SPEAKER LOCALIZATION USING FROBENIUS NORM WITH A FOCUS ON CLOSE SPEAKER AND NOISE SOURCE

Ofer Schwartz
Sound Business Unit CEVA Inc.
Herzelia-Pituah, Israel
Ofer.Schwartz@ceva-dsp.com

***Abstract*—Speaker direction of arrival (DOA) estimation is an important task for beamforming-based noise reduction and camera steering. Common DOA estimation techniques suffer from biased DOA estimates when the speaker is close to the noise source. In this paper, a novel speaker DOA estimation is presented based on Frobenius norm minimization. In our model, multiple possible speakers are located in each one of a predefined set of candidate DOAs. Instead of estimating the DOA, the power spectral density (PSD)s of the speakers are mutually estimated and the dominant DOA is determined by the speaker with the maximal PSD. The PSDs estimation task is then employed by minimizing the Frobenius norm of the matrix-difference between the estimated PSD matrix of the received signals and the model-matrix described the multiple speaker presence. An experimental study demonstrates the benefits of the proposed Frobenius-based DOA algorithm in simulated dataset w.r.t. a maximum likelihood (ML) based DOA estimator, especially when the speaker is angularly close to the noise source.**

I. Introduction

Online speaker localization is required in many applications, including beamforming, camera steering, multi-speaker separation, navigation, and target acquisition. This task becomes challenging when additive directional inference sources are captured by the microphone array.

In this paper, the DOA estimation problem is attributed. In the audio-processing community, baseline DOA estimators are based on the generalized cross correlation (GCC) [1] or the multiple signals classification (MUSIC) algorithm [2]. These techniques are not optimal in the presence of directional inference. In [3], the ML estimator of multiple-source DOAs in the presence of colored noise is derived where the number of assumed speakers is confined to the number of microphones. Some papers use the sparsity assumption of the speech nature [4] and assume a single speaker at each frequency bin from an overall number of active speakers. In [5], [6], a single dominant speaker in each time-frequency (TF) bin was assumed, and the interaural phase difference (IPD)s from all TF bins were clustered into groups associated with a candidate speaker using the estimate maximize (EM)-Mixture of Gaussians (MoG) model. The DOA of the active speakers was estimated using the groups with the highest probability. In [6]–[9], the microphone signal vectors from each TF bin were clustered in same manner while the noise was implicitly modeled. In [9], it was shown that the dominant DOA estimator is obtained by looking after the DOA with the maximum posterior signal-to-noise ratio (SNR) at the output of the minimum variance distortionless response (MVDR) beamformer.

Based on our examination, these algorithms suffer from bias in the DOA estimation in cases where the speaker DOA is close to the noise source DOA due to the maximum posterior SNR criterion. In cases where the noise source is close to the speaker from one side, the DOA with the higher posterior SNR might be biased to the other side of the speaker. This property of the DOA estimators can be problematic for beamforming dedicated for ASR (because bias from the actual DOA of the speaker can cause speech distortion) or for camera steering. It is claimed that this problem is caused by only a single speaker being assumed in each TF bin. Thus, the obtained DOA estimate is the DOA with the maximum posterior SNR.

In this paper, it is proposed to assume multiple activity of possible speakers from all possible DOAs and estimate their PSDs. Then, the dominant DOA can be determined by the DOA associated with the maximum PSD. By this criterion, the actual speaker PSD may be the maximal even when the noise source is closed to the speaker and the bias problem is avoided. However, using an ML estimator (as in [3]), the number of speakers is mathematically confined to the number of microphones, which is a problematic restriction for a low number of available microphones. In [10]–[14], the authors estimated the speech and/or the reverberation and/or the ambient noise PSD by minimizing the Frobenius norm of the difference between the received signals' PSD matrix and its statistical matrix-model. In this paper, the PSDs of multiple possible speakers are estimated using Frobenius norm minimization. Thus, there is no mathematical restriction on the number of speakers and a low resolution of DOA search can be made. Additionally, relative to the ML estimation of the DOA, no matrix inversion is required and therefore the computational burden can be reduced. In the experimental section, which consists of simulated microphone signals, it is shown that the DOA estimates of the proposed algorithm have less mean absolute error (MAE) relative to the ML-based DOA estimator in cases where the speaker and the noise source are angularly close.

II. Signal Model

Consider N microphone observations consisting of reverberant speech and additive noise. The speaker beams from DOA θ_S, which can be chosen from a set of predefined DOA candidates with the required resolution $\theta_S \in [0^o : \beta^o : 360^o]$ and overall $J = \frac{360}{\beta}$ DOA candidates. Although the speaker beams only from DOA θ_S, multi-activity of speakers is assumed from each possible DOA. The i-th microphone observation can then be expressed as:

$$Y_i(m,k) = \sum_{j}^{J} X_{j,i}(m,k) + V_i(m,k), \tag{1}$$

where $Y_i(m,k)$ denotes the i-th microphone observation with time-index m and frequency index k, $X_{j,i}(m,k)$ denotes the j-th speech as observed at the i-th microphone, and $V_i(m,k)$ denotes the ambient noise. Here $X_{j,i}(m,k)$ is modeled as a multiplication of the speech $X_{j,1}(m,k)$ (as received by the first microphone that was arbitrarily chosen as the reference microphone) and the relative direct-path transfer function of the i-th microphone $G_i(\theta_j,k)$, i.e.:

$$X_{j,i}(m,k) = G_i(\theta_j,k) X_{j,1}(m,k). \tag{2}$$

The transfer function $G_i(\theta_j,k)$ is a pure phase depending on the time difference of arrival between the i-th microphone and the first microphone:

$$G_i(\theta_j,k) = \exp\left(-\iota \frac{2\pi k}{K} \frac{\tau_i(\theta_j)}{T_\mathrm{s}}\right), \tag{3}$$

where $\tau_i(\theta_j)$ is the time difference of arrival (TDOA) between the i-th microphone and first microphone of the acoustic wave that comes from DOA θ_j, T_s is the sampling time, and K is the number of frequency bins. Considering only the horizontal plane, and given the two-dimensional positions of the microphones, the TDOA $\tau_i(\theta_j)$ is given by:

$$\tau_i(\theta_j) = \frac{1}{c} \cdot [\cos(\theta_j) \quad \sin(\theta_j)] (\mathbf{p}_i - \mathbf{p}_1), \tag{4}$$

where c is the sound velocity, $\mathbf{p}_i$ is the horizontal position of microphone i, and θ_j is the DOA of speaker j. The N microphone signals can be concatenated in a vector form:

$$\begin{aligned}\mathbf{y}(m,k) &= \sum_j \mathbf{g}(\theta_j,k) X_{j,1}(m,k) + \mathbf{v}(m,k) \\ &= \mathbf{G}(k)\mathbf{x}(m,k) + \mathbf{v}(m,k)\end{aligned} \tag{5}$$

where:

$$\begin{aligned}\mathbf{y}(m,k) &= \left[\begin{array}{ccc} Y_1(m,k) & \dots & Y_N(m,k) \end{array}\right]^\mathrm{T}, \\ \mathbf{g}(\theta_j,k) &= \left[\begin{array}{ccc} G_1(\theta_j,k) & \dots & G_N(\theta_j,k) \end{array}\right]^\mathrm{T}, \\ \mathbf{G}(k) &= \left[\begin{array}{ccc} \mathbf{g}(\theta_1,k) & \dots & \mathbf{g}(\theta_J,k) \end{array}\right], \\ \mathbf{x}(m,k) &= \left[\begin{array}{ccc} X_{1,1}(m,k) & \dots & X_{J,1}(m,k) \end{array}\right]^\mathrm{T}, \\ \mathbf{v}(m,k) &= \left[\begin{array}{ccc} V_1(m,k) & \dots & V_N(m,k) \end{array}\right]^\mathrm{T}.\end{aligned}$$

The speech signals are modeled as a complex-Gaussian process with $X_{j,1}(m,k) \sim \mathcal{N}_C(X_{j,1}(m,k), \phi_{X_j}(m,k))$ where $\phi_{X_j}(m,k)$ is the PSD of the j-th speaker and:

$$\mathcal{N}^C(\mathbf{z},\Phi) = \frac{1}{\pi^N |\Phi|} \exp\left(-\mathbf{z}^\mathrm{H}\, \Phi^{-1} \mathbf{z}\right), \tag{6}$$

where $\mathbf{z}$ denotes a Gaussian vector, Φ is a PSD matrix, and $|\cdot|$ denotes the matrix-determinant operation. The PSD matrix of the noise denoted by $\Phi_\mathbf{v}(k)$ is assumed to be time-invariant and known in advance (or can be accurately estimated during speech-absent periods). Accordingly, the observed signal vector $\mathbf{y}(m,k)$ is also a Gaussian stochastic vector with the probability density function (p.d.f.):

$$\mathbf{y}(m,k) \sim \mathcal{N}_C\left(\mathbf{y}(m,k), \mathbf{G}(k)\Phi_X(m,k)\mathbf{G}^H(k) + \Phi_\mathbf{v}(k)\right). \tag{7}$$

The PSD matrix of the speech is modeled as a diagonal matrix $\Phi_X(m,k) = \mathrm{Diag}\left[\begin{array}{ccc} \phi_{X_1}(m,k) & \dots & \phi_{X_J}(m,k) \end{array}\right]$, due to the lack of correlation between different speakers.

The goal of this work is to estimate the dominant speaker DOA θ_S. For the sake of comparison, two ML-based DOA estimators are first derived for 1) a single-speaker activity model where the dominant DOA θ_S is estimated directly and 2) a multi-speaker activity model where the speech power for each speaker $\phi_{X_j}(m,k)$ is estimated and the dominant DOA is determined by the speaker with the maximum power. Note that the latter estimator is restricted to N speakers.

Then, the proposed way is presented, which estimates the speech powers by minimizing the Frobenius norm of the matrix-difference between the estimated PSD matrix of the received signals and the model-matrix. This estimator has no restriction for the number of speakers.

III. Dominant DOA estimation

In this section, the estimators for the dominant DOA are derived. Whenever possible, the frequency k and time m indexes are omitted for brevity.

A. ML-based DOA estimator assuming single-speaker activity

Only in this section is single-speaker activity assumed. The DOA of the speaker is directly estimated using the ML criterion. Assuming that only a single speaker is active, the p.d.f. of the microphone observations is modelled by $\mathbf{y} \sim \mathcal{N}_C\left(\mathbf{y}, \phi_{X_S}\mathbf{g}(\theta_S)\mathbf{g}^\mathrm{H}(\theta_S) + \Phi_\mathbf{y}\right)$ where ϕ_{X_S} is the PSD of the speaker. The DOA can be estimated by

$$\widehat{\theta}_S = \underset{\theta_S}{\mathrm{argmax}} \log \prod_k \mathcal{N}_C\left(\mathbf{y}, \phi_{X_S}\mathbf{g}(\theta_S)\mathbf{g}^\mathrm{H}(\theta_S) + \Phi_\mathbf{y}\right) \tag{8}$$

Because the DOA and the speaker PSD are both unknown, the PSD should be estimated first for each possible DOA candidate. Using the Fisher-Neyman factorization, the above p.d.f. can be factorized to:

$$\begin{aligned}\mathcal{N}_C\left(\mathbf{y}, \phi_X \mathbf{g}\mathbf{g}^\mathrm{H} + \Phi_\mathbf{y}\right) &= \mathcal{N}_C\left(X_\mathrm{MVDR}, \phi_X + \phi_{\tilde{\mathbf{v}}}\right) \\ &\pi^{N-1} \frac{|\phi_{\tilde{\mathbf{v}}}|}{|\Phi_\mathbf{y}|} \exp\left(-\mathbf{y}^H \Phi_\mathbf{y}^{-1}\mathbf{y} + \frac{|X_\mathrm{MVDR}|^2}{\phi_{\tilde{\mathbf{v}}}}\right)\end{aligned} \tag{9}$$

where $X_{\mathrm{MVDR}} \equiv \frac{\mathbf{g}_j^H \Phi_{\mathbf{v}}^{-1} \mathbf{y}}{\mathbf{g}^H \Phi_{\mathbf{v}}^{-1} \mathbf{g}}$ is the output of the MVDR beamformer steered to the DOA represented by $\mathbf{g}$ and $\phi_{\tilde{\mathbf{v}}} \equiv \frac{1}{\mathbf{g}^H \Phi_{\mathbf{v}}^{-1} \mathbf{g}}$ is the noise power at the output of the MVDR beamformer. Using the ML estimator, the PSD of the speaker is obtained by taking the derivative of the above p.d.f. w.r.t. ϕ_X and equaling it to zero, $\widehat{\phi}_{X_j} = |X_{\mathrm{MVDR}}|^2 - \phi_{\tilde{\mathbf{v}}}$. Finally, inserting the estimate of the speech power into the ML in (8) yields the following simplified expression containing the posterior SNR in each DOA:

$$\widehat{\theta}_S = \underset{\theta}{\operatorname{argmax}} \sum_k \frac{|X_{\mathrm{MVDR}}(\theta)|^2}{\phi_{\tilde{\mathbf{v}}}(\theta)} - \log \frac{|X_{\mathrm{MVDR}}(\theta)|^2}{\phi_{\tilde{\mathbf{v}}}(\theta)} \quad (10)$$

To get smoothed DOA estimates over time, it is recommended to smooth the posterior SNR over time according to:

$$\mathrm{pSNR}(m) = \alpha \cdot \mathrm{pSNR}(m-1) + (1-\alpha) \frac{|X_{\mathrm{MVDR}}(\theta)|^2}{\phi_{\tilde{\mathbf{v}}}(\theta)} \quad (11)$$

where $0 \leq \alpha < 1$ is a smoothing factor and use $\mathrm{pSNR}(m)$ in (10) instead of $\frac{|X_{\mathrm{MVDR}}(\theta)|^2}{\phi_{\tilde{\mathbf{v}}}(\theta)}$. Note that the dominant DOA is actually determined by the DOA with the maximum posterior SNR[1]. As is shown in the experiments, this ML criterion sometimes implies a DOA that maximizes the posterior SNR and not the oracle DOA of the dominant speaker. For camera steering usage or for ASR usage (which requires undistorted speech during the noise reduction operation such as beamforming), this characteristic might be problematic. Additionally, each time the noise PSD matrix is updated it should be inverted (which increases the computational burden).

B. ML-based DOA estimator assuming multi-speaker activity

Assuming the p.d.f. in (7), the ML-based estimation of the speakers PSD matrix is obtained by:

$$\widehat{\Phi}_X = \underset{\Phi_X}{\operatorname{argmax}} \mathcal{N}_C \left(\mathbf{G} \Phi_X \mathbf{G}^H + \Phi_{\mathbf{v}} \right). \quad (12)$$

Using the Fisher-Neyman factorization the above p.d.f. can be factorized to:

$$\mathcal{N}_C \left(\mathbf{y}, \mathbf{G} \Phi_X \mathbf{G}^H + \Phi_{\mathbf{v}} \right) = \mathcal{N}_C \left(\mathbf{x}_{\mathrm{LCMV}}, \Phi_X + \Phi_{\tilde{\mathbf{v}}} \right)$$
$$\pi^{N-J} \frac{|\Phi_{\tilde{\mathbf{v}}}|}{|\Phi_{\mathbf{v}}|} \exp \left(-\mathbf{y}^H \Phi_{\mathbf{v}}^{-1} \mathbf{y} + \mathbf{x}_{\mathrm{LCMV}}^H \Phi_{\tilde{\mathbf{v}}}^{-1} \mathbf{x}_{\mathrm{LCMV}} \right) \quad (13)$$

where $\mathbf{x}_{\mathrm{LCMV}} \equiv \left(\mathbf{G}^H \Phi_{\mathbf{v}}^{-1} \mathbf{G} \right)^{-1} \mathbf{G}^H \Phi_{\mathbf{v}}^{-1} \mathbf{y}$ is the multi-speaker LCMV outputs and $\Phi_{\tilde{\mathbf{v}}} \equiv \left(\mathbf{G}^H \Phi_{\mathbf{v}}^{-1} \mathbf{G} \right)^{-1}$ is the noise PSD matrix of the residual noises obtained at the LCMV outputs. Taking the derivative of the above p.d.f. w.r.t. Φ_X and equaling it to zero yields $\Phi_X = \mathbf{x}_{\mathrm{LCMV}} \mathbf{x}_{\mathrm{LCMV}}^H - \Phi_{\tilde{\mathbf{v}}}$. To get smoothed PSD estimates, the estimated PSDs can be smoothed across the time by:

$$\widehat{\Phi}_X(m) = \alpha \widehat{\Phi}_X(m-1) + (1-\alpha) \left(\mathbf{x}_{\mathrm{LCMV}} \mathbf{x}_{\mathrm{LCMV}}^H - \Phi_{\tilde{\mathbf{v}}} \right). \quad (14)$$

Finally, the dominant DOA is associated with the DOA with the maximum power over all frequencies, namely $\widehat{\theta}_S = \theta_{\widehat{j}}$ where $\widehat{j} = \operatorname{argmax}_j \sum_k \widehat{\Phi}_{X,jj}$ and $\widehat{\Phi}_{X,jj}$ is the j-th diagonal element of $\widehat{\Phi}_X$.

Note that the matrix $\mathbf{G}^H \Phi_{\mathbf{v}}^{-1} \mathbf{G}$ is a $J \times J$ matrix and can be inverted only when $J \leq N$ because its maximal rank equals N. This restriction can be problematic for a desired low-resolution of DOA search (namely large number of searched DOAs J). Thus, this algorithm is described here only for the sake of completeness and not experimented with in this paper. In the next section, the proposed Frobenius-norm-based PSD estimation is derived, which has no limit on the number of speakers.

C. Multi-speaker PSD estimation using Frobenius norm minimization

In this section, the speech PSDs are estimated by matching the short-term estimate of the received signal PSD matrix with its matrix-model. The matrix-model of the PSD matrix of $\mathbf{y}$ is given by $\Phi_{\mathbf{y}} = \sum_j \phi_{X_j} \Sigma_j + \Phi_{\mathbf{v}}$ where $\Sigma_j \equiv \mathbf{g}(\theta_j) \mathbf{g}^H(\theta_j)$ and a short-term estimate of $\Phi_{\mathbf{y}}$ can be recursively given by:

$$\widehat{\Phi}_{\mathbf{y}}(m) = \alpha \widehat{\Phi}_{\mathbf{y}}(m-1) + (1-\alpha)\, \mathbf{y}(m) \mathbf{y}^{\mathrm{H}}(m). \quad (15)$$

Matching the modelled PSD matrix and the estimated PSD matrix of $\mathbf{y}$, the problem at hand can be recast as a system of N^2 equations in J variables. Because there might be more or fewer equations than variables, the best fitting parameter set that minimizes the total squared error can be found by minimizing the Frobenius norm between $\widehat{\Phi}_{\mathbf{y}}(m)$ in (15) and its matrix-model. Accordingly, estimates of the speech PSD can therefore be the minimizers of the following cost-function:

$$\widehat{\boldsymbol{\phi}}_X = \underset{\boldsymbol{\phi}_X}{\operatorname{argmin}} \left\| \widehat{\Phi}_{\mathbf{y}}(m) - \left(\sum_j \phi_{X_j} \Sigma_j + \Phi_{\mathbf{v}} \right) \right\|_{\mathrm{F}}^2, \quad (16)$$

with $||\cdot||_{\mathrm{F}}^2$ being the squared Frobenius norm given for any arbitrary matrix $\mathbf{Z}$ by $||\mathbf{Z}||_{\mathrm{F}}^2 = \sum_{i,j} |\mathbf{Z}_{i,j}|^2 = \mathrm{Tr}\left[\mathbf{Z}^{\mathrm{H}} \mathbf{Z}\right]$. Denote $\boldsymbol{\phi}_X = \left[\begin{array}{ccc} \phi_{X_1} & \ldots & \phi_{X_J} \end{array} \right]$. Following some algebraic steps, the cost function in (16) can be written as:

$$||\Phi_{\mathrm{e}}(m)||_{\mathrm{F}}^2 = \boldsymbol{\phi}_X^{\mathrm{T}} \mathbf{A} \boldsymbol{\phi}_X - 2\mathbf{b}^{\mathrm{T}}(m) \boldsymbol{\phi}_X + C(m), \quad (17)$$

where $\mathbf{A}$ is time-invariant $J \times J$ matrix defined by:

$$\mathbf{A} \equiv \left(\begin{array}{ccc} \mathrm{Tr}\left[\Sigma_1^H \Sigma_1\right] & \ldots & \mathrm{Tr}\left[\Sigma_1^H \Sigma_J\right] \\ \vdots & \ddots & \vdots \\ \mathrm{Tr}\left[\Sigma_J^H \Sigma_1\right] & \ldots & \mathrm{Tr}\left[\Sigma_J^H \Sigma_J\right] \end{array} \right), \quad (18)$$

$\mathbf{b}(m)$ is time-varying vector defined as:

$$\mathbf{b}(m) \equiv \left(\begin{array}{c} \mathrm{Tr}\left[\Sigma_1^H \left(\widehat{\Phi}_{\mathbf{y}}(m) - \Phi_{\mathbf{v}} \right)\right] \\ \vdots \\ \mathrm{Tr}\left[\Sigma_J^H \left(\widehat{\Phi}_{\mathbf{y}}(m) - \Phi_{\mathbf{v}} \right)\right] \end{array} \right) \quad (19)$$

and $C(m)$ is defined as:

$$C(m) \equiv \mathrm{Tr}\left[\left(\widehat{\Phi}_{\mathbf{y}}(m) - \Phi_{\mathbf{v}} \right)^{\mathrm{H}} \left(\widehat{\Phi}_{\mathbf{y}}(m) - \Phi_{\mathbf{v}} \right) \right]. \quad (20)$$

[1]Note that the function $x - \log x$ is monotonic increasing function with x when $x > 1$. Thus the logarithm expression can be neglected in (10)

Because the cost function $||\Phi_{\mathrm{e}}(m)||_{\mathrm{F}}^2$ has a quadratic form, setting its gradient w.r.t. $\boldsymbol{\phi}_X$ to zero yields the following minimum-point:

$$\widehat{\boldsymbol{\phi}}_X = \mathbf{A}^{-1}\mathbf{b}(m). \tag{21}$$

Note that this estimator has no restriction on the number of searched DOAs J. Additionally, there are no matrix inversions within the online implementation because $\mathbf{A}^{-1}$ is independent of the noise PSD and can be calculated in advance. Finally, the dominant DOA is associated with the direction with the maximum PSD, namely $\widehat{\theta}_S = \theta_{\widehat{j}}$ where $\widehat{j} = \mathrm{argmax}_j \sum_k \widehat{\boldsymbol{\phi}}_{X,j}$ and $\widehat{\boldsymbol{\phi}}_{X,j}$ is the j-th element of $\widehat{\boldsymbol{\phi}}_X$.

IV. Performance Evaluation

The performance of the proposed algorithm is evaluated on simulated signals by estimating the DOAs of a single speaker and calculating the MAE w.r.t. the oracle DOA. Four algorithms are compared by their MAE results:

1) The GCC-Hanan-Thompson (HT) [1] algorithm. Because the GCC-HT is designed only for dual-microphone cases, and this paper assumes any microphone array configuration, the GCCs were summed for each possible pair of microphones. Given only the q_1-th and q_2-th microphones and using our notations, the GCC-HT for dual microphones is given by $\widehat{\theta}_S = \theta_{\widehat{j}}$ where:

$$\widehat{j} = \underset{j}{\mathrm{argmax}} \sum_k \frac{\widehat{\Phi}_{\mathbf{y},q_1q_2}}{\left|\widehat{\Phi}_{\mathbf{y},q_1q_2}\right|} \frac{\psi_{q_1q_2}}{1-\psi_{q_1q_2}} \frac{G_{j,q_1}}{G_{j,q_2}}, \tag{22}$$

where $\psi_{q_1q_2} = \frac{\left|\widehat{\Phi}_{\mathbf{y},q_1q_2}\right|^2}{\widehat{\Phi}_{\mathbf{y},q_1q_1}\widehat{\Phi}_{\mathbf{y},q_2q_2}}$ is the coherence between the microphone signals and $\widehat{\Phi}_{\mathbf{y},q_1q_2}$ is the q_1, q_2 element of $\widehat{\Phi}_{\mathbf{y}}$. Note that the GCC-HT assumes a spatial white noise field, which is its main disadvantage w.r.t. the other algorithms that assume a known PSD matrix of the ambient noise.
2) A DOA estimator obtained by maximization of the energy of the D&S beamformer output steered to each possible direction. Using our notations, the D&S-based estimator is given by $\widehat{\theta}_S = \theta_{\widehat{j}}$ where $\widehat{j} = \mathrm{argmax}_j \sum_k \mathbf{g}_j^{\mathrm{H}} \widehat{\Phi}_{\mathbf{y}} \mathbf{g}_j$.
3) The single-speaker-assumption-based DOA estimator given in Sec. III-A.
4) The proposed Frobenius-norm-minimization-based DOA estimator given in Sec. III-C.

A. Experimental setup

A circular array with a 5 cm diameter consisting of three microphones at the perimeter and one at the center was used. Anechoic speech and noise signals were convolved by room impulse responses (RIRs) produced by an open-source RIRs simulator [2]. The reverberation time was adjusted to $T_{60} = 0.3$. The modelled observed speech and the directional noise were summed with wanted SNRs before inputting to the algorithms. To evaluate the DOA estimates with various angular distances between the speaker and the noise source, the DOA of the noise was set to various angles while the DOA of the speaker was fixed to 60^o. The length of each simulated recording was 60sec. The sampling frequency was 16 kHz, and the frame length of the short-time Fourier transform (STFT) was 32 ms with an 8 ms overlap. The resolution of the candidate DOAs was $\beta = 5^o$ ($J = 72$ DOA candidates). The frequency band $300 - 3000$ Hz was used for the DOA estimation.

[2]The RIRs simulator can be freely downloaded from https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.

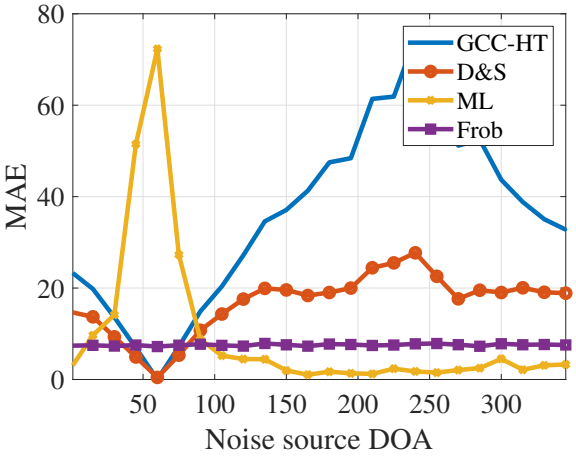

Fig. 1. MAE results for speaker at 60^o and various noise source DOA

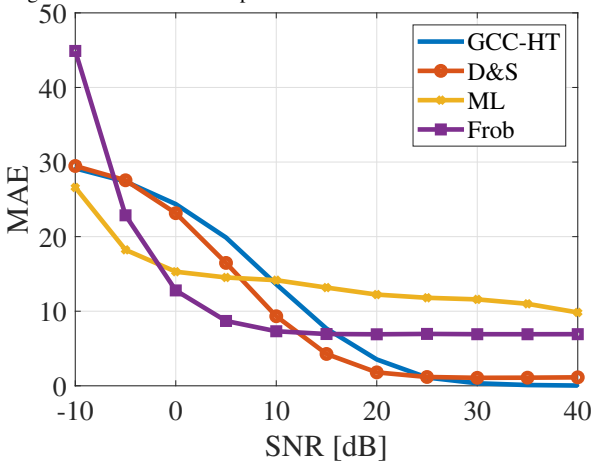

Fig. 2. MAE results for SNR=10dB, speaker at 60^o, noise source at 30^o and various SNR

B. Results and discussion

The localization algorithms estimate the dominant DOA for each frame. Given $\widehat{\theta}_S$ produced by each algorithm and the oracle DOA of the speaker $\theta_S = 60^o$, the MAE is calculated by:

$$\mathrm{MAE} = \frac{1}{M}\sum_m \min\left(\left|\widehat{\theta}_S - \theta_S\right|, 360^o - \left|\widehat{\theta}_S - \theta_S\right|\right). \tag{23}$$

The MAEs for SNR=10dB, fixed speaker DOA 60^o and various noise source DOAs are presented in Fig. 1. It can be

verified that the MAE for the GCC-HT is generally high, apart from where the speaker and the noise are from the same DOA. The GCC-HT does not consider the noise PSD, and therefore implies the most significant DOA (the DOA of the speaker or the noise source or some average DOA of them). Note that the GGC-HT enhances any TF bins with high coherence ψ and thus the noisy TF bins are also enhanced. The D&S has lower MAE (relative to GCC-HT) because the energy of the speech is higher than the noise energy (SNR=10dB). The ML-based DOA estimator has the lower MAE where the speaker and the noise source are angularly far away, but has a high MAE when the speaker is close to the noise source because the ML estimator actually biases the DOA estimate away from the noise source DOA to where the best posterior SNR at the output of the MVDR beamformer is achieved. The proposed Frobenius-norm-based DOA has a constant MAE of 7^o in all cases.

The MAEs for fixed-speaker DOA at 60^o, noise-source DOA at 30^o, and various SNR are presented in Fig. 2. It can be verified that for 20dB $\leq$ SNR, the GCC-HT and D&S outperform the ML- and Frobenius-norm-based DOAs because the speech is significant enough relative to the noise and thus neglecting the noise is preferable. In 0dB $\leq$ SNR $\leq$ 10dB, the proposed Frobenius norm outperforms the other algorithms, and in -10dB $\leq$ SNR $\leq$ 0dB, all of the algorithms fail.

Example signal of fixed-speaker DOA at 60^o, noise-source DOA at 30^o, SNR=10 db and the DOA estimates of the various algorithms are presented in Fig. 3. It can be verified that both

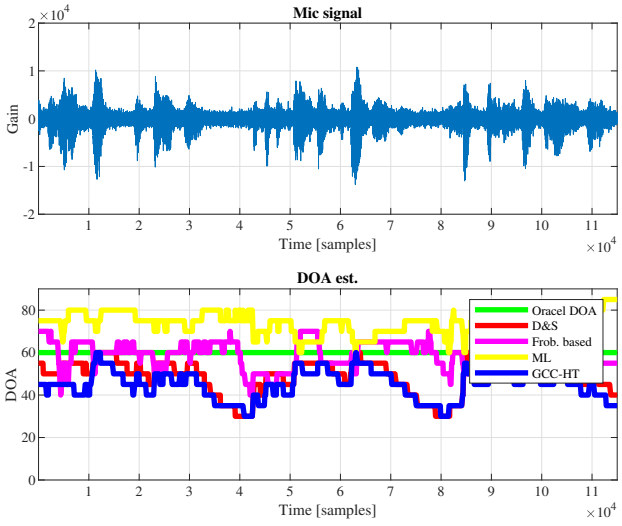

Fig. 3. (Up) signals with speaker at 60^o, noise source at 30^o and SNR=10dB. (Down) DOA estimates of the various algorithms

the DOA estimates of D&S and the GCC-HT are biased to the noise-source DOA. The DOA estimates of ML are biased to other side of the speaker (DOA$= 70^o - 80^o$). The DOA estimates of the proposed Frobenius-norm-based algorithm are more focused on the oracle speaker DOA.

V. Conclusions

In this work, a single-speaker DOA localization algorithm using a microphone array is presented. Multiple possible speakers were located in each one of a predefined set of candidate DOAs. Instead of directly estimating the DOA, the PSDs of the speakers were mutually estimated and the dominant DOA was determined by the speaker with the maximal PSD. The PSD estimation task was employed by minimizing the Frobenius norm of the matrix-difference between the estimated PSD matrix of the received signals and the model-matrix described by the multiple speaker presence. The experimental study demonstrated the benefits of the proposed Frobenius-based DOA algorithm in a simulated data set w.r.t. an ML-based DOA estimator, especially where the speaker is close to the noise source.

References

[1] Charles Knapp and Glifford Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[2] Ralph Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[3] Hao Ye and D DeGroat, "Maximum likelihood DOA estimation and asymptotic cramér-rao bounds for additive unknown colored noise," *IEEE Transactions on Signal Processing*, vol. 43, no. 4, pp. 938–949, 1995.

[4] Ozgur Yilmaz and Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[5] Michael I Mandel, Ron J Weiss, and Daniel PW Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[6] Ofer Schwartz, Yuval Dorfan, Emanuël AP Habets, and Sharon Gannot, "Multi-speaker DOA estimation in reverberation conditions using expectation-maximization," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.

[7] Yuval Dorfan, Ofer Schwartz, Boaz Schwartz, Emanuël AP Habets, and Sharon Gannot, "Multiple DOA estimation and blind source separation using estimation-maximization," in *IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, 2016.

[8] Ofer Schwartz, Yuval Dorfan, Maja Taseska, Emanuël AP Habets, and Sharon Gannot, "DOA estimation in noisy environment with unknown noise power using the EM algorithm," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 86–90.

[9] Koby Weisberg, Sharon Gannot, and Ofer Schwartz, "An online multiple-speaker DOA tracking using the cappé-moulines recursive expectation-maximization algorithm," in *ICASSP 2019*. IEEE, pp. 656–660.

[10] Hai Quang Dam, Siow Yong Low, Hai Huyen Dam, and Sven Nordholm, "Space constrained beamforming with source psd updates," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004, vol. 4, pp. iv–iv.

[11] Ofer Schwartz, Sharon Gannot, and Emanuël AP Habets, "Joint estimation of late reverberant and speech power spectral densities in noisy environments using frobenius norm," in *EUSIPCO 2016*. IEEE, pp. 1123–1127.

[12] Ina Kodrasi and Simon Doclo, "Joint late reverberation and noise power spectral density estimation in a spatially homogeneous noise field," in *ICASSP 2018*. IEEE, pp. 441–445.

[13] Daniele Mirabilii and Emanuël AP Habets, "Multi-channel wind noise reduction using the corcos model," in *ICASSP 2019*. IEEE, pp. 646–650.

[14] Thomas Dietzen, Simon Doclo, Marc Moonen, and Toon van Waterschoot, "Square root-based multi-source early PSD estimation and recursive retf update in reverberant environments by means of the orthogonal procrustes problem," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 755–769, 2020.