2022 International Workshop on Acoustic Signal Enhancement (IWAENC) | 978-1-6654-6867-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWAENC53105.2022.9914768

COHERENCE-BASED FREQUENCY SUBSET SELECTION FOR BINAURAL RTF-VECTOR-BASED DIRECTION OF ARRIVAL ESTIMATION FOR MULTIPLE SPEAKERS

Daniel Fejgin and Simon Doclo

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

ABSTRACT

Recently, a method has been proposed to estimate the direction of arrival (DOA) of a single speaker by minimizing the frequency-averaged Hermitian angle between an estimated relative transfer function (RTF) vector and a database of prototype anechoic RTF vectors. In this paper, we extend this method to multi-speaker localization by introducing the frequency-averaged Hermitian angle spectrum and selecting peaks of this spatial spectrum. To construct the Hermitian angle spectrum, we consider only a subset of frequencies, where it is likely that one speaker is dominant. We compare the effectiveness of the generalized magnitude squared coherence and two coherent-to-diffuse ratio (CDR) estimators as frequency selection criteria. Simulation results for estimating the DOAs of two speakers in a reverberant environment with diffuse-like babble noise using binaural hearing devices show that using the binaural effective-coherence-based CDR estimate as a frequency selection criterion yields the best performance.

Index Terms— direction of arrival estimation, relative transfer function, binaural hearing aids, coherent-to-diffuse ratio

1. INTRODUCTION

In many speech communication applications, such as teleconferencing systems and hearing devices, estimating the direction of arrival (DOA) of speech sources in the acoustic scene is of crucial importance [1]. In this paper we specifically consider binaural hearing devices, for which several learning- and non-learning-based methods for multi-speaker DOA estimation have been proposed, e.g., based on interaural time and level differences [2], generalized cross correlation functions [3], or using the subspace-based multiple signal classification (MUSIC) approach [4]. In this paper we consider relative transfer function (RTF) vectors, which have been used for single-speaker binaural DOA estimation [5–8], or for multi-speaker DOA estimation [9] (although not specifically in the context of binaural hearing devices).

In [8] we proposed an RTF-vector-based binaural DOA estimation method for a single speaker by selecting the direction for which the frequency-averaged Hermitian angle between the estimated RTF vector and a database of prototype anechoic RTF vectors is minimized. In this paper we extend the DOA estimation method from [8] to the multi-speaker case. Assuming that the number of speakers J is known, multi-speaker DOA estimation could in principle be simply achieved by selecting J peaks of the frequency-averaged Hermitian angle spectrum. Instead of averaging the Hermitian angle over all frequency bins (as in [8]), we consider only a subset of frequency bins, where it is likely that one speaker dominates over all other speakers, noise, and reverberation. Common criteria to perform frequency bin subset selection in the context of DOA estimation are based on, e.g., signal-to-noise ratio (SNR) [6, 10], onsets [10], and coherence-based quantities such as the coherent-to-diffuse ratio (CDR) [11–13].

In this paper we compare the effectiveness of coherence-based quantities for frequency bin subset selection, more in particular, the generalized magnitude squared coherence from [14] and two recently proposed binaural CDR estimators from [15] which are based on quantities to which we refer to as binaural generalized coherence and effective coherence. This means that the Hermitian angle is only averaged using frequency bins where the estimated binaural coherence-based quantity exceeds a certain threshold. For an acoustic scenario with two static speakers in a reverberant room with diffuse-like babble noise we analyze the performance of the binaural DOA estimation method using the proposed frequency-averaged Hermitian angle spectrum based on simulations with measured binaural room impulse responses (BRIRs). We compare the effectiveness and the influence of the threshold of the coherence-based selection criteria for frequency bin subset selection. Experimental results show that using the binaural effective-coherence-based CDR estimate yields the best DOA estimation performance of all considered cohere-based quantities, both for the Hermitian angle spectrum as well as for the MUSIC spectrum.

2. SIGNAL MODEL AND NOTATION

We consider a binaural hearing aid setup with M microphones, i.e., $M/2$ microphones on each hearing aid. We consider an acoustic scenario with J simultaneously active speakers $S_{1:J}$ located at DOAs $\theta_{1:J}$ (in the azimuthal plane) in a noisy and reverberant environment, where J is assumed to be known. In the STFT domain, the m-th microphone signal can be written as

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

where $k \in \{1,...,K\}$ and $l \in \{1,...,L\}$ denote the frequency bin index and the frame index, respectively, and $X_{m,j}(k,l)$ and $N_m(k,l)$ denote the j-th speech component and the noise component in the m-th microphone, respectively. Assuming one dominant speaker (indexed by d) per time-frequency bin because of the sparseness of speech signals in the STFT domain [16], and stacking all microphone signals in an M-dimensional vector $\mathbf{y}(k,l) = [Y_1(k,l),...,Y_M(k,l)]^T$, where $(\cdot)^T$ denotes transposition, the vector $\mathbf{y}(k,l)$ is given by

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

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

Assuming that the speech component for each (dominant) speaker $\mathbf{x}_d(k,l)$ can be split into a direct-path component $\mathbf{x}_d^{\mathrm{DP}}(k,l)$ and a

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 B 2.

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

reverberant component $\mathbf{x}_d^{\mathrm{R}}(k,l)$ and assuming that the multiplicative transfer function approximation [17] holds for the direct-path component, $\mathbf{x}_d(k,l)$ can be written as

$$\mathbf{x}_d(k,l)=\mathbf{x}_d^{\mathrm{DP}}(k,l)+\mathbf{x}_d^{\mathrm{R}}(k,l)=\mathbf{a}_d(k,\theta_d(l))S_d(k,l)+\mathbf{x}_d^{\mathrm{R}}(k,l) \quad (3)$$

where $\mathbf{a}_d(k,\theta_d(l))$ denotes the direct-path acoustic transfer function (ATF) vector between the dominant speaker with DOA $\theta_d(l)$ and the microphones. Choosing the first microphone as the reference microphone (without loss of generality), $\mathbf{x}_d(k,l)$ can also be written as

$$\mathbf{x}_d(k,l)=\mathbf{g}_d(k,\theta_d(l))X_{1,d}^{\mathrm{DP}}(k,l)+\mathbf{x}_d^{\mathrm{R}}(k,l), \quad (4)$$

where

$$\mathbf{g}_d(k,\theta_d(l))=[1,G_{2,d}(k,\theta_d(l)),...,G_{M,d}(k,\theta_d(l))]^T \quad (5)$$

denotes the direct-path RTF vector and $X_{1,d}^{\mathrm{DP}}(k,l)$ denotes the direct-path speech component of the dominant speaker in the reference microphone. The noise and reverberation components are condensed into the undesired component $\mathbf{u}_d(k,l)=\mathbf{n}(k,l)+\mathbf{x}_d^{\mathrm{R}}(k,l)$ such that $\mathbf{y}(k,l)\approx\mathbf{x}_d^{\mathrm{DP}}(k,l)+\mathbf{u}_d(k,l)$.

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

$$\mathbf{\Phi}_{\mathrm{y}}(k,l)=\mathcal{E}\Big\{\mathbf{y}(k,l)\mathbf{y}^H(k,l)\Big\}=\mathbf{\Phi}_{\mathrm{x}_d}^{\mathrm{DP}}(k,\theta_d(l))+\mathbf{\Phi}_{\mathrm{u}}(k,l), \quad (6)$$

with

$$\mathbf{\Phi}_{\mathrm{x}_d}^{\mathrm{DP}}(k,\theta_d(l))=\mathbf{g}_d(k,\theta_d(l))\mathbf{g}_d^H(k,\theta_d(l))\Phi_{X_d}^{\mathrm{DP}}(k,l), \quad (7)$$

$$\mathbf{\Phi}_{\mathrm{u}}(k,l)=\mathcal{E}\Big\{\mathbf{u}_d(k,l)\mathbf{u}_d^H(k,l)\Big\}, \quad (8)$$

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

3. BINAURAL RTF-VECTOR-BASED DOA ESTIMATION

To estimate the DOAs $\theta_{1:J}$ of all speakers, in this section we propose a multi-speaker extension of the binaural RTF-vector-based single-speaker DOA estimation method from [8]. In Section 3.1 we briefly explain the single-speaker DOA estimation method, where the DOA is estimated by comparing the estimated direct-path RTF vector with a database of prototype anechoic RTF vectors based on the Hermitian angle. In Section 3.2 we propose a method to estimate the DOAs of multiple speakers based on the frequency-averaged Hermitian angle spectrum, where we consider several binaural coherence-based quantities for frequency bin subset selection.

3.1. Single-speaker DOA estimation

To obtain an estimate of the direct-path RTF vector of the dominant speaker in each time-frequency bin, we use the state-of-the-art covariance whitening (CW) method [18]. First, the estimated noisy covariance matrix $\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l)$ is prewhitened using a square-root decomposition (e.g., Cholesky decomposition) of the estimated covariance matrix $\hat{\mathbf{\Phi}}_{\mathrm{u}}(k,l)$ of the undesired component , i.e.,

$$\hat{\mathbf{\Phi}}_{\mathrm{u}}(k,l)=\hat{\boldsymbol{L}}_{\mathrm{u}}(k,l)\hat{\boldsymbol{L}}_{\mathrm{u}}^H(k,l), \quad (9)$$

$$\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}}(k,l)=\hat{\boldsymbol{L}}_{\mathrm{u}}^{-1}(k,l)\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l)\hat{\boldsymbol{L}}_{\mathrm{u}}^{-H}(k,l). \quad (10)$$

The direct-path RTF vector is then estimated as the normalized de-whitened principal eigenvector of the prewhitened noisy covariance matrix, i.e.

$$\hat{\mathbf{g}}(k,l)=\frac{\hat{\boldsymbol{L}}_{\mathrm{u}}(k,l)\mathcal{P}\Big\{\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}}(k,l)\Big\}}{\mathbf{e}_1^T\hat{\boldsymbol{L}}_{\mathrm{u}}(k,l)\mathcal{P}\Big\{\hat{\mathbf{\Phi}}_{\mathrm{y}}^{\mathrm{w}}(k,l)\Big\}}, \quad (11)$$

where $\mathcal{P}\{\cdot\}$ denotes the principal eigenvector of a matrix and $\mathbf{e}_1=[1,0,...,0]^T$ is an M-dimensional selection vector.

For each time-frequency bin, the estimated direct-path RTF vector $\hat{\mathbf{g}}(k,l)$ is compared against a database of prototype anechoic RTF vectors $\bar{\mathbf{g}}(k,\theta_i)$ for different discrete directions $\theta_i, i=1,...,I$ using the so-called Hermitian angle [19], i.e.,

$$p(k,l,\theta_i)=\arccos\left(\frac{|\bar{\mathbf{g}}^H(k,\theta_i)\hat{\mathbf{g}}(k,l)|}{\|\bar{\mathbf{g}}(k,\theta_i)\|_2\|\hat{\mathbf{g}}(k,l)\|_2}\right). \quad (12)$$

The DOA of the speaker is then estimated as the direction for which the Hermitian angle averaged over all frequencies (except DC) is maximized, i.e.,

$$P'(l,\theta_i)=-\sum_{k=2}^{K}p(k,l,\theta_i), \quad \hat{\theta}_1(l)=\mathrm{argmax}_{\theta_i}P'(l,\theta_i). \quad (13)$$

3.2. Multi-speaker DOA estimation

When J speakers are simultaneously active, the DOAs $\theta_{1:J}$ could in principle be estimated by selecting J peaks of the frequency-averaged Hermitian angle $P'(l,\theta_i)$ in (13). However, it should be realized that not all time-frequency bins are dominated by one speaker. Aiming at including only time-frequency bins where the estimated RTF vector in (11) is a good estimate for the direct-path RTF vector in (5) (of one of the speakers), we consider only a subset $\mathcal{K}(l)$ of frequency bins, for which it is likely that the direct-path of one speaker dominates over all other speakers, noise and reverberation. We define the frequency-averaged Hermitian angle spectrum as

$$\boxed{P(l,\theta_i)=-\sum_{k\in\mathcal{K}(l)}p(k,l,\theta_i)} \quad (14)$$

The DOAs $\hat{\theta}_{1:J}(l)$ are estimated by determining the J peaks of this spatial spectrum (assuming J be known).

To determine the subset $\mathcal{K}(l)$, several selection criteria have been proposed [6, 10–13], many of which are coherence-based. More in particular, in this paper we consider the generalized magnitude squared coherence (GMSC) [14] as well as two binaural CDR estimates presented in [15] to which we refer to as binaural generalized-coherence-based CDR estimate and binaural effective-coherence-based CDR estimate.

According to [14], the generalized coherence generalizes the notion of coherence to $M\geq 2$ microphone signals and is defined as

$$\widehat{\mathrm{GC}}(k,l)=\frac{\lambda_{\max}\Big\{\widehat{\mathbf{\Gamma}}_{\mathrm{y}}(k,l)\Big\}-1}{M-1}, \quad (15)$$

with $\widehat{\mathbf{\Gamma}}_{\mathrm{y}}(k,l)$ containing the estimated coherence between the microphone signals, i.e.,

$$\widehat{\Gamma}_{\mathrm{y}_{i,j}}(k,l)=\hat{\Phi}_{\mathrm{y}_{i,j}}(k,l)/\sqrt{\hat{\Phi}_{\mathrm{y}_{i,i}}(k,l)\ \hat{\Phi}_{\mathrm{y}_{j,j}}(k,l)} \quad (16)$$

and $\lambda_{\max}\{\cdot\}$ denoting the principal eigenvalue of a matrix. The generalized magnitude squared coherence (GMSC) is then obtained as

$$\widehat{\mathrm{GMSC}}(k,l)=\Big|\widehat{\mathrm{GC}}(k,l)\Big|^2. \quad (17)$$

In [15] two binaural CDR estimates have been proposed. The first estimate, to which we refer to as binaural generalized-coherence-based CDR estimate, is defined as

$$\widehat{\mathrm{CDR}}_1(k,l) = \frac{\widetilde{\mathrm{GC}}_{\mathrm{u}}(k) - \widehat{\mathrm{GC}}(k,l)}{\widehat{\mathrm{GC}}(k,l) - 1}, \tag{18}$$

where the (time-invariant) generalized coherence of the undesired component $\widetilde{\mathrm{GC}}_{\mathrm{u}}(k)$ is obtained similarly as $\widehat{\mathrm{GC}}(k,l)$ in (15), but using $\widetilde{\mathbf{\Gamma}}_{\mathrm{u}}(k)$ instead of $\widehat{\mathbf{\Gamma}}_{\mathrm{y}}(k,l)$. The model coherence matrix $\widetilde{\mathbf{\Gamma}}_{\mathrm{u}}(k)$ models the coherence of the undesired component $\mathbf{u}_d(k,l)$ as a diffuse sound field, assuming that both the noise component $\mathbf{n}(k,l)$ as well as the reverberation component $\mathbf{x}_d^{\mathrm{R}}(k,l)$ can be modeled as a diffuse sound field. Depending on the considered microphone pair, either a free-field sinc-model [20] for microphones i and j or a modified sinc-model accounting for head shadow effects [21] for microphones i' and j' is employed

$$\widetilde{\Gamma}_{\mathrm{u}_{i,j}}(k) = \mathrm{sinc}\left(\frac{\omega_k d_{i,j}}{c}\right), \tag{19}$$

$$\widetilde{\Gamma}_{\mathrm{u}_{i',j'}}(k) = \mathrm{sinc}\left(\alpha \frac{\omega_k d_{i',j'}}{c}\right) \frac{1}{\sqrt{1 + \left(\beta \frac{\omega_k d_{i',j'}}{c}\right)^4}}, \tag{20}$$

where ω_k denotes the discrete angular frequency, $d_{i,j}$ denotes the distance between microphones i and j, c denotes the speed of sound, and $\alpha = 0.5$ and $\beta = 2.2$ [21].

The second estimate, to which we refer to as binaural effective-coherence-based CDR estimate, is defined as:

$$\widehat{\mathrm{CDR}}_2(k,l) = f\left(\widehat{\Gamma}_{\mathrm{y,eff}}(k,l), \widetilde{\Gamma}_{\mathrm{u}_{i',j'}}(k)\right), \tag{21}$$

$$\widehat{\Gamma}_{\mathrm{y,eff}}(k,l) = \frac{1}{|\mathcal{M}|} \sum_{i,j \in \mathcal{M}} \widehat{\Gamma}_{\mathrm{y}_{i,j}}(k,l), \tag{22}$$

where the effective coherence $\widehat{\Gamma}_{\mathrm{y,eff}}(k,l)$ in (22) represents the average coherence between all possible microphone pairs between the left and the right hearing aid (denoted as the microphone set $\mathcal{M}$), and the CDR functional f in (27) has been introduced in [22].

Given the coherence-based quantities $\widehat{\mathrm{GMSC}}(k,l)$, $\widehat{\mathrm{CDR}}_1(k,l)$, and $\widehat{\mathrm{CDR}}_2(k,l)$, we propose to define the subset $\mathcal{K}(l)$ of frequency bins to be included in the computation of the Hermitian angle spectrum in (14) as

$$\boxed{\begin{aligned} \mathcal{K}_{\mathrm{GMSC}}(l) &= \left\{k : \widehat{\mathrm{GMSC}}(k,l) \geq \mathrm{GMSC}_{\min}\right\} & (23) \\ \mathcal{K}_{\mathrm{CDR}_{1,2}}(l) &= \left\{k : \widehat{\mathrm{CDR}}_{1,2}(k,l) \geq \mathrm{CDR}_{\min}\right\} & (24) \end{aligned}}$$

where $\mathrm{GMSC}_{\min}$ and $\mathrm{CDR}_{\min}$ denote frequency- and frame-independent thresholds.

3.3. Baseline method: MUSIC

As baseline method for multi-speaker DOA estimation we consider MUSIC [4], which is based on the orthogonality between the acoustic transfer function vector $\mathbf{a}_d(k,\theta_d(l))$ and the noise subspace of $\mathbf{\Phi}_{\mathrm{y}}(k,l)$ (see (3) and (6)). The narrowband MUSIC cost function is given by

$$p^{(\mathrm{MUSIC})}(k,l,\theta_i) = \frac{1}{\|\bar{\mathbf{a}}^H(k,\theta_i)\hat{\boldsymbol{Q}}_{\mathrm{y,u}}(k,l)\|_2^2}. \tag{25}$$

with $\bar{\mathbf{a}}(k,\theta_i)$ denoting the prototype anechoic acoustic transfer function vector for direction θ_i and $\hat{\boldsymbol{Q}}_{\mathrm{y,u}}(k,l)$ denoting the estimated noise subspace of $\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l)$. Performing the incoherent frequency averaging method as described in [23] and considering frequency bin subset selection as in Section 3.2, the frequency-averaged normalized MUSIC spectrum is defined as

$$P^{(\mathrm{MUSIC})}(l,\theta_i) = \sum_{k \in \mathcal{K}(l)} \frac{p^{(\mathrm{MUSIC})}(k,l,\theta_i)}{\max_{\theta_j} p^{(\mathrm{MUSIC})}(k,l,\theta_j)}. \tag{26}$$

The DOAs $\hat{\theta}_{1:J}^{(\mathrm{MUSIC})}(l)$ are estimated by determining the J peaks of this spatial spectrum (assuming J to be known).

4. EXPERIMENTAL RESULTS

For an acoustic scenario with two static speakers in a reverberant room with diffuse-like babble noise, in this section we compare the performance of the coherence-based selection criteria discussed in Section 3.2, both for the Hermitian angle spectrum as well as for the baseline MUSIC spectrum. The experimental setup and implementation details of the algorithms are described in Section 4.1. The results in terms of localization accuracy are presented and discussed in Section 4.2.

4.1. Experimental setup and implementation details

To simulate the binaural microphone signals, we use measured binaural room impulse responses (BRIRs) from the `office_i` scenario (reverberation time $T_{60} \approx 300$ms) of the database of [24]. This database contains measured BRIRs for DOAs in the range $[-80°, 80°]$ with an angular resolution of $5°$. Although the used hearing aids contain three microphone each, we only consider the front and rear microphones on each hearing aid ($M = 4$). We simulate several static two-speaker scenarios ($J = 2$), where for each possible DOA combination with a minimum angular spacing of $15°$ (in total 930 DOA combinations) clean speech signals (male, female) from the DNS Challenge dataset [25] are convolved with the corresponding BRIRs. The speech signals are constantly active and are approximately 4s long. The average broadband speech power across all microphones is set to the same value for both speakers. Diffuse-like multi-channel babble noise is generated using the method in [26] and added to the speech components of the microphone signals. The signal-to-noise ratio (SNR) is set to $\{0,5,20\}$ dB, where the SNR is defined as the average broadband speech power of one speaker across all microphones to the average broadband noise power across all microphones. The microphone signals are simulated at a sampling rate of 16kHz.

The simulated microphone signals are processed in the STFT domain using 32 ms square-root Hann windows with 50 % overlap. The anechoic BRIRs from [24] with an angular resolution of $5°$ in the range $[-180°, 175°]$ are used to generate the database of prototype anechoic ATF vectors $\{\bar{\mathbf{a}}(k,\theta_i)\}_{i=1}^{I}$ and RTF vectors $\{\bar{\mathbf{g}}(k,\theta_i)\}_{i=1}^{I}$, with $I = 72$. For each time-frequency bin the noisy covariance matrix $\mathbf{\Phi}_{\mathrm{y}}(k,l)$ and

$$f\left(\widehat{\Gamma}_{\mathrm{y,eff}}, \widetilde{\Gamma}_{\mathrm{u}_{i',j'}}\right) = \frac{\widetilde{\Gamma}_{\mathrm{u}_{i',j'}} \Re\{\widehat{\Gamma}_{\mathrm{y,eff}}\} - |\widehat{\Gamma}_{\mathrm{y,eff}}|^2 - \sqrt{\widetilde{\Gamma}^2_{\mathrm{u}_{i',j'}} \Re\{\widehat{\Gamma}_{\mathrm{y,eff}}\}^2 - \widetilde{\Gamma}^2_{\mathrm{u}_{i',j'}} |\widehat{\Gamma}_{\mathrm{y,eff}}|^2 + \widetilde{\Gamma}^2_{\mathrm{u}_{i',j'}} - 2\widetilde{\Gamma}_{\mathrm{u}_{i',j'}} \Re\{\widehat{\Gamma}_{\mathrm{y,eff}}\} + |\widehat{\Gamma}_{\mathrm{y,eff}}|^2}}{|\widehat{\Gamma}_{\mathrm{y,eff}}|^2 - 1} \tag{27}$$

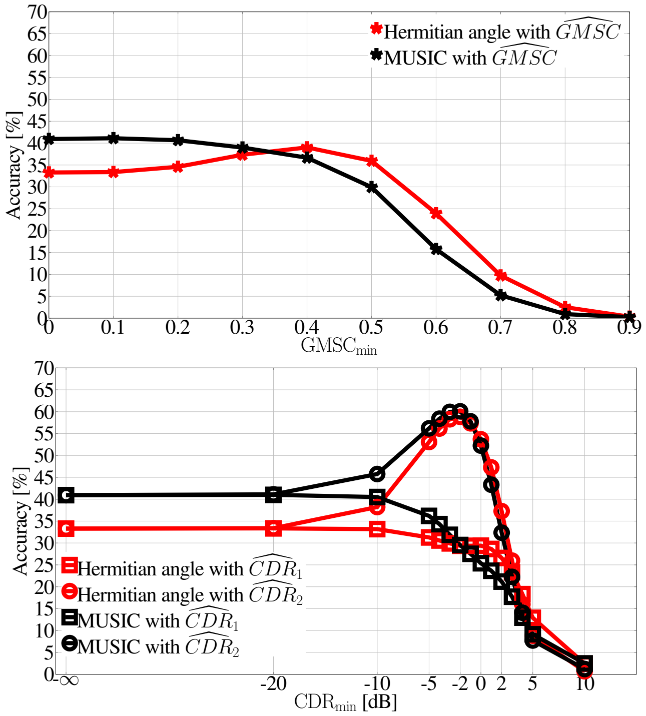

Fig. 1. Average localization accuracy using the Hermitian angle spectrum (red) and MUSIC spectrum (black), with frequency bin subset selection based on generalized magnitude squared coherence (top) or coherent-to-diffuse ratio (bottom).

the undesired covariance matrix $\mathbf{\Phi}_{\mathrm{u}}(k,l)$ are estimated using recursive smoothing during speech-and-noise periods and noise-only periods, respectively as

$$\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l) = \alpha_{\mathrm{y}}\hat{\mathbf{\Phi}}_{\mathrm{y}}(k,l-1) + (1-\alpha_{\mathrm{y}})\mathbf{y}(k,l)\mathbf{y}^H(k,l) \qquad (28)$$

$$\hat{\mathbf{\Phi}}_{\mathrm{u}}(k,l) = \alpha_{\mathrm{u}}\hat{\mathbf{\Phi}}_{\mathrm{u}}(k,l-1) + (1-\alpha_{\mathrm{u}})\mathbf{y}(k,l)\mathbf{y}^H(k,l), \qquad (29)$$

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 periods and noise-only periods are determined based on the thresholded speech presence probability [27], averaged over all microphones.

Performance is assessed in terms of the localization accuracy, i.e. the percentage of correctly localized frames. Similarly as in [9], we consider a frame to be correctly localized only if both estimated DOAs are within $\pm 5^\circ$ of the true DOAs.

4.2. Results

For the two-speaker scenarios described in Section 4.1, we investigate the localization accuracy of the multi-speaker DOA estimation methods proposed in Section 3.2. More in particular, we compare the influence of selecting frequency bins based on several coherence-based quantities (GMSC and CDR). Since it is unrealistic to assume that the thresholds in (23) and (24) can be chosen scenario-dependent, the localization accuracy is averaged over all considered DOA combinations and SNRs (see Section 4.1).

For the Hermitian angle spectrum in (14) and the MUSIC spectrum in (26), Fig. 1 depicts the average localization accuracy for different thresholds of either the generalized magnitude squared coherence $\widehat{\mathrm{GMSC}}$ in (23) or the coherent-to-diffuse ratios $\widehat{\mathrm{CDR}}_1$ and $\widehat{\mathrm{CDR}}_2$ in (24). For all selection criteria, a small threshold corresponds to selecting many frequency bins, whereas a large threshold corresponds to selecting few frequency bins. When the threshold is equal to zero ($-\infty\,$dB), all frequency bins are selected. For the selection criteria $\widehat{\mathrm{GMSC}}$ and $\widehat{\mathrm{CDR}}_1$, it can be observed that setting the threshold $\mathrm{GMSC}_{,\mathrm{min}} > 0$ and $\mathrm{CDR}_{\mathrm{min}} > 0$ does not enable to significantly increase the localization accuracy compared to selecting all frequency bins, both for the Hermitian angle spectrum as well as for the MUSIC spectrum. In contrast, for the selection criterion $\widehat{\mathrm{CDR}}_2$ based on the binaural effective coherence a significant influence of the CDR threshold can be observed, which is in line with [11–13]. DOA estimation using either the Hermitian angle spectrum or the MUSIC spectrum works best when choosing a threshold of $\mathrm{CDR}_{\mathrm{min}} \approx -2\mathrm{dB}$. Using this threshold, the localization accuracy can be increased from about $35\,\%$ to $60\,\%$ for the Hermitian angle spectrum and from about 40% to 60% for the MUSIC spectrum. When considering the SNR-dependent localization accuracies (not depicted here) for the $\widehat{\mathrm{CDR}}_2$ selection criterion, there is always a distinct peak at $\mathrm{CDR}_{\mathrm{min}} \approx -2\mathrm{dB}$ implying an SNR-independent optimal threshold value. Although the proposed Hermitian angle spectrum as a functional for DOA estimation does not outperform the MUSIC spectrum (which for the estimation of the noise subspace requires an eigenvalue decomposition), it represents a viable alternative, especially given the fact that low-complexity RTF vector estimation methods have been proposed assuming the availability of one or more external microphones [28, 29].

5. CONCLUSIONS

In this paper we proposed an extension of a recently proposed RTF-vector-based DOA estimation method for a single speaker to RTF-vector-based DOA estimation for multiple speakers by introducing the Hermitian angle spectrum. To construct this spatial spectrum, we consider only a subset of frequency bins, where it is likely that one speaker dominates over all other speakers, noise, and reverberation. In this paper we compared the effectiveness of the generalized magnitude squared coherence, the binaural generalized-coherence-based estimate of the CDR, and the binaural effective-coherence-based estimate of the CDR as criteria for frequency bin subset selection. Using measured BRIRs, we simulated acoustic scenarios with two static speakers in a reverberant room with diffuse-like babble noise. Simulation results for DOA estimation using binaural hearing devices show no significant increase in the localization accuracy when using either the generalized magnitude squared coherence or the binaural generalized-coherence-based estimate of the CDR as selection criteria compared to selecting all frequency bins. In contrast, the localization accuracy can be significantly increased when using the binaural effective-coherence-based estimate of the CDR as a selection criterion. Using the optimal threshold value of about -2dB, enables to increase the average localization accuracy for the proposed Hermitian angle spectrum as a functional for DOA estimation from about 35% to 60% compared to when selecting all frequency bins. Using the same optimal threshold value when using the MUSIC spectrum as a functional for DOA estimation, localization accuracy can be increased from about 40% to 60% compared to when selecting all frequency bins. For the binaural effective-coherence-based estimate of the CDR as a frequency bin subset selection criterion, experimental results also imply an optimal threshold value that is independent of the SNR.

6. REFERENCES

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

[2] 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.

[3] 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.

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

[5] 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.

[6] X. Li, R. Horaud, L. Girin, and S. Gannot, “Local relative transfer function for sound source localization,” in *Proc. European Signal Processing Conference (EUSIPCO)*, Nice, France, Aug. 2015, pp. 399–403.

[7] B. Yang, H. Liu, and X. Li, “Learning deep direct-path relative transfer function for binaural sound source localization,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 3491–3503, Oct. 2021.

[8] 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.

[9] 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.

[10] 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.

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

[12] C. Evers, E. A. P. Habets, S. Gannot, and P. A. Naylor, “DoA reliability for distributed acoustic tracking,” *IEEE Signal Processing Letters*, vol. 25, no. 9, pp. 1320–1324, Sep. 2018.

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

[14] D. Ramirez, J. Via, and I. Santamaria, “A generalization of the magnitude squared coherence spectrum for more than two signals: definition, properties and estimation,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Mar. 2008, pp. 3769–3772.

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

[16] 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.

[17] 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.

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

[20] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, “Measurement of correlation coefficients in reverberant sound fields,” *The Journal of the Acoustical Society of America*, vol. 27, no. 6, pp. 1072–1077, Nov. 1955.

[21] 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.

[22] A. Schwarz and W. Kellermann, “Coherent-to-diffuse power ratio estimation for dereverberation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, Jun. 2015.

[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] 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.

[25] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, “The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results,” in *arxiv.2005.13981*, May 2020.

[26] E. A. P. Habets, I. Cohen, and S. Gannot, “Generating nonstationary multisensor signals under a spatial coherence constraint,” *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[27] 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.

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

[29] 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.