arguments of the function in (22) are the estimated coherence $\widehat{\Gamma}_{\mathrm{y},i,j}$ of the noisy signal

$$\widehat{\Gamma}_{y_{i,j}}(k,l) = \hat{\Phi}_{\mathrm{y}_{i,j}}(k,l) / \sqrt{\hat{\Phi}_{\mathrm{y}_{i,i}}(k,l)\, \hat{\Phi}_{\mathrm{y}_{j,j}}(k,l)} \tag{13}$$

with $\hat{\Phi}_{\mathrm{y}_{i,j}}$ denoting an estimate of the (i,j)-th element of the covariance matrix of the noisy microphone signals and a model $\widetilde{\Gamma}_{\mathrm{u},i,j}$ of the coherence of the undesired component. To consider more than just a single microphone pair for the estimation of the CDR, the coherence of the noisy signals between multiple microphone pairs (denoted as the microphone set $\mathcal{M}$) between the left and the right hearing aid is averaged prior to evaluating the CDR-functional in (22), resulting in the binaural effective coherence [8, 19], i.e.,

$$\widehat{\Gamma}_{y,\mathrm{eff}}(k,l) = \frac{1}{|\mathcal{M}|} \sum_{i,j\in\mathcal{M}} \widehat{\Gamma}_{y_{i,j}}(k,l), \tag{14}$$

Thus, the binaural effective coherence represents the average coherence between the head-mounted microphone signals. Due to the arbitrary position of the external microphone, we consider only the head-mounted microphones (with fixed relative positions) for the estimation of the binaural effective coherence $\widehat{\Gamma}_{y,\mathrm{eff}}(k,l)$.

To model the coherence of the undesired component for the estimation of the CDR in (22) between the head-mounted microphone signals, head shadow effects need to be included. Assuming a diffuse sound field for both the noise and

[1] As previously stated, we only consider the estimated head-mounted RTF vector $\hat{\mathbf{g}}_{\mathrm{H}_d}(k,l)$ for DOA estimation and not the extended RTF vector $\hat{\mathbf{g}}_d(k,l)$ that depends both on the speaker DOA and the (unknown) position of the external microphone.

reverberation component, a modified sinc-model [20] is employed, i.e.,

$$\widetilde{\Gamma}_u(k) = \mathrm{sinc}\left(\alpha\frac{\omega_k r}{c}\right)\frac{1}{\sqrt{1+\left(\beta\frac{\omega_k r}{c}\right)^4}}, \tag{15}$$

where ω_k denotes the discrete angular frequency, r denotes the distance between the microphones of left and right hearing aid which is approximated as the diameter of a head, c denotes the speed of sound, and $\alpha = 0.5$ and $\beta = 2.2$ denote empirically determined parameters of the modified sinc-model.

In this paper we compare the influence of different RTF vector estimation methods on constructing the frequency-averaged Hermitian angle spectrum in (10). In [8] no external microphone was used and therefore the DOAs were estimated from the spatial spectrum as in (16) constructed from head-mounted RTF vectors that were estimated using the CW method as in (18), i.e.,

$$P^{(\mathrm{CW})}(l,\theta_i) = -\sum_{k\in\mathcal{K}(l)} h\left(\hat{\mathbf{g}}_{\mathrm{H}_d}^{(\mathrm{CW})}(k,l), \bar{\mathbf{g}}(k,\theta_i)\right). \tag{16}$$

In this paper, we propose to exploit the availability of the external microphone and estimate the DOAs from the spatial spectrum constructed as in (17) constructed from head-mounted RTF vectors that are estimated using the SC method as in (21), i.e.,

$$\boxed{P^{(\mathrm{SC})}(l,\theta_i) = -\sum_{k\in\mathcal{K}(l)} h\left(\hat{\mathbf{g}}_{\mathrm{H}_d}^{(\mathrm{SC})}(k,l), \bar{\mathbf{g}}(k,\theta_i)\right)} \tag{17}$$

A summary on the covariance whitening (CW) method [9] and the spatial coherence (SC) method [10] is provided in the next section.

4. RTF VECTOR ESTIMATION

In order to estimate DOAs of multiple speakers, a frequency-averaged Hermitian angle spectrum is constructed, which assess the similarity between the estimated M-dimensional head-mounted RTF vector $\hat{\mathbf{g}}_{\mathrm{H}_d}(k,l)$ and a database of prototype anechoic RTF vectors for different directions. In this section, we review two RTF vector estimation methods. The computationally expensive state-of-the-art covariance whitening (CW) method [9] is summarized in Section 4.1. The computationally inexpensive spatial coherence (SC) method [10] is discussed in Section 4.2.

4.1 Covariance whitening (CW)

To apply the CW method [9], estimates $\hat{\mathbf{\Phi}}_{\mathrm{y}}$ and $\hat{\mathbf{\Phi}}_{\mathrm{u}}$ of the covariance matrices of the noisy signal and the undesired signal component are required. Based on these estimates, the head-mounted direct-path RTF vector $\mathbf{g}_{\mathrm{H}_d}$ can be estimated using only the head-mounted microphone signals as

$$\hat{\mathbf{g}}_{\mathrm{H}_d}^{(\mathrm{CW})} = f\left(\mathbf{E}_{\mathrm{H}}\hat{\mathbf{\Phi}}_{\mathrm{y}}\mathbf{E}_{\mathrm{H}}^H, \mathbf{E}_{\mathrm{H}}\hat{\mathbf{\Phi}}_{\mathrm{u}}\mathbf{E}_{\mathrm{H}}^H\right), \tag{18}$$

$$f\left(\check{\mathbf{\Phi}}_{\mathrm{y}}, \check{\mathbf{\Phi}}_{\mathrm{u}}\right) = \frac{\check{\mathbf{\Phi}}_{\mathrm{u}}^{1/2}\mathcal{P}\left\{\check{\mathbf{\Phi}}_{\mathrm{u}}^{-1/2}\check{\mathbf{\Phi}}_{\mathrm{y}}\check{\mathbf{\Phi}}_{\mathrm{u}}^{-H/2}\right\}}{\check{\mathbf{e}}_1^T\check{\mathbf{\Phi}}_{\mathrm{u}}^{1/2}\mathcal{P}\left\{\check{\mathbf{\Phi}}_{\mathrm{u}}^{-1/2}\check{\mathbf{\Phi}}_{\mathrm{y}}\check{\mathbf{\Phi}}_{\mathrm{u}}^{-H/2}\right\}}, \tag{19}$$

where $\mathcal{P}\{\cdot\}$ denotes the principal eigenvector of a matrix, $\check{\mathbf{\Phi}}_{\mathrm{u}}^{1/2}$ denotes a square-root decomposition (e.g., Cholesky decomposition) of the $\check{M}$-dimensional matrix $\check{\mathbf{\Phi}}_{\mathrm{u}}$ and $\check{\mathbf{e}}_1 = [1,0,\ldots,0]^T$ denotes an $\check{M}$-dimensional selection vector. Note that $\mathbf{g}_{\mathrm{H}_d}$ can be estimated likewise from the head-mounted microphone signals *and* the external microphone signal together, via $\mathbf{E}_{\mathrm{H}} f\left(\hat{\mathbf{\Phi}}_{\mathrm{y}}, \hat{\mathbf{\Phi}}_{\mathrm{u}}\right)$, differing in general from the estimate $\hat{\mathbf{g}}_{\mathrm{H}_d}^{(\mathrm{CW})}$ as in (18). However, based on the results of [5] and [6], we will consider only the estimate as in (18) obtained from the head-mounted microphone signals only as no significant benefit in DOA estimation performance was reported when all microphone signals were used.

4.2 Spatial coherence (SC)

The SC method [10] requires an external microphone and relies on the assumption of a low spatial coherence between the undesired component U_{M+1} in the external microphone signal and the undesired components $U_m, m \in \{1,\ldots,M\}$, in the head-mounted microphone signals, i.e.

$$\mathcal{E}\left\{U_m U_{M+1}^*\right\} \approx 0, \quad m \in \{1,\ldots,M\}. \tag{20}$$

As shown in [10], this assumption holds quite well, for example, when the distance between the external microphone and the head-mounted microphones is large enough and the undesired component is spatially diffuse-like. Exploiting this assumption, results in $\mathcal{E}\left\{Y_m Y_{M+1}^*\right\} = \mathcal{E}\left\{X_m X_{M+1}^*\right\}, m \in \{1,\ldots,M\}$, thus the RTF vector can be efficiently estimated without expensive matrix decompositions as

$$\hat{\mathbf{g}}_{\mathrm{H}_d}^{(\mathrm{SC})} = \mathbf{E}_{\mathrm{H}}\frac{\hat{\mathbf{\Phi}}_{\mathrm{y}}\mathbf{e}_{M+1}}{\mathbf{e}_1^T\hat{\mathbf{\Phi}}_{\mathrm{y}}\mathbf{e}_{M+1}}, \tag{21}$$

$$f\left(\widehat{\Gamma}_{\mathrm{y},i,j}, \widetilde{\Gamma}_{\mathrm{u},i,j}\right) = \frac{\widetilde{\Gamma}_{\mathrm{u},i,j}\Re\{\widehat{\Gamma}_{\mathrm{y},i,j}\} - |\widehat{\Gamma}_{\mathrm{y},i,j}|^2 - \sqrt{\widetilde{\Gamma}_{\mathrm{u},i,j}^2\Re\{\widehat{\Gamma}_{\mathrm{y},i,j}\}^2 - \widetilde{\Gamma}_{\mathrm{u},i,j}^2|\widehat{\Gamma}_{\mathrm{y},i,j}|^2 + \widetilde{\Gamma}_{\mathrm{u},i,j}^2 - 2\widetilde{\Gamma}_{\mathrm{u},i,j}\Re\{\widehat{\Gamma}_{\mathrm{y},i,j}\} + |\widehat{\Gamma}_{\mathrm{y},i,j}|^2}}{|\widehat{\Gamma}_{\mathrm{y},i,j}|^2 - 1} \tag{22}$$

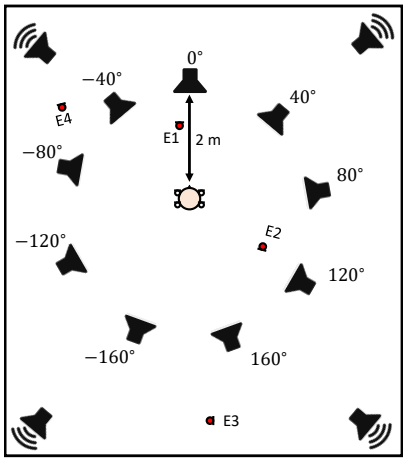

Figure 1. Experimental setup with a head-mounted binaural hearing setup and an external microphone depicted in red at four different positions (E1-E4).

with $\mathbf{e}_m$ denoting an $(M+1)$-dimensional selection vector selecting the m-th element.

5. EXPERIMENTAL RESULTS

Applying the CW and SC method for RTF vector estimation, in this section we compare the DOA estimation performance when using the SC-based frequency-averaged Hermitian angle spectrum as in (17) against the DOA estimation performance when using the CW-based frequency-averaged Hermitian angle spectrum as in (16). We evaluate the methods with recorded signals for an acoustic scenario with two static speakers in a reverberant room with diffuse-like babble noise. The experimental setup and implementation details of the algorithms are described in Section 5.1. The results in terms of localization accuracy are presented and discussed in Section 5.2.

5.1 Experimental setup and implementation details

For the experiments we used signals that were recorded in a laboratory at the University of Oldenburg with dimensions of about $7\times6\times2.7\,\mathrm{m}^3$, where the reverberation time can be adjusted by means of absorber panels, which are mounted to the walls and the ceiling. The reverberation time was set to approximately $T_{60}\approx250\mathrm{ms}$. Fig. 1 depicts the experimental setup. A dummy head with a binaural hearing aid setup ($M=4$) was placed approximately in the center of the laboratory. For this hearing aid setup a database of prototype anechoic RTF vectors is obtained from measured anechoic binaural room impulse responses [21] with an angular resolution of $5°$ ($I=72$). A single external microphone was placed at four different positions (denoted as E1 - E4), which was not restricted to be close to a speaker. Two speakers from the EBU SQAM CD corpus [22] (male and female, English language) were played back via loudspeakers that were located at approximately $2\,\mathrm{m}$ distance from the dummy head. For the evaluation, all 72 pairs of DOAs of non-collocated speakers (each of the 9 DOAs in the range $[-160°,-120°,...,160°]$) were considered. The speech signals were constantly active and had a duration of approximately $5\,\mathrm{s}$. Diffuse-like noise was generated with four loudspeakers facing the corners of the laboratory, playing back different multi-talker recordings. The speech and noise components were recorded separately and were mixed at $\{-5\mathrm{dB},0\mathrm{dB},5\mathrm{dB}\}$ broadband signal-to-noise ratio (SNR) averaged over all head-mounted microphones of the hearing aid setup. All microphone signals were recorded simultaneously, hence neglecting synchronization and latency aspects.

The microphone signals were processed in the STFT-domain using a $32\,\mathrm{ms}$ square-root Hann window with 50 % overlap at a sampling frequency of $16\,\mathrm{kHz}$. The covariance matrices $\mathbf{\Phi}_\mathrm{y}$ and $\mathbf{\Phi}_\mathrm{u}$ were estimated recursively during detected speech-and-noise and noise-only TF bins, respectively, using smoothing factors corresponding to time constants of $250\,\mathrm{ms}$ for $\hat{\mathbf{\Phi}}_\mathrm{y}$ and $500\,\mathrm{ms}$ for $\hat{\mathbf{\Phi}}_\mathrm{u}$, respectively. The speech-and-noise TF bins were discriminated from noise-only TF bins based on the speech presence probability [23], averaged and thresholded over all head-mounted microphone signals.

We assess the DOA estimation performance by averaging the localization accuracy over the considered DOA pairs and SNRs. For the localization accuracy we average the per-frame-accuracies over all frames, where we define the per-frame accuracy as

$$\mathrm{ACC}(l)=j_\mathrm{correct}(l)/J, \tag{23}$$

with $j_\mathrm{correct}\,(l)$ denoting the number of speakers that are correctly localized within a range of $\pm5°$ in the l-th frame and $J=2$.

5.2 Results

Fig. 2 depicts the average localization accuracies that are obtained from the spatial spectrum as in (16), denoted by "CW", and the accuracies obtained from the spatial spectrum as in (17), denoted by "SC-EX", where "X" stands for one of the four positions of the external microphone. To show the effectiveness of the subset selection, we considered two

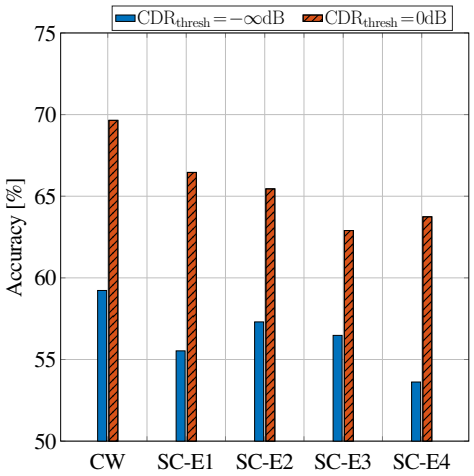

Figure 2. Average localization accuracy without (unhatched blue) and with (hatched orange) frequency subset selection using an external microphone placed at one of four different positions (SC-E1 – SC-E4) or not using an external microphone (CW) for the construction of the spatial spectrum.

threshold values, $\text{CDR}_{\text{thresh}} = -\infty\,\text{dB}$ (corresponding to selecting all frequencies) and $\text{CDR}_{\text{thresh}} = 0\text{dB}$, shown as unhatched blue bars and hatched orange bars, respectively.

First, for every condition a large improvement in the localization accuracy of up to 11 % due to the frequency subset selection can be observed. This result is in line with the results reported in [8]. Second, considering the spatial spectrum obtained from (17), it can be observed that the position of the external microphone has a minor effect on the estimated DOA, resulting in localization accuracies in the range 62 % - 66 % using a threshold value of $\text{CDR}_{\text{thresh}} = 0\text{dB}$. For the external microphone placed at positions E3 or E4, i.e., close to the loudspeakers playing back the noise, a slightly lower DOA estimation accuracy can be observed when comparing to the external microphone placed at positions E1 or E2. Third, comparing the DOA estimation performance when using the CW method against the SC method for estimating the head-mounted RTF vector, a difference up to around 5 % - 7 % can be observed. Thus, the low-complexity SC method yields a comparable DOA estimation performance for multiple speakers as the CW method, which is line with the single speaker DOA estimation results reported in [5].

6. CONCLUSIONS

Based on two RTF vector estimation methods, in this paper we compared the DOA estimation performance for multiple speakers for a binaural hearing aid setup exploiting an external microphone or not. We did not restrict the position of the external microphone to be close to the target speaker. Estimating the RTF vector using either the CW method without exploiting the external microphone or using the SC method exploiting the external microphone, we constructed a frequency-averaged Hermitian angle spectrum from which the DOAs of the speakers were estimated as the directions that maximized the spatial spectrum. We evaluated the approach using simulations with recorded two speaker scenarios in acoustic environments with mild reverberation and diffuse-like babble noise scaled to low SNRs for different positions of the external microphone. The results show that using the SC method for the construction of the frequency-averaged Hermitian angle spectrum yields a DOA estimation accuracy (62 % - 66 %) that is comparable to the CW method (≈70%) at a lower computational complexity.

7. REFERENCES

[1] Y. Huang, J. Benesty, and J. Chen, "Time delay estimation and source localization," in *Springer Handbook of Speech Processing* (J. Benesty, M. M. Sondhi, and Y. Huang, eds.), pp. 1043–1063, Berlin, Heidelberg, Germany: Springer, 2008.

[2] P.-A. Grumiaux, S. Kitic, L. Girin, and A. Guerin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, pp. 107–151, Jul. 2022.

[3] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Bias-compensated informed sound source localization using relative transfer functions," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, pp. 1275–1289, Jul. 2018.

[4] U. Kowalk, S. Doclo, and J. Bitzer, "Signal-informed DNN-based DOA estimation combining an external microphone and GCC-PHAT features," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, (Bamberg, Germany), pp. 1–5, Sep. 2022.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2177/1 - Project ID 390895286 and Project ID 352015383 - SFB 1330 B2.

[5] D. Fejgin and S. Doclo, "Comparison of binaural RTF-vector-based direction of arrival estimation methods exploiting an external microphone," in *Proc. European Signal Processing Conference (EUSIPCO)*, (Dublin, Ireland), pp. 241–245, Aug. 2021.

[6] D. Fejgin and S. Doclo, "Assisted RTF-vector-based binaural direction of arrival estimation exploiting a calibrated external microphone array," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Rhodes, Greece), Jun. 2023.

[7] J. Mecklenburger and T. Groth, "Wireless technologies and hearing aid connectivity," in *Hearing Aids* (G. R. Popelka, B. C. J. Moore, R. R. Fay, and A. N. Popper, eds.), pp. 131–149, Cham, Switzerland: Springer, 2016.

[8] D. Fejgin and S. Doclo, "Coherence-based frequency subset selection for binaural RTF-vector-based direction of arrival estimation for multiple speakers," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, (Bamberg, Germany), pp. 1–5, Sep. 2022.

[9] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, pp. 1071–1086, Aug. 2009.

[10] 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), pp. 146–150, Sep. 2018.

[11] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, pp. 1830–1847, Jul. 2004.

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

[13] 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, p. 3048–3058, Nov. 1998.

[14] M. Taseska and E. A. P. Habets, "DOA-informed source extraction in the presence of competing talkers and background noise," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, pp. 1–13, Aug. 2017.

[15] A. Brendel, C. Huang, and W. Kellermann, "STFT bin selection for localization algorithms based on the sparsity of speech signal spectra," in *Proc. Euronoise*, (Crete, Greece), pp. 2561–2568, May 2018.

[16] C. Evers, E. A. P. Habets, S. Gannot, and P. A. Naylor, "DoA reliability for distributed acoustic tracking," *IEEE Signal Processing Letters*, vol. 25, pp. 1320–1324, Sep. 2018.

[17] R. Lee, M.-S. Kang, B.-H. Kim, K.-H. Park, S. Q. Lee, and H.-M. Park, "Sound source localization based on GCC-PHAT with diffuseness mask in noisy and reverberant environments," *IEEE Access*, vol. 8, pp. 7373–7382, Jan. 2020.

[18] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, pp. 1006–1018, Jun. 2015.

[19] H. W. Löllmann, A. Brendel, and W. Kellermann, "Generalized coherence-based signal enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Barcelona, Spain), pp. 201–205, May 2020.

[20] I. M. Lindevald and A. H. Benade, "Two-ear correlation in the statistical sound fields of rooms," *The Journal of the Acoustical Society of America*, vol. 80, pp. 661–664, Aug. 1986.

[21] 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, Jul. 2009.

[22] European Broadcasting Union, "Sound quality assessment material - recordings for subjective tests: User's handbook for the EBU SQUAM CD," 2008. [Online]. Available: https://tech.ebu.ch/publications/sqamcd.

[23] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, pp. 1383–1393, May 2012.