ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889277

Completing Sets of Prototype Transfer Functions for Subspace-based Direction of Arrival Estimation of Multiple Speakers

Daniel Fejgin and Simon Doclo
Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all,
Carl von Ossietzky Universität Oldenburg, Germany

***Abstract*—To estimate the direction of arrival (DOA) of multiple speakers, subspace-based prototype transfer function matching methods such as multiple signal classification (MUSIC) or relative transfer function (RTF) vector matching are commonly employed. In general, these methods require calibrated microphone arrays, which are characterized by a known array geometry or a set of known prototype transfer functions for several directions. In this paper, we consider a partially calibrated microphone array, composed of a calibrated binaural hearing aid and a (non-calibrated) external microphone at an unknown location with no available set of prototype transfer functions. We propose a procedure for completing sets of prototype transfer functions by exploiting the orthogonality of subspaces, allowing to apply matching-based DOA estimation methods with partially calibrated microphone arrays. For the MUSIC and RTF vector matching methods, experimental results for two speakers in noisy and reverberant environments clearly demonstrate that for all locations of the external microphone DOAs can be estimated more accurately with completed sets of prototype transfer functions than with incomplete sets.**

***Index Terms*—direction of arrival estimation, subspaces, binaural hearing aids, external microphone**

I. Introduction

Many speech communication applications such as hearing aids and hands-free conferencing systems require estimates of the direction of arrival (DOA) of multiple speakers in noisy and reverberant environments. Over the last decades, many model-based and machine-learning-based methods have been developed for DOA estimation, typically requiring knowledge about the geometrical configuration of the microphone array [1]–[11]. However, applying these methods to DOA estimation using microphone arrays with a partially unknown geometrical configuration, e.g., a hearing aid linked to one or more external microphones (eMics) at unknown locations, is not straightforward.

In [12]–[17] DOA estimation methods have been proposed which exploit the availability of one or more eMics. The methods in [12] and [14] utilize the eMic signal to estimate a clean speech reference signal and a voice activity detector, however, severely restricting the location of the eMic to the vicinity of the target speaker. Without restricting the location of the eMic, the methods in [13], [15], and [16] utilize the eMic signal to estimate relative transfer function (RTF) vectors or generalized cross correlations which are subsequently used to construct a spatial spectrum (spatial map). However, since the location of the eMic is unknown only a spatial spectrum for the microphones of the calibrated array instead of a spatial spectrum for all microphones is constructed. In this context, calibrated arrays are defined as arrays for which a set of (anechoic) prototype transfer function vectors for several directions is available, either because the array configuration is known or a set of measured transfer functions is available. Opposed to these methods, the method in [17] utilizes the signals from a calibrated array of eMics at unknown locations to estimate RTF vectors as well as to construct a spatial spectrum for all available microphones.

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.

To perform DOA estimation using prototype transfer function vector matching-based methods, in this paper we aim at exploiting the eMic for the construction of a spatial spectrum for all available microphones, however, without requiring the availability of prototype transfer functions for the eMic. We propose an optimal procedure in the least-squares sense that completes sets of prototype transfer function vectors with a transfer function corresponding to the eMic. We complete these sets by exploiting the orthogonality of the complementary signal subspace and the noise subspace obtained from the eigenvalue decomposition of the covariance matrix of pre-whitened signals. For a binaural hearing aid setup with an external microphone at an unknown location, we compare the DOA estimation performance with the incomplete set (i.e., using only the microphone signals of the hearing aids) and the proposed completed sets of prototype transfer function vectors using MUSIC [5] and the RTF vector matching method [15]. Results using real-world recordings from the BRUDEX database [18] clearly demonstrate a benefit of the proposed method for multiple eMic locations that are distributed over a large area.

II. Signal model and notation

In a noisy and reverberant acoustic environment, we consider J speakers which are recorded using a binaural hearing aid setup with $M/2$ microphones on each hearing aid and one eMic, i.e., a total of $M+1$ microphones. The eMic is spatially separated from the hearing aids at an unknown location (see Fig. 1). 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 $X_{m,j}$ and N_m denote the speech component of the j-th speaker and the noise component, respectively, where $m \in \{1,...,M+1=\mathrm{E}\}$, $k \in \{1,...,K\}$ and $l \in \{1,...,L\}$ denote the microphone index, the frequency bin index, and the frame index, respectively. Denoting the stacked vector of microphone signals as $\mathbf{y}(k,l) = [Y_1(k,l),...,Y_{\mathrm{E}}(k,l)]^T \in \mathbb{C}^{M+1}$, with $(\cdot)^T$ denoting the transposition operator and assuming disjoint speaker activity in the STFT domain [19], this vector can be approximated as $\mathbf{y}(k,l) \approx \mathbf{x}_{\mathrm{d}}(k,l) + \mathbf{n}(k,l)$, where the vectors $\mathbf{x}_{\mathrm{d}}$ and $\mathbf{n}$ denote the speech component of the dominant speaker and the noise component, respectively, both defined similarly as $\mathbf{y}$. For conciseness, we omit the time-frequency (TF) bin indices k and l when possible in the following.

We assume that the speech component $\mathbf{x}_{\mathrm{d}}$ can be split into a direct-path component $\mathbf{x}_{\mathrm{d}}^{\mathrm{DP}}$ and a reverberant component $\mathbf{x}_{\mathrm{d}}^{\mathrm{R}}$, i.e., $\mathbf{x}_{\mathrm{d}} = \mathbf{x}_{\mathrm{d}}^{\mathrm{DP}} + \mathbf{x}_{\mathrm{d}}^{\mathrm{R}}$. Condensing the noise and reverberation components into the undesired component $\mathbf{u} = \mathbf{n} + \mathbf{x}_{\mathrm{d}}^{\mathrm{R}}$, the vector of microphone signals can be written as $\mathbf{y} = \mathbf{x}_{\mathrm{d}}^{\mathrm{DP}} + \mathbf{u}$.

Approximating the direct-path component with a multiplicative transfer function [20] using the direct-path acoustic transfer function (ATF) vector $\mathbf{a}(\theta_{\mathrm{d}}) = [A_1(\theta_{\mathrm{d}}),...,A_{\mathrm{E}}(\theta_{\mathrm{d}})]^T$, with θ_{d} denoting the

DOA of the dominant speaker relative to the binaural hearing aid setup, the vector $\mathbf{x}_{\mathrm{d}}^{\mathrm{DP}}$ can be written as

$$\mathbf{x}_{\mathrm{d}}^{\mathrm{DP}} = \mathbf{a}(\theta_{\mathrm{d}}) S_{\mathrm{d}} = \mathbf{g}(\theta_{\mathrm{d}}) X_{1,\mathrm{d}}, \tag{2}$$

where S_{d} denotes the dominant speech signal and $\mathbf{g}(\theta_{\mathrm{d}}) = \mathbf{a}(\theta_{\mathrm{d}})/A_1(\theta_{\mathrm{d}})$ denotes the direct-path relative transfer function (RTF) vector, assuming the first microphone to be the reference microphone. For the binaural hearing aid setup only, we assume the availability of a set of anechoic prototype ATF vectors $\bar{\mathbf{a}}_{\mathrm{HA}}(k,\theta_i)$ for different candidate directions $\theta_i, i=1,...,I$. The set of anechoic prototype RTF vectors is obtained as $\bar{\mathbf{g}}_{\mathrm{HA}}(k,\theta_i) = \frac{\bar{\mathbf{a}}_{\mathrm{HA}}(k,\theta_i)}{\mathbf{e}_1^T \bar{\mathbf{a}}_{\mathrm{HA}}(k,\theta_i)}$, where $\mathbf{e}_m = [0,...,1,0,...,0]$ denotes a selection vector with all zeros except the m-th element. Since the location of the eMic is unknown, obviously no sets of anechoic prototype ATFs $\bar{A}_{\mathrm{E}}(k,\theta_i)$ and RTFs $\bar{G}_{\mathrm{E}}(k,\theta_i)$ are available for the eMic. Therefore, the considered microphone array is referred to as partially calibrated.

Assuming uncorrelated direct-path speech and undesired components, the $(M+1)\times(M+1)$ noisy covariance matrix can be written, using (2), as

$$\mathbf{\Phi}_{\mathrm{y}} = \mathcal{E}\left\{\mathbf{y}\mathbf{y}^H\right\} = \mathbf{\Phi}_{\mathrm{x_d}}^{\mathrm{DP}} + \mathbf{\Phi}_{\mathrm{u}} = \phi_{\mathrm{s}} \mathbf{a}(\theta_{\mathrm{d}}) \mathbf{a}^H(\theta_{\mathrm{d}}) + \mathbf{\Phi}_{\mathrm{u}}, \tag{3}$$

where $(\cdot)^H$ and $\mathcal{E}\{\cdot\}$ denote the complex transposition and expectation operators, respectively, $\mathbf{\Phi}_{\mathrm{u}} = \mathcal{E}\{\mathbf{u}\mathbf{u}^H\}$ denotes the covariance matrix of the undesired component, and $\phi_{\mathrm{s}} = \mathcal{E}\{|S_{\mathrm{d}}|^2\}$ denotes the power spectral density of the dominant speaker. Based on (3) and a square-root decomposition (e.g., Cholesky decomposition) of the covariance matrix of the undesired component $\mathbf{\Phi}_{\mathrm{u}} = \mathbf{\Phi}_{\mathrm{u}}^{1/2} \mathbf{\Phi}_{\mathrm{u}}^{H/2}$, the noisy covariance matrix after pre-whitening can be written as

$$\mathbf{\Phi}_{\mathrm{y}}^{\mathrm{w}} = \mathbf{\Phi}_{\mathrm{u}}^{-1/2} \mathbf{\Phi}_{\mathrm{y}} \mathbf{\Phi}_{\mathrm{u}}^{-H/2} = \phi_{\mathrm{s}} \mathbf{a}^{\mathrm{w}}(\theta_{\mathrm{d}}) (\mathbf{a}^{\mathrm{w}}(\theta_{\mathrm{d}}))^H + \mathbf{I}_{M+1\times M+1}, \tag{4}$$

with $\mathbf{a}^{\mathrm{w}}(\theta_{\mathrm{d}}) = \mathbf{\Phi}_{\mathrm{u}}^{-1/2} \mathbf{a}(\theta_{\mathrm{d}})$ denoting the pre-whitened direct-path ATF vector. The $M \times M$ covariance matrices $\mathbf{\Phi}_{\mathrm{y,HA}}$, $\mathbf{\Phi}_{\mathrm{y,HA}}^{\mathrm{w}}$, and $\mathbf{\Phi}_{\mathrm{u,HA}}$ corresponding to the hearing aid signals can be extracted from (3) and (4) as $\mathbf{\Phi}_{\mathrm{y,HA}} = \mathbf{E}_{\mathrm{HA}} \mathbf{\Phi}_{\mathrm{y}} \mathbf{E}_{\mathrm{HA}}^T$, $\mathbf{\Phi}_{\mathrm{y,HA}}^{\mathrm{w}} = \mathbf{E}_{\mathrm{HA}} \mathbf{\Phi}_{\mathrm{y}}^{\mathrm{w}} \mathbf{E}_{\mathrm{HA}}^T$ and $\mathbf{\Phi}_{\mathrm{u,HA}} = \mathbf{E}_{\mathrm{HA}} \mathbf{\Phi}_{\mathrm{u}} \mathbf{E}_{\mathrm{HA}}^T$ with the selection matrix $\mathbf{E}_{\mathrm{HA}} = [\mathbf{I}_{M\times M}, \mathbf{0}_{M\times 1}]$, where $\mathbf{I}_{M\times M}$ denotes an $M \times M$ identity matrix and $\mathbf{0}_{M\times 1}$ denotes a vector with zeros.

III. Subspace-based DOA Estimation Methods

To estimate the DOAs $\theta_{1:J}$ of all speakers, we consider two baseline methods, namely MUSIC [5] and the RTF vector matching method presented in [15]. Both methods consider complementary subspaces obtained from the same subspace decomposition. In Section III-A, we review for both methods the construction of frequency-dependent spatial spectra (SPS), i.e., functions of candidate directions with the location of peaks likely corresponding to the true speaker DOAs. In Section III-B, we review how DOAs are estimated from these frequency-dependent SPS (see Fig. 2).

A. Construction of frequency-dependent SPS

The frequency-dependent SPS for both baseline methods are obtained from the eigenvalue decomposition (EVD) of an estimate of the noisy covariance matrix $\hat{\mathbf{\Phi}}_{\mathrm{y}}$ after pre-whitening with a square-root decomposition (e.g., Cholesky decomposition) of an estimate of the covariance matrix of the undesired component $\hat{\mathbf{\Phi}}_{\mathrm{u}} = \hat{\mathbf{\Phi}}_{\mathrm{u}}^{1/2} \hat{\mathbf{\Phi}}_{\mathrm{u}}^{H/2}$ [5], [15], [21], i.e.

$$\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}} = \hat{\mathbf{\Phi}}_{\mathrm{u}}^{-1/2} \hat{\mathbf{\Phi}}_{\mathrm{y}} \hat{\mathbf{\Phi}}_{\mathrm{u}}^{-H/2} \overset{\mathrm{EVD}}{=} \hat{\mathbf{Q}} \hat{\mathbf{\Lambda}} \hat{\mathbf{Q}}^H, \tag{5}$$

where $\hat{\mathbf{Q}}$ and $\hat{\mathbf{\Lambda}}$ denote the unitary matrix containing the eigenvectors and the diagonal matrix containing the eigenvalues of the pre-whitened matrix $\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}}$, respectively. The matrix of eigenvectors can be partitioned into the signal subspace $\mathcal{P}\{\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}}\}$ and noise subspace $\hat{\mathbf{Q}}_{\mathrm{n}}$ as $\hat{\mathbf{Q}} = \left[\mathcal{P}\{\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}}\}, \hat{\mathbf{Q}}_{\mathrm{n}}\right]$, where $\mathcal{P}\{\cdot\}$ denotes the principal eigenvector of a matrix.

The signal subspace of $\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}}$ is spanned by the principal eigenvector $\mathcal{P}\{\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}}\}$ and based on (4) it is assumed that this vector does not differ too much from the vector $\mathbf{a}^{\mathrm{w}}(\theta_{\mathrm{d}})$. This assumption is utilized by the state-of-the-art covariance whitening (CW) method [22] to estimate RTF vectors via de-whitening and normalization, i.e.,

$$\hat{\mathbf{g}}(k,l) = \frac{\hat{\mathbf{\Phi}}_{\mathrm{u}}^{1/2}(k,l) \mathcal{P}\{\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}}(k,l)\}}{\mathbf{e}_1^T \hat{\mathbf{\Phi}}_{\mathrm{u}}^{1/2}(k,l) \mathcal{P}\{\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}}(k,l)\}}. \tag{6}$$

The noise subspace of $\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}}$ is spanned by the columns of $\hat{\mathbf{Q}}_{\mathrm{n}}$. Both subspaces are orthogonal to each other. For the construction of frequency-dependent SPS, MUSIC considers only the noise subspace whereas the RTF vector matching method considers only the signal subspace.

Exploiting the orthogonality between the signal and noise subspaces, MUSIC estimates the speaker DOAs by searching for vectors from a set of pre-whitened prototype ATF vectors $\bar{\mathbf{a}}^{\mathrm{w}}(k,l,\theta_i) = \hat{\mathbf{\Phi}}_{\mathrm{u}}^{-1/2}(k,l) \bar{\mathbf{a}}(k,\theta_i) = [\bar{\mathbf{a}}_{\mathrm{HA}}^{\mathrm{w}}(k,l,\theta_i), \bar{A}_{\mathrm{E}}^{\mathrm{w}}(k,l,\theta_i)]^T$ with $\bar{\mathbf{a}}(k,\theta_i) = [\bar{\mathbf{a}}_{\mathrm{HA}}(k,\theta_i), \bar{A}_{\mathrm{E}}(k,\theta_i)]^T$ for different candidate directions θ_i that maximize the orthogonality with the estimated noise subspace. Thus, the frequency-dependent SPS is constructed as follows

$$\tilde{p}^{\mathrm{MUSIC}}(k,l,\theta_i) = \frac{1}{\|\hat{\mathbf{Q}}_{\mathrm{n}}^H(k,l) \bar{\mathbf{a}}^{\mathrm{w}}(k,l,\theta_i)\|_2^2}. \tag{7}$$

Instead of considering the SPS in (7) directly for DOA estimation, we will consider the normalized SPS as suggested by [23], i.e.,

$$p^{\mathrm{MUSIC}}(k,l,\theta_i) = \frac{\tilde{p}^{\mathrm{MUSIC}}(k,l,\theta_i)}{\max\limits_{\theta_{i'}} \tilde{p}^{\mathrm{MUSIC}}(k,l,\theta_{i'})}. \tag{8}$$

Exploiting the assumed parallelity of the estimated signal subspace with the pre-whitened direct-path ATF vector, the RTF vector matching method estimates the speaker DOAs by searching for vectors from a set of prototype RTF vectors $\bar{\mathbf{g}}(k,\theta_i) = [\bar{\mathbf{g}}_{\mathrm{HA}}(k,\theta_i), \bar{G}_{\mathrm{E}}(k,\theta_i)]^T$ for different candidate directions θ_i that maximize the parallelity with the estimated RTF vectors. Considering the Hermitian angle [24] between the estimated and prototype RTF vector as a measure for parallelity, the frequency-dependent SPS is constructed as follows

$$p^{\mathrm{RTF}}(k,l,\theta_i) = -\arccos\left(\frac{|\bar{\mathbf{g}}^H(k,\theta_i) \hat{\mathbf{g}}(k,l)|}{\|\bar{\mathbf{g}}(k,\theta_i)\|_2 \|\hat{\mathbf{g}}(k,l)\|_2}\right). \tag{9}$$

It should be noted that in order to compute the SPS according to (8) and (9), prototype transfer functions $\bar{\mathbf{a}}^{\mathrm{w}}(k,l,\theta_i)$ and $\bar{\mathbf{g}}(k,\theta_i)$ for *all* microphones must be available. Since for the considered partially calibrated microphone array only $\bar{\mathbf{a}}_{\mathrm{HA}}^{\mathrm{w}}(k,l,\theta_i)$ and $\bar{\mathbf{g}}_{\mathrm{HA}}(k,\theta_i)$ are available, these SPS cannot be calculated. Hence, to estimate DOAs using the considered methods one can either consider the signals of the binaural hearing aid setup only (leading to the construction of SPS obtained from the subspace decomposition of $\hat{\mathbf{\Phi}}_{\mathrm{y,HA}}^{\mathrm{w}} = \mathbf{E}_{\mathrm{HA}} \hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}} \mathbf{E}_{\mathrm{HA}}^T$) or one needs to complete the sets of prototype transfer function vectors $\bar{\mathbf{a}}_{\mathrm{HA}}^{\mathrm{w}}(k,l,\theta_i)$ and $\bar{\mathbf{g}}_{\mathrm{HA}}(k,\theta_i)$ with elements corresponding to the eMic, which is the topic of this paper (see Section IV).

B. Fusion of frequency-dependent SPS

To exploit spatial information across multiple frequencies, the spatial spectra $p(k,l,\theta_i)$ in (8) or (9) are combined as proposed in [25] using the speaker-grouped frequency fusion mechanism [26]. Assuming the

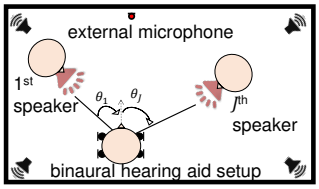

Fig. 1: Considered noisy and reverberant acoustic scenario with J speakers, a binaural hearing aid setup and an external microphone.

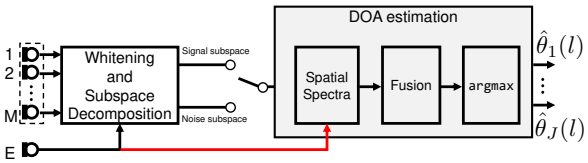

Fig. 2: Block diagram of the baseline (without red arrow) and proposed (with arrow) subspace-based DOA estimation methods.

number of speakers J to be known, the speaker DOAs are estimated from frequency-averaged SPS, each associated with a single speaker, as

$$\hat{\theta}_j(l) = \underset{\theta_i}{\operatorname{argmax}} \sum_{k\in\mathcal{K}(l)} \mathbb{1}_j(k,l)p(k,l,\theta_i), \quad j=1,\ldots,J, \tag{10}$$

where $\mathcal{K}$ denotes the subset of frequencies where one speaker is assumed to dominate over all other speakers, noise, and reverberation, and $\mathbb{1}_j(k,l)$ denotes an indicator function that denotes the association between the (k,l)-th TF bin and the j-th speaker, i.e.,

$$\mathbb{1}_j(k,l) = \begin{cases} 1 & \text{TF bin } (k,l) \text{ associated with speaker } j \\ 0 & \text{else.} \end{cases} \tag{11}$$

As proposed in [27], to estimate the frequency subset $\mathcal{K}(l)$ we consider a criterion based on the binaural effective-coherence-based coherent-to-diffuse ratio (CDR), i.e.,

$$\mathcal{K}(l) = \left\{k : \widehat{\mathrm{CDR}}(k,l) \geq \mathrm{CDR}_{\mathrm{thresh}}\right\}, \tag{12}$$

where $\widehat{\mathrm{CDR}}$ denotes an estimate of the CDR and $\mathrm{CDR}_{\mathrm{thresh}}$ denotes a threshold value. As proposed in [25], to estimate the indicator function $\mathbb{1}_j(k,l)$ we discriminate the speakers spatially using estimated interaural time differences $\hat{\tau}_j(l)$ and a score function $\Psi(k,l,\hat{\tau}_j(l))$, i.e.

$$\mathbb{1}_j(k,l) = \begin{cases} 1 & \text{if } j = \underset{j'\in\{1,\ldots,J\}}{\operatorname{argmax}} \Psi(k,l,\hat{\tau}_{j'}(l)) \\ 0 & \text{else.} \end{cases} \tag{13}$$

Details regarding the estimation of $\mathcal{K}$ and $\mathbb{1}_j(k,l)$ can be found in [27] and [25], respectively.

IV. Completing Sets of Prototype Transfer Functions

In order to construct frequency-dependent SPS according to (8) and (9), in this section we propose a procedure that completes sets of prototype transfer function vectors.

Performing an EVD on the noisy covariance matrix after pre-whitening $\mathbf{\Phi}_{\mathrm{y}}^{\mathrm{w}}$ in (4), results in $\mathbf{\Phi}_{\mathrm{y}}^{\mathrm{w}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H$, where $\mathbf{Q} = \left[\mathbf{a}^{\mathrm{w}}(\theta_{\mathrm{d}}), \mathbf{Q}_{\mathrm{n}}\right]$ and $\mathbf{Q}_{\mathrm{n}}$ denotes the noise subspace of $\mathbf{\Phi}_{\mathrm{y}}^{\mathrm{w}}$. Partitioning the noise subspace as $\mathbf{Q}_{\mathrm{n}} = [\mathbf{Q}_{\mathrm{n,HA}}^T, \mathbf{q}_{\mathrm{n,E}}]^T$ with $\mathbf{Q}_{\mathrm{n,HA}} = \mathbf{E}_{\mathrm{HA}}\mathbf{Q}_{\mathrm{n}}$ and $\mathbf{q}_{\mathrm{n,E}} = \mathbf{Q}_{\mathrm{n}}^T\mathbf{e}_{\mathrm{E}}$, allows to write the orthogonality of the signal subspace with the noise subspace as

$$\mathbf{Q}_{\mathrm{n}}^H\mathbf{a}^{\mathrm{w}}(\theta_{\mathrm{d}}) = \mathbf{Q}_{\mathrm{n,HA}}^H\mathbf{a}_{\mathrm{HA}}^{\mathrm{w}}(\theta_{\mathrm{d}}) + A_{\mathrm{E}}^{\mathrm{w}}(\theta_{\mathrm{d}})\mathbf{q}_{\mathrm{n,E}}^* = \mathbf{0}_{M\times 1}, \tag{14}$$

with $\mathbf{a}_{\mathrm{HA}}^{\mathrm{w}}(\theta_{\mathrm{d}}) = \mathbf{E}_{\mathrm{HA}}\mathbf{a}^{\mathrm{w}}(\theta_{\mathrm{d}})$ and $A_{\mathrm{E}}^{\mathrm{w}}(\theta_{\mathrm{d}}) = \mathbf{e}_{\mathrm{E}}^T\mathbf{a}^{\mathrm{w}}(\theta_{\mathrm{d}})$ and with $(\cdot)^*$ denoting the element-wise complex conjugation operation. We stress that the orthogonality relation in (14) relates *all* pre-whitened ATFs with the *full* noise subspace. Please note that in general $\mathbf{Q}_{\mathrm{n,HA}}^H\mathbf{a}_{\mathrm{HA}}^{\mathrm{w}}(\theta_{\mathrm{d}}) \neq \mathbf{0}_{M\times 1}$.

We propose to exploit the orthogonality relation in (14) for the computation of the pre-whitened ATF $A_{\mathrm{E}}^{\mathrm{w}}(\theta_{\mathrm{d}})$ given $\mathbf{Q}_{\mathrm{n,HA}}$, $\mathbf{q}_{\mathrm{n,E}}$ and $\mathbf{a}_{\mathrm{HA}}^{\mathrm{w}}(\theta_{\mathrm{d}})$ using a least-squares optimization problem. The particular optimization problem that we consider after rearranging (14) is

$$\alpha_{\mathrm{opt}}(\theta_{\mathrm{d}}) = \underset{\alpha}{\operatorname{argmin}} \left\|\mathbf{Q}_{\mathrm{n,HA}}^{-H}\mathbf{q}_{\mathrm{n,E}}^* - \alpha\mathbf{a}_{\mathrm{HA}}^{\mathrm{w}}(\theta_{\mathrm{d}})\right\|_2^2 = -\frac{1}{A_{\mathrm{E}}^{\mathrm{w}}(\theta_{\mathrm{d}})}. \tag{15}$$

The solution to (15) is given by

$$\boxed{\alpha_{\mathrm{opt}}(\theta_{\mathrm{d}}) = \frac{(\mathbf{a}_{\mathrm{HA}}^{\mathrm{w}}(\theta_{\mathrm{d}}))^H\mathbf{Q}_{\mathrm{n,HA}}^{-H}\mathbf{q}_{\mathrm{n,E}}^*}{\left\|\mathbf{a}_{\mathrm{HA}}^{\mathrm{w}}(\theta_{\mathrm{d}})\right\|_2^2} \Rightarrow A_{\mathrm{E}}^{\mathrm{w}}(\theta_{\mathrm{d}}) = \frac{-1}{\alpha_{\mathrm{opt}}(\theta_{\mathrm{d}})}} \tag{16}$$

Thus, using (16) one can complete the vector $\mathbf{a}_{\mathrm{HA}}^{\mathrm{w}}(\theta_{\mathrm{d}})$ by exploiting the orthogonality relation between the pre-whitened ATFs and the noise subspace and obtain $\mathbf{a}^{\mathrm{w}}(\theta_{\mathrm{d}}) = \left[\mathbf{a}_{\mathrm{HA}}^{\mathrm{w}}(\theta_{\mathrm{d}}), A_{\mathrm{E}}^{\mathrm{w}}(\theta_{\mathrm{d}})\right]^T$.

Based on the least squares solution in (16), we propose the following completed set of pre-whitened prototype ATF vectors

$$\boxed{\bar{\mathbf{a}}_{\mathrm{completed}}^{\mathrm{w}}(k,l,\theta_i) = \begin{bmatrix} \bar{\mathbf{a}}_{\mathrm{HA}}^{\mathrm{w}}(k,l,\theta_i) \\ -\dfrac{\left\|\bar{\mathbf{a}}_{\mathrm{HA}}^{\mathrm{w}}(k,l,\theta_i)\right\|_2^2}{(\bar{\mathbf{a}}_{\mathrm{HA}}^{\mathrm{w}})^H(k,l,\theta_i)\hat{\mathbf{Q}}_{\mathrm{n,HA}}^{-H}(k,l)\hat{\mathbf{q}}_{\mathrm{n,E}}^*(k,l)} \end{bmatrix}} \tag{17}$$

Based on (6) and (17), we propose the following completed set of prototype RTF vectors

$$\boxed{\bar{\mathbf{g}}_{\mathrm{completed}}(k,l,\theta_i) = \frac{\hat{\mathbf{\Phi}}_{\mathrm{u}}^{1/2}(k,l)\,\bar{\mathbf{a}}_{\mathrm{completed}}^{\mathrm{w}}(k,l,\theta_i)}{\mathbf{e}_1^T\hat{\mathbf{\Phi}}_{\mathrm{u}}^{1/2}(k,l)\,\bar{\mathbf{a}}_{\mathrm{completed}}^{\mathrm{w}}(k,l,\theta_i)}} \tag{18}$$

Using the completed sets of prototype transfer function vectors in (17) and (18), allows to construct frequency-dependent SPS according to (8) and (9) when used with partially calibrated arrays. Fig. 2 summarizes the novel DOA estimation methods.

V. Experimental Results

For acoustic scenarios with two static speakers in multiple reverberant environments with diffuse-like babble noise, in this section we compare the DOA estimation performance with MUSIC and the RTF vector matching method when using the incomplete sets $\bar{\mathbf{a}}_{\mathrm{HA}}^{\mathrm{w}}(k,l,\theta_i)$ and $\bar{\mathbf{g}}_{\mathrm{HA}}(k,\theta_i)$ and the completed sets $\bar{\mathbf{a}}_{\mathrm{completed}}^{\mathrm{w}}(k,l,\theta_i)$ in (17) and $\bar{\mathbf{g}}_{\mathrm{completed}}(k,l,\theta_i)$ in (18). In Section V-A, we describe the experimental setup and implementations details. In Section V-B, we present and discuss the results.

A. Experimental setup and implementation details

For the experiments, we consider separate recordings of speech and diffuse-like babble noise from the BRUDEX database [18]. The signals have been recorded in a laboratory at the University of Oldenburg with dimensions of about $(7 \times 6 \times 2.7)\,\mathrm{m}^3$ with binaural hearing aids ($M = 4$) on a dummy head and an eMic placed at 36 possible locations which are uniformly distributed in the laboratory. We consider three reverberation environments ('low', 'medium', and 'high'), corresponding to median reverberation times $\mathrm{T}_{60} \approx [310, 510, 1300]\,\mathrm{ms}$. Excluding co-located speakers, we consider a female and a male speaker ($J = 2$, both constantly active, duration = 5s) located at 132 possible two-speaker DOA combinations in the range $[-150:30:180]^\circ$ at a distance of approximately 2m relative to the dummy head. For both speakers we consider equal average broadband speech power across all signals of the microphones of

the hearing aids. The noise component is added to the reverberant speech component after scaling the broadband noise power across all signals of the microphones of the hearing aids to signal-to-noise ratios (SNR) in the range $[-5:5:20]\,\mathrm{dB}$. All microphone signals are assumed to be exchanged without quantization errors and are assumed to be synchronized.

To assess the benefit of DOA estimation with the completed sets $\bar{\mathbf{a}}^{\mathrm{w}}_{\mathrm{completed}}(k,l,\theta_i)$ and $\bar{\mathbf{g}}_{\mathrm{completed}}(k,l,\theta_i)$ over the incomplete sets $\bar{\mathbf{a}}^{\mathrm{w}}_{\mathrm{HA}}(k,l,\theta_i)$ and $\bar{\mathbf{g}}_{\mathrm{HA}}(k,\theta_i)$, we compare the following conditions for MUSIC and the RTF vector matching method:

- H/H: EVD of $\hat{\mathbf{\Phi}}^{\mathrm{w}}_{\mathrm{y,HA}}$ (no eMic) and prototype matching with $\bar{\mathbf{a}}^{\mathrm{w}}_{\mathrm{HA}}(k,l,\theta_i)$ and $\bar{\mathbf{g}}_{\mathrm{HA}}(k,\theta_i)$ (no eMic),
- H+E/H: EVD of $\hat{\mathbf{\Phi}}^{\mathrm{w}}_{\mathrm{y}}$ (with eMic) and prototype matching with $\bar{\mathbf{a}}^{\mathrm{w}}_{\mathrm{HA}}(k,l,\theta_i)$ and $\bar{\mathbf{g}}_{\mathrm{HA}}(k,\theta_i)$ (no eMic),
- H+E/H+E: EVD of $\hat{\mathbf{\Phi}}^{\mathrm{w}}_{\mathrm{y}}$ (with eMic) and prototype matching with $\bar{\mathbf{a}}^{\mathrm{w}}_{\mathrm{completed}}(k,l,\theta_i)$ and $\bar{\mathbf{g}}_{\mathrm{completed}}(k,l,\theta_i)$ (with eMic).

All microphone signals are downsampled to 16kHz. The algorithms are implemented within an STFT framework with 32ms square-root Hann windows with 50 % overlap. We estimate the covariance matrices $\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l)$ and $\hat{\mathbf{\Phi}}_{\mathrm{u}}(k,l)$ for each TF bin using a first order recursion during speech-and-noise periods and noise-only periods, respectively, as

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

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

with smoothing factors α_{y} and α_{u} corresponding to time constants of 250ms and 500ms, respectively. To discriminate speech-and-noise periods from noise-only periods, speech presence probabilities [28] are estimated using the hearing aid microphone signals, averaged and thresholded. Based on the results reported in [25] for the selection of the frequency subset $\mathcal{K}$, we set $\mathrm{CDR}_{\mathrm{thresh}} = -3\mathrm{dB}$ for MUSIC and $\mathrm{CDR}_{\mathrm{thresh}} = -5\mathrm{dB}$ for the RTF vector matching method. The indicator function $\mathbb{1}_j(k,l)$ is estimated as described in [25].

The set of anechoic prototype ATF vectors $\bar{\mathbf{a}}_{\mathrm{HA}}(k,\theta_i)$ is obtained from measured anechoic binaural room impulse responses [29] with an angular resolution of 5° in the range $[-180:5:175]^\circ$ ($I = 72$).

We assess the DOA estimation performance using the following definition of accuracy:

$$\mathrm{ACC} = \frac{1}{JL}\sum_{l=1}^{L} j_{\mathrm{correct}}(l), \tag{21}$$

where j_{correct} denotes the number of speakers for which the DOA is estimated within $\pm 5^\circ$ correctly. For both DOA estimation methods, we average the accuracies over all acoustic scenarios, i.e., DOA combinations, SNRs, and reverberation conditions.

B. Results

Fig. 3 depicts the average localization accuracy for the investigated conditions for the two DOA estimation methods. The horizontal line in the H/H condition (no eMic) shows the average localization accuracy when considering only the hearing aid microphone signals. Considering in addition to these signals also the eMic signal, the violin plots in the H+E/H condition (eMic signal included only in the EVD of $\hat{\mathbf{\Phi}}^{\mathrm{w}}_{\mathrm{y}}$ but not in the sets of prototype transfer function vectors) and in the H+E/H+E condition (eMic signal included in the EVD of $\hat{\mathbf{\Phi}}^{\mathrm{w}}_{\mathrm{y}}$ and in the sets of prototype transfer function vectors) show the distribution of average localization accuracy due to different locations of the eMic. First, it can be observed that for both DOA estimation methods the performance with the H/H condition is similar (about 74%). We interpret this observation as a consequence that the signal

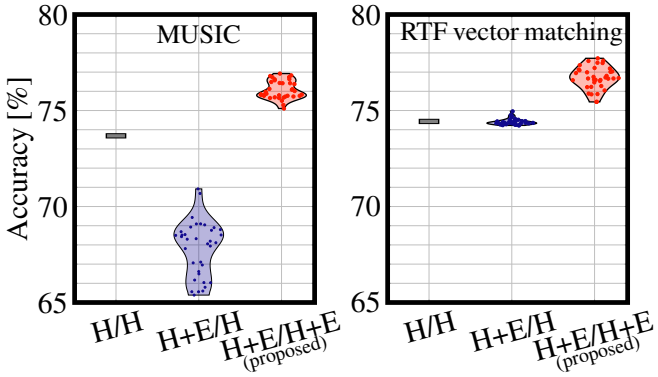

Fig. 3: Average localization accuracy of DOA estimation using MUSIC (left) and the RTF vector matching method (right) with the conditions H/H (no eMic), H+E/H (eMic included only in the EVD of $\hat{\mathbf{\Phi}}^{\mathrm{w}}_{\mathrm{y}}$ but not in the prototype matching), and H+E/H+E (eMic included in the EVD of $\hat{\mathbf{\Phi}}^{\mathrm{w}}_{\mathrm{y}}$ and in the prototype matching).

subspace and the noise subspace are similarly meaningful for DOA estimation since they are obtained from the same covariance matrix $\hat{\mathbf{\Phi}}^{\mathrm{w}}_{\mathrm{y,HA}}$. Second, it can be observed that for the RTF vector matching method there is only a minor performance difference between the H/H and the H+E/H condition. We interpret this result, which is in line with the results reported in [15] and [17], as a consequence of the parallelity of the vectors $\mathcal{P}\{\phi_{\mathrm{s}}\mathbf{a}^{\mathrm{w}}_{\mathrm{HA}}(\theta_{\mathrm{d}})(\mathbf{a}^{\mathrm{w}}_{\mathrm{HA}}(\theta_{\mathrm{d}}))^{H}\}$ and $\mathbf{E}_{\mathrm{HA}}\mathcal{P}\{\phi_{\mathrm{s}}\mathbf{a}^{\mathrm{w}}(\theta_{\mathrm{d}})(\mathbf{a}^{\mathrm{w}}(\theta_{\mathrm{d}}))^{H}\}$. For MUSIC, however, there is a large performance difference between the H/H and the H+E/H condition. We interpret the higher performance at the H/H condition compared to the H+E/H condition as a consequence of the orthogonality relation in (14), which holds for all signals and not just a subset of signals. Also note that in MUSIC the condition H+E/H can be understood as setting $A^{\mathrm{w}}_{\mathrm{E}}(\theta_{\mathrm{d}})$ to 0, which clearly deviates from the optimal solution in (16). Third, considering the H+E/H+E condition (corresponding to the proposed completion procedure), the results clearly show that for both methods DOAs can be estimated more accurately than with the H/H and H+E/H conditions. This result together with those reported in [17] further supports the advantage in exploiting eMic signals for the construction of spatial spectra. Fourth, for both DOA estimation methods the results from the H+E/H+E condition show the highest average localization accuracies for *all* eMic locations. Based on these results, the potential and robustness to the location of the eMic of the proposed procedure for completing sets of prototype transfer functions is clearly demonstrated.

VI. Conclusions

In this paper we estimated DOAs of multiple speakers with partially calibrated microphone arrays, composed of a calibrated binaural hearing aid and a non-calibrated external microphone at an unknown location. We proposed an optimal procedure in the least-squares sense that exploits the external microphone for the completion of sets of prototype transfer function vectors. We compared DOA estimation with the incomplete and completed sets of prototype transfer function vectors for the subspace-based MUSIC and the RTF vector matching method. Experimental results with two speakers in multiple reverberant environments with diffuse-like noise from the BRUDEX database clearly demonstrate that DOAs can be estimated more accurately with the proposed completed sets of prototype transfer function vectors than with incomplete sets. Moreover, we showed that the procedure is robust to the location of the external microphone.

References

[1] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, no. 1, pp. 1–24, Aug. 2017.

[2] 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.

[3] 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, Apr. 2020.

[4] 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.

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

[6] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Berlin, Heidelberg, Germany: Springer, 2001, pp. 157–180.

[7] V. Tourbabin and B. Rafaely, "Direction of arrival estimation using microphone array processing for moving humanoid robots," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2046–2058, Nov. 2015.

[8] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, Mar. 2019.

[9] S. Hafezi, A. H. Moore, and P. A. Naylor, "Spatial consistency for multiple source direction-of-arrival estimation and source counting," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4592–4603, Dec. 2019.

[10] H. Hammer, S. E. Chazan, J. Goldberger, and S. Gannot, "Dynamically localizing multiple speakers based on the time-frequency domain," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–10, Apr. 2021.

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

[12] 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.

[13] D. Fejgin and S. Doclo, "Exploiting an external microphone for binaural RTF-vector-based direction of arrival estimation for multiple speakers," in *Proc. Forum Acusticum*, Turin, Italy, Sep. 2023, pp. 1741–1747.

[14] U. Kowalk and S. Doclo, "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, Sep. 2022, pp. 1–5.

[15] 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.

[16] K. Brümann and S. Doclo, "Steered response power-based direction-of-arrival estimation exploiting an auxiliary microphone," in *Proc. European Signal Processing Conference (EUSIPCO)*, Lyon, France, Aug. 2024, pp. 1–5.

[17] 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, pp. 1–5.

[18] D. Fejgin, W. Middelberg, and S. Doclo, "BRUDEX database: Binaural room impulse responses with uniformly distributed external microphones," in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, Sep. 2023, pp. 126–130.

[19] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[20] 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.

[21] A. L. Swindlehurst and T. Kailath, "A performance analysis of subspace-based methods in the presence of model errors, Part I: The MUSIC algorithm," *IEEE Trans. on Signal Processing*, vol. 40, no. 7, pp. 1758–1774, Jul. 1992.

[22] 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.

[23] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 581–585, May 2014.

[24] K. Scharnhorst, "Angles in complex vector spaces," *Acta Applicandae Mathematica*, vol. 69, no. 1, pp. 95–103, Oct. 2001.

[25] D. Fejgin, E. Hadad, S. Gannot, Z. Koldovský, and S. Doclo, "Comparison of frequency-fusion mechanisms for binaural direction-of-arrival estimation for multiple speakers," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Republic of Korea, Apr. 2024, pp. 731–735.

[26] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, Aug. 2012.

[27] 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.

[28] 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.

[29] 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.