ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095634

ASSISTED RTF-VECTOR-BASED BINAURAL DIRECTION OF ARRIVAL ESTIMATION EXPLOITING A CALIBRATED EXTERNAL MICROPHONE ARRAY

Daniel Fejgin and Simon Doclo

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

ABSTRACT

Recently, a relative transfer function (RTF) vector-based method has been proposed to estimate the direction of arrival (DOA) of a target speaker for a binaural hearing aid setup, assuming the availability of external microphones. This method exploits the external microphones to estimate the RTF vector corresponding to the binaural hearing aid and constructs a one-dimensional spatial spectrum by comparing the estimated RTF vector against a database of anechoic prototype RTF vectors for several directions. In this paper, we assume the availability of a calibrated array of external microphones, which is characterized by a second database of anechoic prototype RTF vectors. We propose a method where the external microphones are not only exploited for RTF vector estimation but also assist in estimating the DOA of the target speaker. Based on the estimated RTF vector for all microphones and the prototype RTF databases of the binaural hearing aid and the external microphone array, a two-dimensional spatial spectrum is constructed from which the DOA is estimated. Experimental results for a reverberant environment with diffuse-like noise show that assisted DOA estimation outperforms DOA estimation where the prototype database characterizing the external microphone array is not used.

***Index Terms*—** direction of arrival estimation, binaural hearing aids, assisted localization, external microphones

1. INTRODUCTION

With the advent of mobile devices that are equipped with microphones, wirelessly linking hearing aids to these devices has become increasingly popular [1]. In such an acoustic sensor network, jointly processing all available microphones, i.e. the hearing aid microphones in conjunction with the external microphones, has been shown to be beneficial for, e.g., noise reduction [2–6] as well as for direction of arrival (DOA) estimation [7–9].

Large research effort has already been dedicated to DOA estimation [10–14], in particular also for binaural hearing aid applications [7–9, 15–17]. In [7] we proposed a DOA estimation method for a binaural hearing aid setup, which exploits an external microphone at an unknown position. In this method all available microphone signals are used to estimate the so-called relative transfer function (RTF) vector between all microphones and a reference microphone on one of the hearing aids. The estimated RTF vector corresponding to only the hearing aid microphones is then compared against a database of anechoic prototype RTF vectors (e.g., obtained through measurement), by constructing a one-dimensional spatial spectrum for different directions. It should be noted that the element in the estimated RTF vector corresponding to the external microphone cannot be used in this comparison, since the position of the external microphone is generally unknown. The DOA of the target speaker relative to the binaural hearing aid setup is then estimated as the direction for which the spatial spectrum is maximized. Contrary to [8], in [7] the external microphone does not need to be in the vicinity of the target speaker in order to capture a nearly clean speech reference, while contrary to [9] no pre-trained representation of clean speech spectral envelopes is required for DOA estimation.

Whereas the methods in [7, 8] exploit a single external microphone, in this paper we assume the availability of a calibrated array of external microphones. With calibrated array we mean that (similarly as for the binaural hearing aid) a database of anechoic prototype RTF vectors for several directions is available for this array. The prototype RTF vectors can be obtained, e.g., through measurement or when the geometry of the external microphone array is known (and assuming free-field sound propagation). It should be noted that the relative position of the external microphone array with respect to the binaural hearing aid setup is obviously unknown.

We propose a method where the calibrated array of external microphones assists in estimating the DOA of the target speaker. Similarly as in [7], the RTF vector between all available microphones and a reference microphone on one of the hearing aids is first estimated using the state-of-the-art covariance whitening method [18]. Using the entire RTF vector and the prototype databases of both arrays, i.e. the binaural hearing aid as well as the external microphone array, we now define a two-dimensional spatial spectrum. The DOA of the target speaker is then estimated by determining the location of the main peak of this two-dimensional spatial spectrum. To further investigate the potential of the proposed method, we also consider a special case, by transforming the two-dimensional spatial spectrum into a one-dimensional spatial spectrum using a coordinate system transformation which requires prior information. The performance of the proposed binaural DOA estimation method is evaluated using recordings for a single static speaker in a reverberant acoustic scenario with diffuse-like noise. Experimental results show the benefit of exploiting a calibrated external microphone array compared to exploiting an uncalibrated external microphone array or using only the hearing aid microphones for DOA estimation.

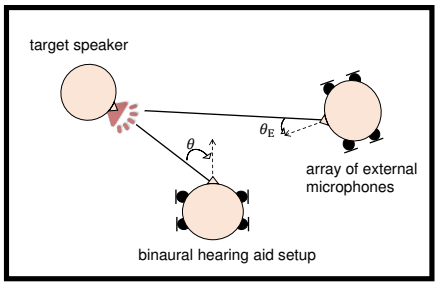

Fig. 1. Considered acoustic scenario with one target speaker, a binaural hearing setup and an external microphone array, which in this work corresponds to the same binaural hearing aid setup.

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.

2. SIGNAL MODEL AND NOTATION

We consider an acoustic sensor network composed of two microphone arrays with a total of $M = M_{\mathrm{H}} + M_{\mathrm{E}}$ microphones (see Fig. 1): a binaural hearing aid setup consisting of M_{H} head-mounted microphones (i.e., $M_{\mathrm{H}}/2$ microphones on each hearing aid) and an array consisting of M_{E} external microphones, which is assumed to be spatially separated from the binaural hearing aid setup. We consider a noisy and reverberant acoustic scenario with a single speaker that is located at DOA θ relative to the local coordinate system of the binaural hearing setup and angle θ_{E} relative to the local coordinate system of the external microphone array (in the azimuthal plane).

In the short-time Fourier transform (STFT) domain, the m-th microphone signal can be written as

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

where $X_m(k,l)$ and $N_m(k,l)$ denote the speech and noise component in the m-th microphone signal, respectively, and $k \in \{1,...,K\}$ and $l \in \{1,...,L\}$ denote the frequency bin index and the frame index, respectively. Since all frequency bins are assumed to be independent and are hence processed independently, the index k will be omitted in the remainder of the paper where possible. Stacking all microphone signals into the vector $\mathbf{y}(l) = [Y_1(l),...,Y_{M_{\mathrm{H}}}(l),...,Y_M(l)]^T \in \mathbb{C}^M$, where $(\cdot)^T$ denotes transposition, the noisy microphone signals can be written as $\mathbf{y}(l) = \mathbf{x}(l) + \mathbf{n}(l)$, where the speech vector $\mathbf{x}(l)$ and the noise vector $\mathbf{n}(l)$ are defined similarly as $\mathbf{y}(l)$.

Assuming that the speech vector can be split into a direct-path component $\mathbf{x}^{\mathrm{DP}}(l)$ and a reverberant component $\mathbf{x}^{\mathrm{R}}(l)$ and assuming that the multiplicative transfer function approximation [19] holds for the direct-path component, the speech vector $\mathbf{x}(l)$ can be written as

$$\mathbf{x}(l) = \mathbf{x}^{\mathrm{DP}}(l) + \mathbf{x}^{\mathrm{R}}(l) = \begin{bmatrix} \mathbf{a}_{\mathrm{H}}(\theta) \\ \mathbf{a}_{\mathrm{E}}(\theta_{\mathrm{E}}) \end{bmatrix} S(l) + \mathbf{x}^{\mathrm{R}}(l), \tag{2}$$

where the M_{H}-dimensional vector $\mathbf{a}_{\mathrm{H}}(\theta)$ and the M_{E}-dimensional vector $\mathbf{a}_{\mathrm{E}}(\theta_{\mathrm{E}})$ denote the direct-path acoustic transfer function vectors between the speaker S and the hearing aid microphones and the external microphones, respectively. By introducing the direct-path RTF vectors $\mathbf{g}_{\mathrm{H}}(\theta) = \mathbf{a}_{\mathrm{H}}(\theta)/A_1(\theta)$ and $\mathbf{g}_{\mathrm{E}}(\theta_{\mathrm{E}}) = \mathbf{a}_{\mathrm{E}}(\theta_{\mathrm{E}})/A_{\mathrm{E},1}(\theta_{\mathrm{E}})$, where the first microphone of each array is chosen as reference microphone without loss of generality, we can rewrite the direct-path speech component in (2) as

$$\mathbf{x}^{\mathrm{DP}}(l) = \begin{bmatrix} A_1(\theta)\mathbf{g}_{\mathrm{H}}(\theta) \\ A_{\mathrm{E},1}(\theta_{\mathrm{E}})\mathbf{g}_{\mathrm{E}}(\theta_{\mathrm{E}}) \end{bmatrix} S(l) = \mathbf{g} X_1^{\mathrm{DP}}(l), \tag{3}$$

where the M-dimensional vector $\mathbf{g}$ denotes the direct-path RTF vector between all microphones and the reference microphone of the hearing aid and $X_1^{\mathrm{DP}}(l)$ denotes the direct-path speech component in the reference microphone of the hearing aid. Condensing the noise and reverberation components into the undesired component $\mathbf{u}(l) = \mathbf{n}(l) + \mathbf{x}^{\mathrm{R}}(l)$, the vector of noisy microphone signals can be written as $\mathbf{y}(l) = \mathbf{x}^{\mathrm{DP}}(l) + \mathbf{u}(l)$.

The RTF vectors $\mathbf{g}_{\mathrm{H}}(\theta)$ and $\mathbf{g}_{\mathrm{E}}(\theta_{\mathrm{E}})$ can be both extracted from the RTF vector $\mathbf{g}$ in (3) as

$$\boxed{\mathbf{g}_{\mathrm{H}}(\theta) = \mathbf{E}_{\mathrm{H}}\mathbf{g}, \quad \mathbf{g}_{\mathrm{E}}(\theta_{\mathrm{E}}) = \frac{\mathbf{E}_{\mathrm{E}}\mathbf{g}}{\mathbf{e}_{1,\mathrm{E}}^T \mathbf{E}_{\mathrm{E}}\mathbf{g}}} \tag{4}$$

$$\mathbf{E}_{\mathrm{H}} = [\mathbf{I}_{M_{\mathrm{H}} \times M_{\mathrm{H}}}, \mathbf{0}_{M_{\mathrm{H}} \times M_{\mathrm{E}}}], \quad \mathbf{E}_{\mathrm{E}} = [\mathbf{0}_{M_{\mathrm{E}} \times M_{\mathrm{H}}}, \mathbf{I}_{M_{\mathrm{E}} \times M_{\mathrm{E}}}], \tag{5}$$

where $\mathbf{I}_{N \times N}$ denotes an $N \times N$-dimensional identity matrix, $\mathbf{0}_{N \times N'}$ denotes an $N \times N'$ matrix of zeros, and $\mathbf{e}_{1,\mathrm{E}} = [1,0,...,0]^T$ denotes the M_{E}-dimensional selection vector. Both for the binaural hearing aid setup as well as for the external microphone array, we assume that a database of anechoic prototype RTF vector is available (referred to as calibrated array). The database for the binaural hearing aid setup is denoted as $\bar{\mathbf{g}}_{\mathrm{H}}(k,\theta_i)$, $i = 1,...,I$, while the database for the external microphone array is denoted as $\bar{\mathbf{g}}_{\mathrm{E}}(k,\theta_{\mathrm{E},j})$, $j = 1,...,J$.

Assuming that the direct-path speech component is uncorrelated with the undesired component, the $M \times M$-dimensional covariance matrix of the noisy microphone signals can be written as

$$\mathbf{\Phi}_{\mathrm{y}}(l) = \mathcal{E}\left\{\mathbf{y}(l)\mathbf{y}^H(l)\right\} = \mathbf{g}\mathbf{g}^H \Phi_{\mathrm{x}}^{\mathrm{DP}}(l) + \mathbf{\Phi}_{\mathrm{u}}(l), \tag{6}$$

where $(\cdot)^H$ and $\mathcal{E}\{\cdot\}$ denote complex transposition and expectation operators, respectively, $\Phi_{\mathrm{x}}^{\mathrm{DP}}(l) = \mathcal{E}\{|X_1^{\mathrm{DP}}(l)|^2\}$ denotes the power spectral density of the direct-path speech component in the reference microphone, and $\mathbf{\Phi}_{\mathrm{u}}(l) = \mathcal{E}\{\mathbf{u}(l)\mathbf{u}^H(l)\}$ denotes the covariance matrix of the undesired component. The $M_{\mathrm{H}} \times M_{\mathrm{H}}$-dimensional covariance matrices $\mathbf{\Phi}_{\mathrm{y,H}}(l)$ and $\mathbf{\Phi}_{\mathrm{u,H}}(l)$ corresponding to the noisy microphone signals and the undesired components in the hearing aid microphones can be extracted from (6) as

$$\mathbf{\Phi}_{\mathrm{y,H}}(l) = \mathbf{E}_{\mathrm{H}}\mathbf{\Phi}_{\mathrm{y}}(l)\mathbf{E}_{\mathrm{H}}^T, \quad \mathbf{\Phi}_{\mathrm{u,H}}(l) = \mathbf{E}_{\mathrm{H}}\mathbf{\Phi}_{\mathrm{u}}(l)\mathbf{E}_{\mathrm{H}}^T. \tag{7}$$

3. RTF-VECTOR-BASED DOA ESTIMATION

To estimate the DOA θ of the target speaker relative to the binaural hearing setup, in this section we propose an extension of the RTF-vector-based DOA estimation method from [7]. In Section 3.1 we review the existing method from [7], where either no external microphone array or an uncalibrated external microphone array is used. In Section 3.2 we propose a method to jointly use both calibrated arrays, i.e. the binaural hearing aid setup and the external microphone array for DOA estimation.

3.1. Baseline RTF-vector-based DOA estimation

To estimate the M_{H}-dimensional RTF vector $\mathbf{g}_{\mathrm{H}}$ corresponding to the binaural hearing aid setup, we use the state-of-the-art covariance whitening (CW) method [18] in each time-frequency bin. This RTF vector can be estimated from only the hearing aid microphone signals as

$$\hat{\mathbf{g}}_{\mathrm{H}}^{(\mathrm{CW})}(l) = f\left(\hat{\mathbf{\Phi}}_{\mathrm{y,H}}(l), \hat{\mathbf{\Phi}}_{\mathrm{u,H}}(l)\right), \tag{8}$$

$$f\left(\tilde{\mathbf{\Phi}}_{\mathrm{y}}, \tilde{\mathbf{\Phi}}_{\mathrm{u}}\right) = \frac{\tilde{\mathbf{\Phi}}_{\mathrm{u}}^{1/2} \mathcal{P}\left\{\tilde{\mathbf{\Phi}}_{\mathrm{u}}^{-1/2} \tilde{\mathbf{\Phi}}_{\mathrm{y}} \tilde{\mathbf{\Phi}}_{\mathrm{u}}^{-H/2}\right\}}{\tilde{\mathbf{e}}_1^T \tilde{\mathbf{\Phi}}_{\mathrm{u}}^{1/2} \mathcal{P}\left\{\tilde{\mathbf{\Phi}}_{\mathrm{u}}^{-1/2} \tilde{\mathbf{\Phi}}_{\mathrm{y}} \tilde{\mathbf{\Phi}}_{\mathrm{u}}^{-H/2}\right\}}, \tag{9}$$

where $\mathcal{P}\{\cdot\}$ denotes the principal eigenvector of a matrix, $\tilde{\mathbf{\Phi}}_{\mathrm{u}}^{1/2}$ denotes a square-root decomposition (e.g., Cholesky decomposition) of $\tilde{\mathbf{\Phi}}_{\mathrm{u}}$ and $\tilde{\mathbf{e}}_1 = [1,0,...,0]^T$ denotes a selection vector of same dimensionality as the matrix $\tilde{\mathbf{\Phi}}_{\mathrm{u}}$. Alternatively, the RTF vector $\mathbf{g}_{\mathrm{H}}$ can be estimated from all microphone signals, i.e. the hearing aid and external microphone signals, as

$$\hat{\mathbf{g}}^{(\mathrm{CW-E})}(l) = f\left(\hat{\mathbf{\Phi}}_{\mathrm{y}}(l), \hat{\mathbf{\Phi}}_{\mathrm{u}}(l)\right), \tag{10}$$

$$\hat{\mathbf{g}}_{\mathrm{H}}^{(\mathrm{CW-E})}(l) = \mathbf{E}_{\mathrm{H}}\hat{\mathbf{g}}^{(\mathrm{CW-E})}(l). \tag{11}$$

Due to the involved eigenvalue and square-root decomposition, it can be easily shown that in general $\hat{\mathbf{g}}_{\mathrm{H}}^{(\mathrm{CW-E})}(l) \neq \hat{\mathbf{g}}_{\mathrm{H}}^{(\mathrm{CW})}(l)$. In [20] it was experimentally shown that the RTF vector $\mathbf{g}_{\mathrm{H}}$ can be more accurately estimated using all available microphone signals than using only the hearing aid microphone signals.

Based on the estimated RTF vector $\hat{\mathbf{g}}_{\mathrm{H}}(k,l)$ corresponding to only the hearing aid microphone signals and the database of anechoic

prototype RTF vectors $\bar{\mathbf{g}}_{\mathrm{H}}(k,\theta_i)$, a one-dimensional spatial spectrum is constructed as

$$P(l,\theta_i) = -\sum_{k=2}^{K-1} d(\hat{\mathbf{g}}_{\mathrm{H}}(k,l), \bar{\mathbf{g}}_{\mathrm{H}}(k,\theta_i)). \tag{12}$$

The DOA of the speaker is estimated as the location of the main peak of this spatial spectrum. It should be noted that the spatial spectrum in (12) is obtained via frequency-averaging of narrowband spatial spectra in order to make the DOA estimation more robust against estimation errors in the RTF vector at individual frequencies. Each narrowband spatial spectrum is obtained by assessing the similarity between the estimated RTF vector and an anechoic prototype RTF vector. Inspired by [21], in this paper we consider a similarity measure based on the ℓ_2-norm, i.e.

$$d(\mathbf{a},\mathbf{b}) = \|\exp(i\angle\mathbf{a}) - \exp(i\angle\mathbf{b})\|_2, \tag{13}$$

where the operators $\angle\cdot$ and $\exp(\cdot)$ are applied element-wise to extract the phase and apply the exponential.

3.2. Assisted RTF-vector-based DOA estimation

When considering an uncalibrated external microphone array as in Section 3.1, it should be realized that only the database $\bar{\mathbf{g}}_{\mathrm{H}}(k,\theta_i)$ and thus only the estimated RTF vector $\hat{\mathbf{g}}_{\mathrm{H}}(k,l)$ corresponding to the hearing aid microphones can be considered for the construction of a one-dimensional spatial spectrum as in (12). Since in this paper the external microphone array is assumed to be calibrated, i.e. a database $\bar{\mathbf{g}}_{\mathrm{E}}(k,\theta_{\mathrm{E},j})$ of anechoic prototype RTF vectors is available, the entire estimated RTF vector $\hat{\mathbf{g}}(k,l)$ and both prototype databases can - and should - be utilized. We first propose to construct a two-dimensional spatial spectrum that exploits both RTF vector databases $\bar{\mathbf{g}}_{\mathrm{H}}(k,\theta_i)$ and $\bar{\mathbf{g}}_{\mathrm{E}}(k,\theta_{\mathrm{E},j})$ jointly with the entire RTF vector $\hat{\mathbf{g}}^{(\mathrm{CW-E})}(k,l)$ in order to exploit spatial correlations between microphone signals of both arrays for improving the DOA estimation. We then consider a special case, requiring prior information about relative position and orientation of both arrays.

We define two M-dimensional concatenated RTF vectors, similarly to $\mathbf{g}$ in (3). The vector

$$\bar{\mathbf{g}}^{\mathrm{joint}}(k,\theta_i,\theta_{\mathrm{E},j}) = \begin{bmatrix} \bar{\mathbf{g}}_{\mathrm{H}}(k,\theta_i) \\ \bar{\mathbf{g}}_{\mathrm{E}}(k,\theta_{\mathrm{E},j}) \end{bmatrix}. \tag{14}$$

concatenates the anechoic prototype RTF vectors from both databases and is obtained for each pair of DOAs θ_i and $\theta_{\mathrm{E},j}$. Similarly, the vector

$$\tilde{\mathbf{g}}(k,l) = \begin{bmatrix} \hat{\mathbf{g}}_{\mathrm{H}}^{(\mathrm{CW-E})}(k,l) \\ \hat{\mathbf{g}}_{\mathrm{E}}^{(\mathrm{CW-E})}(k,l) \end{bmatrix} \tag{15}$$

concatenates the estimated RTF vector corresponding to the hearing aid microphones $\hat{\mathbf{g}}_{\mathrm{H}}^{(\mathrm{CW-E})}(k,l)$ in (11) and the estimated RTF vector corresponding to the external microphones, which is equal to

$$\hat{\mathbf{g}}_{\mathrm{E}}^{(\mathrm{CW-E})}(l) = \frac{\mathbf{E}_{\mathrm{E}}\hat{\mathbf{g}}^{(\mathrm{CW-E})}(l)}{\mathbf{e}_{1,\mathrm{E}}^T\mathbf{E}_{\mathrm{E}}\hat{\mathbf{g}}^{(\mathrm{CW-E})}(l)} \tag{16}$$

similarly to (4). A two-dimensional spatial spectrum is now constructed by assessing the similarity between the concatenated vectors in (15) and (16), i.e.

$$\boxed{P^{\mathrm{joint}}(l,\theta_i,\theta_{\mathrm{E},j}) = -\sum_{k=2}^{K-1} d\Big(\tilde{\mathbf{g}}(k,l), \bar{\mathbf{g}}^{\mathrm{joint}}(k,\theta_i,\theta_{\mathrm{E},j})\Big)} \tag{17}$$

with the similarity measure defined in (13). The DOA of the speaker is estimated as the location of the main peak of this two-dimensional spatial spectrum. Fig. 2 depicts an exemplary spectrum for a static speaker in a reverberant environment with spatially diffuse babble noise. It should be realized that the main peak consists of the estimated DOA $\hat{\theta}(l)$ as well as the estimated angle $\hat{\theta}_{\mathrm{E}}(l)$ of the speaker as "seen" from both local coordinate systems (compare Fig. 1). In this paper, however, we will evaluate the estimation accuracy of $\hat{\theta}(l)$ only, as we are mainly interested in how the array of external microphones assists the binaural hearing aid setup in estimating the DOA θ.

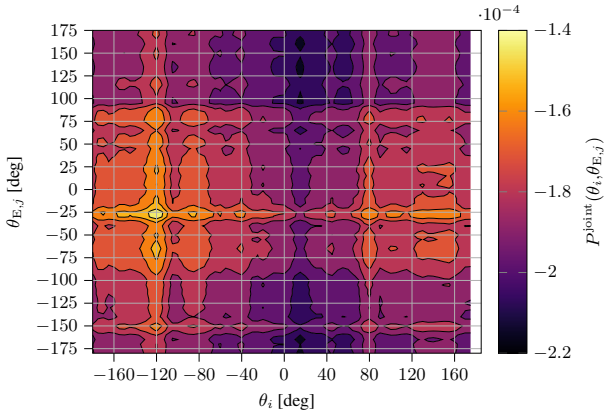

Fig. 2. Example of a two-dimensional spatial spectrum $P^{\mathrm{joint}}(\theta_i,\theta_{\mathrm{E},j})$. Corresponding acoustic scene: $\theta = -120°$ and $\theta_{\mathrm{E}} = -25°$, $T_{60} \approx 1250\mathrm{ms}$, $\mathrm{SNR} = 30\mathrm{dB}$.

To further investigate the potential of the proposed method, we also consider a special case that transforms the two-dimensional spatial spectrum in (17) into a one-dimensional spatial spectrum. Contrary to $P^{\mathrm{joint}}(l,\theta_i,\theta_{\mathrm{E},j})$ that considers each pair of DOAs θ_i and $\theta_{\mathrm{E},j}$ in this special case only a subset of pairs is considered. In particular, only those pairs of DOAs θ_i and $\theta_{\mathrm{E},i}$ are considered where the respective prototype RTF vectors $\bar{\mathbf{g}}_{\mathrm{H}}(k,\theta_i)$ and $\bar{\mathbf{g}}_{\mathrm{E}}(k,\theta_{\mathrm{E},i})$ are spatially matched, i.e. steer towards the same candidate position. To obtain the respective DOA pairs prior information about the relative position and orientation of both arrays is required. Based on this information, a coordinate system transformation between the two local coordinate systems can be applied, i.e. $\theta_{\mathrm{E},i} = g(\theta_i,\theta_{\mathrm{E},j})$. Enforcing this matching condition makes the resulting spatial spectrum one-dimensional, i.e.

$$P^{\mathrm{match}}(l,\theta_i) = -\sum_{k=2}^{K-1} d\Big(\tilde{\mathbf{g}}(k,l), \bar{\mathbf{g}}^{\mathrm{match}}(k,\theta_i,\theta_{\mathrm{E},i})\Big). \tag{18}$$

Similarly to (17), the DOA of the speaker is estimated as the location of the main peak.

4. EXPERIMENTAL RESULTS

In this section we compare the DOA estimation accuracy when the exploited array of external microphone is calibrated or not. The experimental setup is described in Section 4.1 and the considered algorithms and implementation details are described in Section 4.2. The results are presented and discussed Section 4.3.

4.1. Experimental setup and implementation details

For the experiments we used signals that were recorded in a laboratory at the University of Oldenburg with dimensions about $7 \times 6 \times 2.7\,\mathrm{m}^3$, where the reverberation time can be adjusted by means of absorber panels, which are mounted to the walls and the ceiling. The reverberation time is set to approximately $T_{60} \approx 1250\mathrm{ms}$. A dummy head with a binaural hearing aid setup ($M_{\mathrm{H}} = 4$) is placed approximately in the center

of the laboratory. For this hearing aid setup a database of anechoic prototype RTF vectors is obtained from measured anechoic binaural room impulse responses (BRIRs) [22] with an angular resolution of 5° ($I = 72$). As an array of external microphones we consider a second dummy head with the same binaural hearing setup ($M_{\mathrm{E}} = 4$) that is located about $1\,\mathrm{m}$ and about 80° to the right side from the first dummy head (see Fig. 1). As the same binaural hearing aid setup is considered for the external microphone array, $\bar{\mathbf{g}}_{\mathrm{E}}(k,\theta_{\mathrm{E},j}) = \bar{\mathbf{g}}_{\mathrm{H}}(k,\theta_{\mathrm{E},j})$ and $J = I$. The target speaker was a male English speaker played back via a loudspeaker. Relative to the first dummy head 9 different speaker DOAs in the range $[-160^\circ,-120^\circ,...,160^\circ]$ at about $2\,\mathrm{m}$ distance are considered. The speech signal is constantly active and is approximately $4\,\mathrm{s}$ long. Diffuse-like noise is generated with four loudspeakers facing the corners of the laboratory, playing back different multi-talker recordings. The speech and noise components are recorded separately and are mixed at $0\,\mathrm{dB}$ signal-to-noise ratio (SNR) averaged over all microphones of the first dummy head. All microphone signals are assumed to be exchanged without errors and are assumed to be synchronized such that latency aspects can be neglected.

4.2. Algorithms and implementation details

We compare the following algorithms to assess whether a calibrated external microphone array can assist the binaural hearing setup in estimating the speaker DOA θ. The notation "X/Y" means that the microphone array X is used to estimate an RTF vector and that the microphone array Y is used to construct a spatial spectrum, where either the binaural hearing aid setup (denoted as H) alone or jointly with the external microphone array (denoted as H+E) are considered as options.

- H/H: considering only the hearing aid microphones of the first dummy head, both to estimate the RTF vector $\hat{\mathbf{g}}_{\mathrm{H}}(k,l)$ in (8) as well as to construct the spatial spectrum $P(l,\theta_i)$ in (12).
- H+E/H: considering all microphones to estimate the RTF vector $\hat{\mathbf{g}}_{\mathrm{H}}(k,l)$ in (11) but considering only the hearing aid microphones of the first dummy head to construct the spatial spectrum $P(l,\theta_i)$ in (12).
- H+E/H+E (2D): considering all microphones, both to estimate the concatenated RTF vector $\tilde{\mathbf{g}}(k,l)$ in (15) as well as to construct the two-dimensional spatial spectrum $P^{\mathrm{joint}}(l,\theta_i,\theta_{\mathrm{E},j})$ in (17).
- H+E/H+E (match): as a special case of H+E/H+E (2D) also all microphones are considered, both to estimate the concatenated RTF vector $\tilde{\mathbf{g}}(k,l)$ in (15) as well as to construct the one-dimensional spatial spectrum $P^{\mathrm{match}}(l,\theta_i)$ in (18) exploiting prior knowledge about the microphone configuration. Due to the enforced matching condition the spatial spectrum $P^{\mathrm{match}}(l,\theta_i)$ consists of only 20 candidate speaker positions.

The microphone signals (sampling frequency $f_{\mathrm{s}} = 16\,\mathrm{kHz}$) are processed in the STFT domain using $32\,\mathrm{ms}$ square-root Hann windows with $50\,\%$ overlap. For each time-frequency (TF) bin the covariance matrix of the noisy microphone signals $\mathbf{\Phi}_{\mathrm{y}}(k,l)$ and the covariance matrix of the undesired components $\mathbf{\Phi}_{\mathrm{u}}(k,l)$ are estimated recursively during speech-and-noise TF bins and noise-only TF bins as

$$\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l) = \alpha_{\mathrm{y}}\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l-1) + (1-\alpha_{\mathrm{y}})\mathbf{y}(k,l)\mathbf{y}^H(k,l), \quad (19)$$

$$\hat{\mathbf{\Phi}}_{\mathrm{u}}(k,l) = \alpha_{\mathrm{u}}\hat{\mathbf{\Phi}}_{\mathrm{u}}(k,l-1) + (1-\alpha_{\mathrm{u}})\mathbf{y}(k,l)\mathbf{y}^H(k,l), \quad (20)$$

where the smoothing factors α_{y} and α_{u} correspond to time constants of $250\,\mathrm{ms}$ and $500\,\mathrm{ms}$, respectively. The speech-and-noise TF bins are discriminated from noise-only TF bins based on estimated speech presence probabilities [23] in the microphones of the first dummy head, which are averaged and thresholded.

To assess the DOA estimation performance, we average the localization accuracy over the considered 9 speaker DOAs, where the localization accuracy is defined as the percentage of frames that are correctly localized within a range of $\pm 5^\circ$.

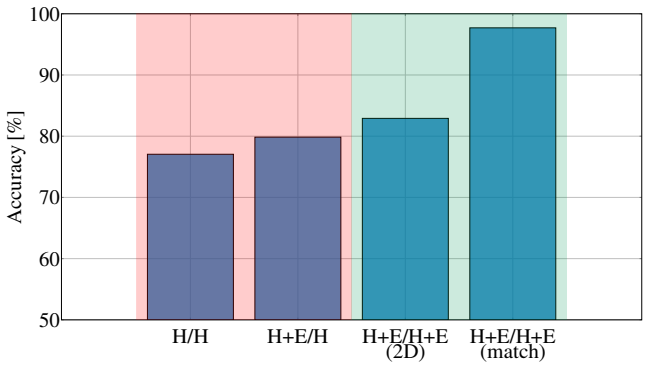

Fig. 3. Average localization accuracy for the investigated algorithms that either exploit calibrated external microphones (green background) or not (red background).

4.3. Results

Fig. 3 depicts the average localization accuracy for the four investigated algorithms. First, it can be observed that for the algorithm H/H using only the hearing aid microphones of the first dummy head and for the algorithm H+E/H, where the external microphones are exploited only to estimate the RTF vector, there is only a minor performance difference (average localization accuracy for both algorithms about $80\,\%$). This result is in line with the high-reverberation results reported in [7]. Second, considering the "H+E/H+E" algorithms, which both exploit the external microphones to estimate the RTF vector as well as to construct a spatial spectrum, the results clearly show that both algorithms yield a larger average localization accuracy than the algorithms that do not consider external microphones to construct a spatial spectrum. In case of the H+E/H+E (2D) algorithm, where a two-dimensional spatial spectrum is constructed, this improved average localization accuracy is due to exploitation of the additional database of prototype RTF vectors $\bar{\mathbf{g}}_{\mathrm{E}}(k,\theta_{\mathrm{E},j})$. Third, in case of the H+E/H+E (match) algorithm, where prior knowledge about the microphone configuration is additionally exploited, the average localization accuracy can be significantly improved. Based on these results, the potential of assisted DOA estimation is clearly demonstrated.

5. CONCLUSIONS

In this paper we explored how calibrated external microphone arrays can be exploited to assist in estimating the DOA of a target speaker. We extended a recently proposed binaural RTF-vector-based DOA estimation method that considered only uncalibrated external microphone arrays to exploit calibrated external microphone arrays. We proposed to exploit the availability of two databases of anechoic prototype RTF vectors, i.e. the binaural hearing aid as well as the external microphone, in order to construct a two-dimensional spatial spectrum from which the DOA of the speaker can be estimated. We compared RTF-vector-based DOA estimation algorithms that either exploit a calibrated array of external microphones or not. Experimental results clearly demonstrate a benefit of exploiting a calibrated external microphone array compared to not exploiting a calibrated external microphone array or using only the hearing aid microphones for DOA estimation. Thus, the benefit of assisted DOA estimation is clearly demonstrated.

6. REFERENCES

[1] J. Mecklenburger and T. Groth, *Wireless Technologies and Hearing Aid Connectivity*, pp. 131–149, Springer, Cham, Switzerland, 2016.

[2] A. Bertrand and M. Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP J. Adv. Signal Processing*, vol. 2009, Oct. 2009.

[3] D. Yee, H. Kamkar-Parsi, R. Martin, and H. Puder, "A noise reduction postfilter for binaurally linked single-microphone hearing aids utilizing a nearby external microphone," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 5–18, Jan. 2018.

[4] R. Ali, G. Bernardi, T. van Waterschoot, and M. Moonen, "Methods of extending a generalized sidelobe canceller with external microphones," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1349–1364, Sep. 2019.

[5] N. Gößling, W. Middelberg, and S. Doclo, "RTF-steered binaural MVDR beamforming incorporating multiple external microphones," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2019, pp. 373–377.

[6] N. Gößling, D. Marquardt, and S. Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 462–476, Dec. 2021.

[7] 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, Aug. 2021, pp. 241–245.

[8] 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, no. 7, pp. 1275–1289, Jul. 2018.

[9] M. S. Kavalekalam, J. K. Nielsen, M. G. Christensen, and J. B. Boldt, "Hearing aid-controlled beamformer for binaural speech enhancement using a model-based approach," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 321–325.

[10] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[11] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Robust Localization in Reverberant Rooms*, pp. 157–180, Springer, Berlin, Heidelberg, Germany, 2001.

[12] S. Gannot, M. Haardt, W. Kellermann, and P. Willett, "Introduction to the issue on acoustic source localization and tracking in dynamic real-life scenes," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 3–7, Mar. 2019.

[13] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.

[14] 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, no. 1, pp. 107–151, Jul. 2022.

[15] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

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

[17] 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, Sep. 2022, pp. 1–5.

[18] 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, no. 6, pp. 1071–1086, Aug. 2009.

[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] A. Hassani, A. Bertrand, and M. Moonen, "Cooperative integrated noise reduction and node-specific direction-of-arrival estimation in a fully connected wireless acoustic sensor network," *Signal Processing*, vol. 107, pp. 68–81, Feb. 2015.

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

[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, no. 1, pp. 1–10, Jul. 2009.

[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, no. 4, pp. 1383–1393, May 2012.