ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446394

COMPARISON OF FREQUENCY-FUSION MECHANISMS FOR BINAURAL DIRECTION-OF-ARRIVAL ESTIMATION FOR MULTIPLE SPEAKERS

Daniel Fejgin[1], *Elior Hadad*[2], *Sharon Gannot*[2], *Zbyněk Koldovský*[3], *Simon Doclo*[1]

[1] Dept. of Medical Physics and Acoustics, Carl von Ossietzky Universität Oldenburg, Germany
[2] Bar-Ilan University, Ramat-Gan, Israel
[3] Technical University of Liberec, Liberec, Czech Republic

ABSTRACT

To estimate the direction of arrival (DOA) of multiple speakers with methods that use prototype transfer functions, frequency-dependent spatial spectra (SPS) are usually constructed. To make the DOA estimation robust, SPS from different frequencies can be combined. According to how the SPS are combined, frequency fusion mechanisms are categorized into narrowband, broadband, or speaker-grouped, where the latter mechanism requires a speaker-wise grouping of frequencies. For a binaural hearing aid setup, in this paper we propose an interaural time difference (ITD)-based speaker-grouped frequency fusion mechanism. By exploiting the DOA dependence of ITDs, frequencies can be grouped according to a common ITD and be used for DOA estimation of the respective speaker. We apply the proposed ITD-based speaker-grouped frequency fusion mechanism for different DOA estimation methods, namely the multiple signal classification, steered response power and a recently published method based on relative transfer function (RTF) vectors. In our experiments, we compare DOA estimation with different fusion mechanisms. For all considered DOA estimation methods, the proposed ITD-based speaker-grouped frequency fusion mechanism results in a higher DOA estimation accuracy compared with the narrowband and broadband fusion mechanisms.

Index Terms— direction of arrival estimation, frequency fusion, speaker grouping, binaural hearing aids

1. INTRODUCTION

In multi-microphone speech communication applications like hearing aids, estimation of the direction of arrival (DOA) of multiple speakers in noisy and reverberant environments is of crucial importance [1]. Due to their popularity, in this paper we specifically consider DOA estimation methods based on prototype transfer function matching. Examples for such methods include the well-established subspace-based multiple signal classification (MUSIC) [2] or power-based steered response power (SRP) methods [3], which both use acoustic transfer function (ATF) vectors, or a recently published method based on matching of anechoic prototype relative transfer function (RTF) vectors [4]. In these methods, frequency-dependent spatial spectra (SPS), i.e., functions of candidate directions, are constructed with the location of peaks likely corresponding to the true speaker DOAs.

To make the DOA estimation of broadband speech sources robust, SPS from different frequencies can be combined. Depending on how the frequency-dependent SPS are combined, frequency fusion mechanisms are categorized into narrowband, broadband, or speaker-grouped [5, 6]. With the narrowband fusion mechanism, i.e., no combination of frequency-dependent SPS, DOAs are first estimated directly for each frequency and subsequently clustered. From these clustered DOA estimates, refined estimates are obtained (e.g., [7]). With the broadband fusion mechanism, DOAs are estimated from a single SPS obtained from pooling frequency-dependent SPS. With the speaker-grouped fusion mechanism, DOAs are estimated from multiple SPS, each obtained from pooling frequency-dependent SPS. As each pooled SPS is associated with a single speaker solely, DOA estimation with the speaker-grouped fusion mechanism may overcome the tending over-emphasis of the dominant speaker and the lobe broadening of the spatial spectrum which may by introduced with the narrowband and broadband fusion mechanisms [5, 6]. However, using the speaker-grouped fusion mechanism requires a speaker-wise grouping of frequencies.

In this paper, we propose an interaural time difference (ITD)-based speaker-grouped frequency fusion mechanism, which we apply to DOA estimation with the MUSIC, SRP and the recently published RTF-vector-matching-based methods. We exploit the DOA dependence of ITDs such that frequencies can be grouped according to a common ITD and, thus, be associated with a single speaker. We compare DOA estimation using the proposed ITD-based speaker-grouped fusion mechanism with the narrowband and broadband fusion mechanisms. To further assess the performance of DOA estimation using the proposed ITD-based speaker-grouped fusion mechanism, we also compare against DOA estimation with separated speakers, where the speakers are separated using successive applications of the independent vector extraction (IVE) algorithm [8]. Experimental results with recorded data from the novel BRUDEX database [9] for multiple reverberant and noisy acoustic scenarios with two static speakers and a binaural hearing aid setup demonstrate the efficiency of the proposed ITD-based speaker-grouped frequency fusion mechanism.

2. SIGNAL MODEL AND NOTATION

We consider a binaural hearing aid setup with M microphones in a noisy and reverberant acoustic scenario with J simultaneously active speakers $S_{1:J}$. We consider stationary DOAs $\theta_{1:J}$ (in the azimuthal plane) and J 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 $m \in \{1,...,M\}$, $k \in \{1,...,K\}$ and $l \in \{1,...,L\}$ denote the index of the microphone, frequency bin, and the frame, respectively, and $X_{m,j}(k,l)$ and $N_m(k,l)$ denote the j-th speech component and

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. The work was also supported by the Israeli Ministry of Science & Technology. The work is also supported by the Department of the Navy, Office of Naval Research Global, through Project No. N62909-23-1-2084.

the noise component in the m-th microphone signal, respectively. Assuming disjoint speaker activity in the STFT domain [10], each time-frequency (TF) bin is dominated by a single speaker across all microphone signals. Stacking all microphone signals into the M-dimensional vector $\mathbf{y}(k,l) = [Y_1(k,l),...,Y_M(k,l)]^T$ with $(\cdot)^T$ denoting the transposition operator, the vector of microphone signals can be approximated as $\mathbf{y}(k,l) \approx \mathbf{x}_{\mathrm{d}}(k,l) + \mathbf{n}(k,l)$, with the vector $\mathbf{x}_{\mathrm{d}}(k,l)$ denoting the dominant speech component and the vector $\mathbf{n}(k,l)$ denoting the noise component, both defined similarly as $\mathbf{y}(k,l)$.

We assume that each speech component j can be split into a direct-path component $\mathbf{x}_j^{\mathrm{DP}}(k,l)$ and a reverberation component $\mathbf{x}_j^{\mathrm{R}}(k,l)$, i.e., $\mathbf{x}_j(k,l) = \mathbf{x}_j^{\mathrm{DP}}(k,l) + \mathbf{x}_j^{\mathrm{R}}(k,l)$. Assuming a multiplicative transfer function for $\mathbf{x}_j^{\mathrm{DP}}(k,l)$ in the STFT domain [11], the direct-path component $\mathbf{x}_{\mathrm{d}}^{\mathrm{DP}}(k,l)$ can be written in terms of the direct-path acoustic transfer function (ATF) vector $\mathbf{a}(k,\theta_{\mathrm{d}})$ (relating the microphone signals with the dominant speaker signal $S_{\mathrm{d}}(k,l)$ with DOA θ_{d}) or equivalently in terms of the direct-path relative transfer function (RTF) vector $\mathbf{g}(k,\theta_{\mathrm{d}})$ (relating the microphone signals with the reference microphone signal $X_{1,\mathrm{d}}(k,l)$) as

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

where the first microphone is considered as the reference microphone (without loss of generality). Condensing the noise and reverberation components into the undesired component $\mathbf{u}(k,l) = \mathbf{n}(k,l) + \mathbf{x}_{\mathrm{d}}^{\mathrm{R}}(k,l)$, where for conciseness the index d of the dominant speaker was omitted in $\mathbf{u}$, the microphone signals can be written as $\mathbf{y}(k,l) = \mathbf{x}_{\mathrm{d}}^{\mathrm{DP}}(k,l) + \mathbf{u}(k,l)$.

Assuming uncorrelated direct-path speech and undesired components, the covariance matrix $\mathbf{\Phi}_{\mathrm{y}}(k,l)$ of the microphone signals can be approximated as

$$\mathbf{\Phi}_{\mathrm{y}}(k,l) = \mathcal{E}\left\{\mathbf{y}(k,l)\mathbf{y}^H(k,l)\right\} \approx \mathbf{\Phi}_{\mathrm{x_d}}^{\mathrm{DP}}(k,l) + \mathbf{\Phi}_{\mathrm{u}}(k,l), \tag{3}$$

where the covariance matrix $\mathbf{\Phi}_{\mathrm{x_d}}^{\mathrm{DP}}(k,l)$ of the direct-path dominant speech component and the covariance matrix $\mathbf{\Phi}_{\mathrm{u}}(k,l)$ of the undesired component are defined similarly as $\mathbf{\Phi}_{\mathrm{y}}(k,l)$ and $(\cdot)^H$ and $\mathcal{E}\{\cdot\}$ denote the complex transposition and expectation operators, respectively.

3. FREQUENCY-DEPENDENT SPATIAL SPECTRA

In this section, we review the construction of frequency-dependent spatial spectra (SPS) for the subspace-based MUSIC [2], the power-based SRP [3, 12] and the RTF-vector-matching-based [4] methods for binaural DOA estimation. For the construction of these frequency-dependent SPS a database of anechoic prototype transfer functions for different candidate directions $\{\theta_i\}_{i=1}^I$ is needed which accounts for head shadow effects. To obtain such a database, analytic diffraction models [13] or measurements can be used.

For the MUSIC and SRP methods, the frequency-dependent SPS are obtained from the estimated covariance matrix of the noisy microphone signals $\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l)$ and a database of anechoic head-related transfer functions $\bar{\mathbf{a}}(k,\theta_i)$. MUSIC is based on the orthogonality between the noise subspace of $\mathbf{\Phi}_{\mathrm{y}}(k,l)$ and the direct-path ATF vector $\mathbf{a}(k,\theta_{\mathrm{d}})$, i.e.,

$$\tilde{p}^{\mathrm{MUSIC}}(k,l,\theta_i) = \frac{1}{\|\hat{\boldsymbol{Q}}_{\mathrm{y,u}}^H(k,l)\bar{\mathbf{a}}(k,\theta_i)\|_2}, \tag{4}$$

where $\hat{\boldsymbol{Q}}_{\mathrm{y,u}}$ denotes the estimated noise subspace of $\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l)$. SRP is based on the maximization of the output power of a beamformer. Accounting for head shadow effects and the signal input power as in [12], the frequency-dependent SPS is given by

$$\tilde{p}^{\mathrm{SRP}}(k,l,\theta_i) = \frac{\bar{\mathbf{a}}^H(k,\theta_i)\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l)\bar{\mathbf{a}}(k,\theta_i)}{\|\bar{\mathbf{a}}(k,\theta_i)\|_2^2\|\mathbf{y}(k,l)\|_2^2}. \tag{5}$$

Similar to [14], in our binaural experiments with MUSIC and SRP we noted the importance of normalizing the values of the frequency-dependent SPS to the range [0,1]. Hence, for binaural DOA estimation we consider the normalized versions $p^{\mathrm{MUSIC}}(k,l,\theta_i)$ and $p^{\mathrm{SRP}}(k,l,\theta_i)$ instead of $\tilde{p}^{\mathrm{MUSIC}}(k,l,\theta_i)$ and $\tilde{p}^{\mathrm{SRP}}(k,l,\theta_i)$ in the following

$$p^{\mathrm{MUSIC}}(k,l,\theta_i) = \frac{\tilde{p}^{\mathrm{MUSIC}}(k,l,\theta_i)}{\max\limits_{\theta_j} \tilde{p}^{\mathrm{MUSIC}}(k,l,\theta_j)} \tag{6}$$

$$p^{\mathrm{SRP}}(k,l,\theta_i) = \frac{\tilde{p}^{\mathrm{SRP}}(k,l,\theta_i)}{\max\limits_{\theta_j} \tilde{p}^{\mathrm{SRP}}(k,l,\theta_j)}. \tag{7}$$

For the RTF-vector-matching-based method, the frequency-dependent SPS is obtained from the comparison between estimated RTF vector $\hat{\mathbf{g}}(k,l)$ and a database of anechoic prototype RTF vectors $\bar{\mathbf{g}}(k,\theta_i)$ for different candidate directions θ_i. Considering the Hermitian angle for the comparison of RTF vectors, the frequency-dependent SPS is obtained as

$$p^{\mathrm{RTF}}(k,l,\theta_i) = -\arccos\left(\frac{|\bar{\mathbf{g}}^H(k,\theta_i)\hat{\mathbf{g}}(k,l)|}{\|\bar{\mathbf{g}}(k,\theta_i)\|_2\|\hat{\mathbf{g}}(k,l)\|_2}\right). \tag{8}$$

To estimate the DOAs $\hat{\theta}_{1:J}(l)$, assuming that the number of speaker J is known, we consider combining the frequency-dependent SPS $p(k,l,\theta_i)$ using the frequency-fusion mechanisms which are discussed in Section 4.

4. COMBINATION OF FREQUENCY-DEPENDENT SPATIAL SPECTRA FOR BINAURAL DOA ESTIMATION

In this section, we first review frequency subset selection criteria and principles of narrowband, broadband, and speaker-grouped frequency-fusion mechanisms for DOA estimation (Sections 4.1 and 4.2). In Section 4.3, we propose for the speaker-grouped frequency-fusion mechanisms a frequency grouping that is based on estimated interaural time differences (ITDs).

4.1. Frequency subset selection

Assuming that DOAs can only be reliably estimated at frequencies where one speaker dominates over all other speakers, noise and reverberation, we restrict DOA estimation to the subset $k \in \mathcal{K}(l)$ which due to time-varying microphone recordings is also time-varying. In the context of DOA estimation, criteria based on, e.g., speaker dominance [15], signal-to-noise ratio (SNR) [7,16], speech onsets [16], or coherence-based quantities such as the coherent-to-diffuse ratio (CDR) [4, 17] have been proposed for frequency subset selection. In this paper, we consider the so-called binaural effective-coherence-based coherent-to-diffuse ratio (CDR) criterion

$$\mathcal{K}(l) = \left\{k : \widehat{\mathrm{CDR}}(k,l) \geq \mathrm{CDR}_{\mathrm{thresh}}\right\}, \tag{9}$$

which is based on a binaural coherence model that accounts for head-shadowing effects [18] and the average coherence between signals from all possible microphone pairs between the left and the right hearing aid. The detailed computation of $\widehat{\mathrm{CDR}}(k,l)$ is described in [4].

4.2. Frequency fusion mechanisms for DOA estimation

With the narrowband fusion mechanism, i.e., no combination of frequency-dependent SPS, the frequency-dependent SPS $p(k,l,\theta_i)$ are maximized directly per frequency, i.e.,

$$\hat{\theta}(k,l) = \operatorname*{argmax}_{\theta_i} p(k,l,\theta_i), \tag{10}$$

where due to the assumed disjoint speaker activity [10] only a single DOA is estimated per frequency. Subsequently, the DOA estimates of the frequency subset $k \in \mathcal{K}(l)$ are combined using, e.g., a histogram operation $\mathcal{H}$ [7] which counts how many DOA estimates correspond to each candidate direction θ_i. The speaker DOAs are estimated as the location of peaks of this histogram, i.e.,

$$\hat{\theta}_{1:J}^{\text{narrow}}(l) = \underset{\theta_i}{\text{findPeaks}}\ \mathcal{H}\Big\{\hat{\theta}(k,l)\Big\}, \quad k \in \mathcal{K}(l). \tag{11}$$

With the broadband frequency-fusion mechanism, the frequency-dependent SPS $p(k,l,\theta_i)$ are combined using a pooling operation, e.g., a summation [4]. The speaker DOAs are estimated as the location of peaks of this pooled SPS, i.e.,

$$\hat{\theta}_{1:J}^{\text{broad}}(l) = \underset{\theta_i}{\text{findPeaks}} \sum_k p(k,l,\theta_i), \quad k \in \mathcal{K}(l). \tag{12}$$

With the speaker-grouped frequency-fusion mechanism, individual frequencies are associated with a single speaker solely, allowing to construct for each individual speaker its own SPS. Let the indicator function $\mathbb{1}_j(k,l)$ denote the association between TF bin (k,l) and the j-th speaker, i.e.,

$$\mathbb{1}_j(k,l) = \begin{cases} 1 & \text{TF bin } (k,l) \text{ associated with speaker } j \\ 0 & \text{else.} \end{cases} \tag{13}$$

Only frequency-dependent SPS $p(k,l,\theta_i)$ that are associated with the j-th speaker are combined using a pooling operation. The j-th speaker DOA is estimated as the location of the maximum of the j-th pooled SPS, i.e., for $j = 1,\ldots,J$

$$\hat{\theta}_j^{\text{group}}(l) = \underset{\theta_i}{\text{argmax}} \sum_k \mathbb{1}_j(k,l)\, p(k,l,\theta_i), \quad k \in \mathcal{K}(l). \tag{14}$$

4.3. ITD-based speaker-grouped frequency grouping

In order to compute the association $\mathbb{1}_j(k,l)$ in (13), we want to exploit the spatial information provided by multiple microphone signals of the binaural hearing aid setup. Since spatially disjoint speakers can be discriminated by their interaural time differences (ITDs), we propose to combine the method for the estimation of ITDs from [19] with the estimation of relative transfer functions (RTFs) and subsequently group frequencies according to common ITDs. It should be noted that ITDs are a popular feature for frequency grouping and binaural DOA estimation [20–22]. However, exploiting ITD-based frequency grouping for DOA estimation using the MUSIC, SRP and RTF-vector-matching-based methods have not been considered.

The main assumption of our ITD-based frequency grouping is that the interaural phase difference between a pair of contra-lateral microphone signals changes linearly with frequency, with the slope of this linear relation corresponding to the ITD which depends on the speaker DOA. For the estimation of ITDs we consider the grid-search approach with candidate ITDs τ_n as in [19], where we compare the phase $\arg\Big\{\hat{G}(k,l)\Big\}$ of the estimated RTF between a pair of contra-lateral microphone signals against a set of phase differences with candidate ITDs τ_n, i.e., $\arg\big\{\mathrm{e}^{\mathrm{i}\omega_k\tau_n}\big\}$, with ω_k denoting the discrete angular frequency at frequency bin with index k. Accounting for the phase wrapping due to the phase extraction operator $\arg\{\cdot\}$ and considering the non-linear transformation $\rho(x) = \mathrm{e}^{\beta\cos(x)}/\mathrm{e}^{\beta}$ with hyperparameter β, a score function $\Psi(k,l,\tau_n)$ in τ_n is constructed, i.e.,

$$\Psi(k,l,\tau_n) = \rho\Big(\arg\Big\{\hat{G}(k,l)\Big\} - \arg\Big\{\mathrm{e}^{\mathrm{i}\omega_k\tau_n}\Big\}\Big). \tag{15}$$

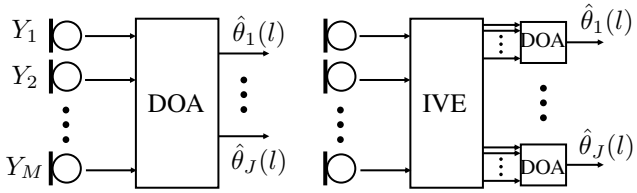

Fig. 1. Comparison of DOA estimation without (left) and with (right) application of a blind speaker-separation algorithm.

Assuming that the number of speakers J is known, the scores are frequency-averaged and the J ITDs, each associated with one of the J speakers, are estimated as the location of peaks of the average score function as

$$\hat{\tau}_{1:J}(l) = \underset{\tau_n}{\text{findPeaks}} \sum_{k \in \mathcal{K}(l)} \Psi(k,l,\tau_n) \tag{16}$$

For the estimation of the indicator function $\mathbb{1}_j(k,l)$ we propose a TF-wise maximization of the similarity scores $\Psi(k,l,\hat{\tau}_j)$ evaluated using each estimated ITD $\hat{\tau}_{1:J}(l)$, i.e.,

$$\boxed{\mathbb{1}_j(k,l) = \begin{cases} 1 & \text{if } j = \underset{j' \in \{1,\ldots,J\}}{\text{argmax}}\ \Psi(k,l,\hat{\tau}_{j'}(l)) \\ 0 & \text{else} \end{cases}} \tag{17}$$

5. SPEAKER-SEPARATION-BASED DOA ESTIMATION

To further assess the performance of DOA estimation using the proposed ITD-based speaker-grouped frequency fusion mechanism, in our experiments in Section 6 we also consider DOA estimation with blindly separated speakers as an upper performance limit. It should be noted that DOA estimation based on blind speaker separation algorithms has previously been investigated in, e.g., [23].

For the speaker separation, we consider successive applications of the independent vector extraction (IVE) algorithm, which assumes independent and known number of speakers and exploits inter-frequency dependencies [8]. The speaker separation is performed offline using the entire mixture. For further details regarding the speaker separation algorithm, we refer the readers to [8] and [24] and references therein. For each separated speaker the DOA is estimated using a method from Section 3 using the broadband fusion mechanism described in Section 4.2. In Fig. 1 the concept of DOA estimation with and without separated speakers is visualized.

6. EXPERIMENTAL RESULTS

For static two-speaker acoustic scenarios in multiple reverberant environments with diffuse-like babble noise, in this section we compare the DOA estimation performance for the MUSIC, SRP, and the RTF-vector-matching-based DOA estimation methods when applied with the narrowband, broadband, and the proposed ITD-based speaker-grouped frequency fusion mechanisms. We also consider DOA estimation with separated speakers using successive applications of the IVE algorithm. In Section 6.1, we describe the experimental setup and implementations details. In Section 6.2 we present and discuss the results in terms of DOA estimation accuracy.

6.1. Experimental setup and implementation details

For the experiments, we consider separate recordings of reverberant speech and diffuse-like babble noise using hearing aid microphones

from the BRUDEX database [9]. We consider three reverberation environments ('low', 'medium', and 'high'), corresponding to median reverberation times $T_{60} \approx [240, 485, 1170]$ ms. Excluding co-located speakers, we consider 132 possible DOA combinations in the range $[-150:30:180]^\circ$ of a female and a male speaker. We cut the speech signals to a duration of approximately 5 s and use an oracle energy-based voice activity detector to obtain constantly active speech signals over the considered duration. We consider equal average broadband speech power across all microphones for both speakers. The noise component is scaled according to SNRs in the range $[-5:5:20]$ dB, where the SNR is set as the ratio of the average broadband speech power of one speaker across all microphones to the average broadband noise power across all microphones.

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

$$\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l) = \alpha_{\mathrm{y}} \hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l-1) + (1-\alpha_{\mathrm{y}}) \mathbf{y}(k,l) \mathbf{y}^H(k,l) \quad (18)$$

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

with smoothing factors α_{y} and α_{u} corresponding to time constants of 250 ms and 500 ms, respectively. With a similar recursion the covariance matrices of the separated speech components $\hat{\mathbf{\Phi}}_{\hat{\mathbf{x}},j}(k,l)$ and noise components $\hat{\mathbf{\Phi}}_{\mathrm{u},j}(k,l)$ are estimated. To discriminate speech-and-noise periods from noise-only periods, a speech presence probability [25] is estimated in all microphones, averaged and thresholded.

We estimate the RTF vector using the state-of-the-art covariance whitening method [26, 27]. Only the entry of the RTF vector that relates the front hearing aid microphone signals with each other is considered for the ITD-based speaker-grouped frequency-fusion mechanism. Psycho-acoustically motivated, we consider candidate ITDs in the range $[-900:1:900]$ µs, and we set the hyperparameter $\beta = 5$. The IVE algorithm is implemented according to [24] with at most 100 iterations.

For the computation of the frequency-dependent SPS, we consider measured anechoic binaural room impulse responses with an angular resolution of 5° in the range $[-180:5:175]^\circ$ [28] from which the sets $\{\bar{\mathbf{a}}(k,\theta_i)\}_{i=1}^{I}$ and $\{\bar{\mathbf{g}}(k,\theta_i)\}_{i=1}^{I}$ of anechoic prototype transfer functions are obtained for $I = 72$ candidate directions.

We assess the DOA estimation performance using the accuracy which we define as:

$$\mathrm{ACC} = \frac{1}{L} \sum_{l=1}^{L} \mathrm{ACC}(l); \quad \mathrm{ACC}(l) = j_{\mathrm{correct}}(l)/J, \quad (20)$$

where $j_{\mathrm{correct}}(l)$ denotes the number of speakers in the l-th frame for which the DOA is estimated within $\pm 5^\circ$ correctly. For each DOA estimation procedure, i.e., the combination of DOA estimation method, frequency fusion mechanism, and (non-) application of speaker separation algorithm, we compute the median of accuracies taken over all acoustic scenes which are specified by the DOA pair, the SNR, and the reverberation condition.

6.2. Results

In Fig. 2, we compare the median DOA estimation accuracy for the considered DOA estimation procedures. For each procedure we also compare the median DOA estimation accuracy when selecting only a subset of frequencies (corresponding to the threshold value

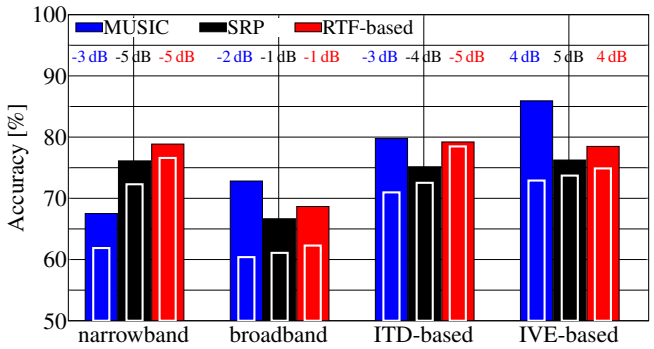

Fig. 2. Median DOA estimation accuracy for the investigated DOA estimation procedures. The white bars denote the median accuracy without frequency subset selection and the number above each bar indicates the CDR threshold value leading to the highest median accuracy.

$\mathrm{CDR}_{\mathrm{thresh}}$ leading to the highest median accuracy, where the respective threshold value is indicated above each bar) against when using the whole frequency set (corresponding to $\mathrm{CDR}_{\mathrm{thresh}} = -\infty$ dB, shown as narrow white bars).

First, it can be observed that similar as in [4] for all considered DOA estimation procedures the frequency subset selection improves DOA estimation, with the largest impact with the MUSIC method. Second, comparing the procedures from the narrowband and broadband frequency fusion methods, variations in the estimation accuracy up to 10 % for the same DOA estimation method can be observed. Third, comparing the proposed ITD-based speaker-grouped frequency fusion mechanism with the narrowband and broadband frequency fusion mechanism, for almost all DOA estimation procedures an increase in localization accuracy can be observed. Comparing the speaker-grouped frequency fusion mechanism against the broadband fusion mechanism, the accuracy can be increased by at least 8 % for all DOA estimation methods. Comparing the speaker-grouped frequency fusion mechanism against the narrowband fusion mechanism, for the MUSIC method the accuracy can be increased by about 12 %, for the RTF-vector-matching-based method a minor improvement can be achieved, and for the SRP method there is no improvement, indicating that both speakers can already be detected in the SPS of the latter methods. Fourth, comparing the procedures using the batch offline IVE algorithm with the procedures using the proposed online ITD-based speaker-grouped frequency fusion mechanism, one sees comparable results for the SRP and RTF-vector-matching-based methods and an improvement of 6 % for the MUSIC method. In summary, our experiments highlight the effectiveness of the proposed ITD-based speaker-grouped frequency fusion mechanism for DOA estimation.

7. CONCLUSION

In this paper, we compared how the combination of frequency-dependent spatial spectra affect DOA estimation for a binaural hearing aid setup. We proposed a speaker-grouped frequency fusion mechanism based on ITDs and compared it against a narrowband and broadband frequency fusion mechanism using the MUSIC, SRP, and RTF-vector-based prototype matching DOA estimation methods. To further assess DOA estimation using the proposed ITD-based speaker-grouped frequency fusion mechanism, we also considered DOA estimation with blindly separated speakers as an upper performance limit. Our experimental results based on noisy and reverberant recordings from the BRUDEX-database demonstrate the effectiveness of the proposed ITD-based speaker-grouped frequency fusion mechanism.

8. REFERENCES

[1] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

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

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

[4] 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.

[5] 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.

[6] S. Thakallapalli, S. V. Gangashetty, and N. Madhu, "NMF-weighted SRP for multi-speaker direction of arrival estimation: robustness to spatial aliasing while exploiting sparsity in the atom-time domain," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–18, Mar. 2021.

[7] E. Hadad and S. Gannot, "Maximum likelihood multi-speaker direction of arrival estimation utilizing a weighted histogram," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 586–590.

[8] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: A deflation approach," *Signal Processing*, vol. 45, no. 1, pp. 59–83, Jul. 1995.

[9] 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.

[10] 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.

[11] 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.

[12] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 515–528, Mar. 2018.

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

[14] 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.

[15] L. Madmoni and B. Rafaely, "Direction of arrival estimation for reverberant speech based on enhanced decomposition of the direct sound," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 131–142, Mar. 2019.

[16] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust DOA estimation of multiple speech sources," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 2287–2291.

[17] 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, no. 1, pp. 1–13, Aug. 2017.

[18] 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, no. 2, pp. 661–664, Aug. 1986.

[19] Z. El Chami, A. Guerin, A. Pham, and C. Servière, "A phase-based dual microphone method to count and locate audio sources in reverberant rooms," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2009, pp. 209–212.

[20] S. Rickard and F. Dietrich, "DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET," in *Proc. Workshop on Statistical Signal and Array Processing*, Pocono Manor, PA, USA, Aug. 2000, pp. 311–314.

[21] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

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

[23] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1490–1503, Aug. 2011.

[24] Z. Koldovský, V. Kautský, P. Tichavský, J. Čmejla, and J. Málek, "Dynamic independent component/vector analysis: Time-variant linear mixtures separable by time-invariant beamformers," *IEEE Trans. on Signal Processing*, vol. 69, pp. 2158–2173, 2021.

[25] 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.

[26] 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.

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

[28] 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.