MULTI-SPEAKER DOA TRACKING ALGORITHM UTILIZING PROBABILITY HYPOTHESIS DENSITY FILTER AND WEIGHTED HISTOGRAM OF SRP-PHAT

Yosef Soussana, Elior Hadad, and Sharon Gannot

Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel
{Yosef.sussana, Elior.Hadad, Sharon.Gannot}@biu.ac.il

ABSTRACT

This contribution presents a concurrent speakers' direction of arrival (DOA) tracking algorithm in reverberant environments. The algorithm is formulated in two stages, leveraging speech sparsity in the short-time Fourier transform (STFT) domain. In the first stage, sets of DOAs per batch of time frames are computed. Initially, a single narrow-band (NB) DOA per time-frequency (TF) bin is selected using the W-disjoint orthogonality property of speech. The NB DOA is obtained as the maximum of the steered response power phase transform (SRP-PHAT) localization spectrum at that TF bin, together with a quality measure describing the confidence in the estimation. A localization spectrum is obtained by combining the NB DOAs using a weighted histogram, with the quality measures serving as weights. The set of DOAs is determined by identifying peaks in the resulting localization spectrum. The collection of DOAs is modeled as a random finite set (RFS). In the second stage, the probability hypothesis density (PHD) filter is applied to estimate and track the speakers' DOAs over a collection of batches. Information from the first stage is utilized to calculate prior knowledge on the appearance of new speakers. Our experimental study demonstrates the superiority of the proposed algorithm over a baseline approach.

1. INTRODUCTION

Speaker localization and tracking play a crucial role in various audio applications, including hearing aids, surveillance systems, and robot audition. Commonly used sound source localizer (SSL) techniques include the generalized cross correlation (GCC)-based methods [1], steered response power (SRP)-based methods (such as SRP-PHAT [2], which is obtained from SRP by applying a phase transform (PHAT) whitening), and maximum likelihood estimation (MLE)-based methods [3, 4]. Nevertheless, the effectiveness of these methods significantly degrades when speakers are concurrently active and in reverberant environments, both of which are typical of real-world situations.

In scenarios involving concurrent speakers, speech sources can be assumed to exhibit disjoint activity in the short-time Fourier transform (STFT) domain (the so-called W-disjoint orthogonality property [5]). According to this property, each time-frequency (TF) bin can be associated with a dominant speaker originating from a specific direction of arrival (DOA), which we denote narrow-band (NB) DOA estimate. Several two-stage SRP-based techniques were employed to estimate the NB DOAs [6–10]. In [7], it was proposed first to estimate single NB DOA by identifying the DOA associated with the highest value in NB SRP-PHAT localization spectrum. Then, a localization spectrum was computed by creating a histogram of the NB DOAs, with confidence measures serving as weights. The DOAs were determined by identifying peaks in the localization spectrum. A similar approach was proposed in [11], where the NB DOA was estimated using the MLE procedure. The accuracy of the DOA localization improves when the number of STFT time-frames increases. In dynamic scenarios, this problem becomes even more prominent.

In a dynamic scenario, tracking becomes essential. A multichannel extension of [12] was proposed in [13]. The tracking abilities were obtained by employing two alternative recursive EM (REM) procedures [14, 15].

The probability hypothesis density (PHD) filter is a recursive Bayesian filter that propagates the first-order moment of the p.d.f., also known as intensity, of the random finite set (RFS) of state in time [16]. This approximation was suggested in multi-speaker tracking scenarios to alleviate the computational intractability. The PHD filter is applied to the single-speaker state space and circumvents the combinatorial problem that arises from data association. In [17], the authors presented an analytical solution to the PHD recursion for the case of linear Gaussian dynamics in a multiple-target scenario. It was shown that when the initial prior intensity is a mixture of Gaussians (MoG), the posterior intensity at any subsequent time step is also a MoG.

In the current contribution, we propose a two-stage algorithm designed for SSL and tracking problems, tailored to dynamic and multi-speaker scenarios. In the first stage a *batch* DOAs are computed as described hereafter. Initially, single NB DOA estimates are computed by selecting the DOA corresponding to the maximum value in the NB SRP-PHAT spectrum. Following the approach described in [7], a quality measure is calculated for each NB DOA, providing confidence for the estimated decision. Next, a localization spectrum for a batch of STFT time-frames is computed. This *batch* localization spectrum is obtained by combining the weighted NB DOAs using the weighted histogram, where the quality measures serve as weights. Finally, the batch DOAs are determined by identifying peaks in the resulting batch localization spectrum. In the second stage, RFSs of candidate DOAs, obtained from the batch localization spectrum by the application of a threshold, are fed into the PHD filter with *batch* confidence, which serves as a prior information in order to manage and track the unknown number of speakers and overcome spectrum detection errors. These errors may occur in the first stage due to small batch size or spontaneous errors caused by reverberation. Thanks to the PHD filter, we increase the tracking ability while working with noisier input data.

2. PROBLEM FORMULATION

Consider an M-microphone array receiving signals from Q concurrent speakers in a noisy and reverberant environment. Let $Z_m(\ell, k)$ be the signals received by the mth microphone (located at (x_m, y_m)). The signals in the STFT domain are given by:

$$Z_m(\ell, k) = \sum_{q=1}^{Q} A_{m,q}(k)\ S_q(\ell, k) + V_m(\ell, k), \qquad (1)$$

where $\ell = 0, \ldots, L-1$ denotes the frame-index and $k = 0, \ldots, K-1$ the frequency index. $A_{m,q}(k)$ denotes the relative transfer function (RTF) associated with source q, which relates the m-th microphone and the reference microphone, $S_q(\ell, k)$ denotes the speech signal uttered by source q as received by the reference microphone, and $V_m(\ell, k)$ denotes the ambient noise.

In this work, we only localize in the azimuthal direction, $\theta \in \Theta = \{0^\circ, 180^\circ\}$, where θ is measured with respect to the array axis. The time difference of arrival (TDOA) of the source signal at the mth microphone with respect to the reference microphone can be expressed as $\tau_m(\theta) = \frac{d_m^x \cos\theta + d_m^y \sin\theta}{c}$, where c is the sound velocity and $d_m^x = x_m - x_r$ and $d_m^y = y_m - y_r$ are the distances between the mth microphone and the reference microphone r.

The relative direct transfer function (RDTF) with respect to θ is defined as

$$\mathbf{d}_\theta(k) = \begin{bmatrix} D_1(k) & \ldots & D_M(k) \end{bmatrix}^\top, \tag{2}$$

where $D_m(k) = e^{-\iota \frac{2\pi k}{K} \frac{\tau_m(\theta)}{T_s}}$ and T_s denotes the sampling period.

We define a grid of J DOA candidates $\theta \in \Theta$ for the localization spectrum. Each direction is associated with the respective RDTF $\mathbf{d}_\theta(k)$. The NB SRP-PHAT spectrum is given by [2, 7]:[1]

$$\mathcal{J}(\theta; \ell, k) = \sum_{m=1}^{M} \sum_{n=1}^{M} \frac{Z_m(\ell, k) Z_n^*(\ell, k)}{|Z_m(\ell, k)||Z_n(\ell, k)|} e^{\iota \frac{2\pi k}{K} \frac{\tau_m - \tau_n}{T_s}}. \tag{3}$$

3. BATCH DIRECTION OF ARRIVAL LOCALIZATION

In the next sections, the proposed tracking algorithm is presented. This section introduces the batch DOA procedure, which provides a set of DOAs for each batch of time-frames. Then, in Section 4, the derivation of the PHD tracker is provided.

The batch DOA estimation task is formulated through the following steps, as described in [7]. Initially, single NB DOA estimates are computed by selecting the DOA corresponding to the maximum value in the NB SRP-PHAT spectrum in (3), i.e.,

$$\hat{\theta}(\ell, k) = \arg\max_\theta \mathcal{J}(\theta; \ell, k). \tag{4}$$

A quality measure is calculated for each NB DOA, providing confidence for the estimated decision. We used a directivity-based confidence measure, which is defined as the peak value of the NB localization spectrum (associated with the selected NB DOA), normalized by a summation of all grid points of the localization spectrum:

$$w_D(\hat{\theta}(\ell, k)) = \frac{\mathcal{J}(\hat{\theta}(\ell, k), \ell, k)}{\sum_{j'=1}^{J} \mathcal{J}(\theta_{j'}; \ell, k)}. \tag{5}$$

Clearly, as $w_D(\hat{\theta}(\ell, k))$ increases, the estimated NB DOA becomes more dominant with respect to the other examined directions. This value is influenced by both the geometry and the reverberation level. Next, a localization spectrum for a batch of STFT time-frames is computed. The indicator $u(j, \ell, k)$ associates each TF bin with a single source emitting from angle θ_j; i.e.,

$$u(j, \ell, k) = \begin{cases} 1 & \hat{\theta}(\ell, k) = \theta_j \\ 0 & \text{otherwise} \end{cases}; \tag{6}$$

such that the indicator $u(j, \ell, k)$ is equal to 1 only when the NB DOA estimate $\hat{\theta}(\ell, k)$ is equal to θ_j in the respective TF bin. The *batch* localization spectrum is defined as a weighted histogram of the NB DOAs over a sliding window of length B time-frames, and with the confidence measures $w_D(\hat{\theta}(\ell, k))$ as weights:

$$\Psi_b(\theta_j) = \sum_{\ell=\ell_b}^{\ell_b+B} \sum_{k=0}^{K-1} w_D(\hat{\theta}(\ell, k)) u(j, \ell, k). \tag{7}$$

Finally, a set of *batch* DOAs are determined by identifying the peaks of the corresponding batch localization spectrum $\Psi_b(\theta_j)$, $j = 1, \ldots, J$. Furthermore, a *batch* confidence measure is calculated for each batch DOA. This confidence measure is used in the tracking phase alongside the candidate DOAs, Ω_b and serving as prior information:

$$v_b^i = \frac{\Psi_b(\theta_i)}{\sum_{\theta \in \Omega_b} \Psi_b(\theta)}. \tag{8}$$

Now, this measure reflects the dominance of the ith DOA relative to other estimated DOAs within the same batch. This resulting value serves as prior information necessary for the GM-PHD filter, as explained in Section 4.3.1.

4. GM-PHD DOA TRACKING

To fully characterize the statistical properties of an active speaker DOA, its posterior p.d.f. should be estimated and propagated in time. SSL may include an unknown number of speakers in the enclosure, and hence, tracking of multiple sources is required. For multi-source tracking, however, the p.d.f. is numerically intractable. Rather than propagating the p.d.f., the posterior can be approximated by its first-order moment, which is known as the intensity function. In [17], the Gaussian mixture probability hypothesis density (GM-PHD) filter and the associated recursive equations were developed under the Gaussian linear motion assumption for multi-target tracking in surveillance systems applications. In this study, we utilize the the GM-PHD filter framework proposed in [17], to estimate the DOA for unknown number of speakers. In our study, we propose calculating new prior information for each recursion step, which is essential for the GM-PHD filter. This new prior information is based on the DOA estimator, introduced in Section 3. In the following subsections, the speakers' state-space model (Section 4.1), measurement model (Section 4.2), and the filtering recursion (Section 4.3) are described.

4.1. State space model

This subsection describes the RFS model with batch evolution for the multi-speaker DOAs state. The model integrates the dynamic changes in the active speaker's DOA along with the addition or removal of a source. Speaker state $\mathbf{x}_{b-1}$ at batch index $b-1$ comprises the speaker DOA and the angular velocity denoted by $[\theta, \rho]^\top$, respectively.

Conditioned on the existence of an arbitrary state speaker for batch b, the p.d.f. of a transition from state $\mathbf{x}_{b-1}$ to state $\mathbf{x}_b$ is given by $f(\mathbf{x}_b|\mathbf{x}_{b-1})$. The state transition is governed by a linear dynamic model [17](see Eq.(17)) and can expressed by the following distribution:

$$f(\mathbf{x}_b|\mathbf{x}_{b-1}) = \mathcal{N}(\mathbf{x}_b; \mathbf{F}\mathbf{x}_{b-1}, \mathbf{Q}) \tag{9}$$

with $\mathbf{F} = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix}$ and $\mathbf{Q} = \sigma_\phi^2 \begin{bmatrix} \frac{\Delta^4}{4} & \frac{\Delta^3}{2} \\ \frac{\Delta^3}{2} & \Delta^2 \end{bmatrix}$, are the dynamic motion model and the process noise, respectively, and Δ is the time

[1] The conventional wide-band SRP-PHAT localization spectrum is computed by averaging the NB SRP-PHAT localization spectrum across all examined TF bins. This process allows us to estimate the wide-band DOAs by identifying the peaks in the wide-band SRP-PHAT localization spectrum.

difference between two adjacent batches. Each $\mathbf{x}_{b-1}$ can be either active at the following batch, b, with probability p_s, or inactive with probability $1-p_s$, and $\mathcal{X}_{b-1}$ is the RFS of all speaker states in the enclosure at batch index $b-1$. Consequently, for a given state $\mathbf{x}_{b-1} \in \mathcal{X}_{b-1}$ at batch $b-1$, its behavior at the next batch is also modeled as an RFS denoted $S_{b|b-1}(\mathbf{x}_{b-1})$. For given multi-speaker states $\mathcal{X}_{b-1}$, the following batch states space $\mathcal{X}_b$ is formulated as

$$\mathcal{X}_b = \left[\bigcup_{\boldsymbol{\zeta} \in \mathcal{X}_{b-1}} S_{b|b-1}(\zeta) \right] \cup \Gamma_b, \tag{10}$$

obtained by the union of the active speakers that remain active from the previous batch and the RFS of the new speaker's appearance, Γ_b, in the current batch. The intesity model of Γ_b will be described in Section 4.3.1.

4.2. Measurements model

Since the number of speakers is unknown, similar to the state space model, the measurement state at each batch index is described as an RFS as follows.

Let the measurements set Ω_b at batch index b, consist of J_b DOAs, $\Omega_b \triangleq \left\{\theta_b^{(j)}\right\}_{j=1}^{J_b}$. Each measurement $\theta_b^{(j)}$, resulting in from either speaker or clutter caused by reverberation, is modeled as a linear Gaussian measurement:

$$g(\theta_b^{(j)}|\mathbf{x}) \sim \mathcal{N}(\theta_b^{(j)}; \mathbf{H}\mathbf{x}, \sigma_r^2), \tag{11}$$

where $\mathbf{H} = [1, 0]$ is the observation matrix. The measurement noise variance σ_r^2, can be deduced from the DOA estimator presented in Section 3. The multi-source measurement process, taking into account both the speakers' DOA detection, missed detection, and the clutter due to reverberation, is also modeled as RFS

$$\Omega_b = \left[\bigcup_{\mathbf{x} \in \mathcal{X}_b} D(\mathbf{x}) \right] \cup C_b, \tag{12}$$

where C_b are noisy mesurments, with uniform intensity over DOA region, together with the detection RFS govenerd by MoG intensity $D(\mathbf{x})$. When active speaker with state $\mathbf{x}_b^{(j)}$ is detected then $D(\mathbf{x}_b^{(j)}) = \theta_b^{(j)}$ otherwise $D(\mathbf{x}_b^{(j)}) = \emptyset$.

4.3. GM-PHD recursion for DOA tracking

This subsection describes the recursion process of the GM-PHD filter. This PHD recursion is applied for propagating the posterior intensity, $\lambda_b(\mathbf{x})$, which is modeled as MoG.

4.3.1. Prediction step

The predicted DOA's intensity, $\lambda_{b|b-1}$ of the state $\mathbf{x}$ at batch index b, is given by [17]

$$\lambda_{b|b-1}(\mathbf{x}) = \lambda_{S,b|b-1}(\mathbf{x}) + \gamma_b(\mathbf{x}), \tag{13}$$

where $\gamma_b(\mathbf{x})$ is the prior information for the appearance of a new speaker in the enclosure, initialized at batch index b, and $\lambda_{S,b|b-1}(\mathbf{x})$ is the predicted intensity of a continuing speaker from batch index $b-1$. Each DOA measurement can arise from an existing speaker, a new speaker, or noise. The new speaker's prior intensity is modeled as MoG, where each DOA from the measurement set Ω_b determines the mean of a single mixture. Additionally, the mixture weight corresponds to the batch confidence value (8):

$$\gamma_b(\mathbf{x}) = \sum_{j=1}^{J_b} v_b^j \mathcal{N}(\mathbf{x}; [\theta_b^{(j)}; 0]^\top, \mathbf{P}_b), \tag{14}$$

where v_b^j is the j-th measurement confidence and $\mathbf{P}_b \triangleq \begin{bmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_a^2 \end{bmatrix}$ is the covariance matrix, with the DOA measurement error and angular velocity variance σ_a^2. The initial angular velocity is assumed to be zero. This prior is calculated at each recursion step, in contrast to [17], where fixed prior intensity for the new appearance is used. Since the measurements also play a role in the update step of the algorithm, we use a softened version of $\gamma_b(\mathbf{x})$. We apply a simple pruning procedure for the mixture $\gamma_b(\mathbf{x})$ with the previous update intensity, λ_{b-1}. Recalling [17], $\lambda_{S,b|b-1}(\mathbf{x})$ also has a MoG form:

$$\lambda_{S,b|b-1}(\mathbf{x}) = \sum_{j=1}^{J_{b-1}} w_{S,b-1}^{(j)} \mathcal{N}(\mathbf{x}; \mathbf{m}_{S,b-1}^{(j)}, \mathbf{P}_{S,b-1}^{(j)}) \tag{15}$$

where J_{b-1} is the number of MoG components at batch $b-1$ with weights $w_{S,b-1}^{(j)} = p_s w_{b-1}^{(j)}$, normalized by the probability of the speaker to continue. The means and covariances of this MoG, $\mathbf{m}_{S,b-1}^{(j)}, \mathbf{P}_{S,b-1}^{(j)}$, are calculated using the Kalman filter prediction equations [17] (see Eqs. (26)(27)).

As both the new speaker intensity $\gamma_b(\mathbf{x})$ and the surviving speaker's intensity are defined, we can express the predicted intensity of the DOA's as the sum of the two MoG:

$$\lambda_{b|b-1}(\mathbf{x}) = \sum_{j=1}^{J_{b|b-1}} w_{b|b-1}^{(j)} \mathcal{N}(\mathbf{x}; \mathbf{m}_{b|b-1}^{(j)}, \mathbf{P}_{b|b-1}^{(j)}) \tag{16}$$

where $J_{b|b-1} = J_b + J_{b-1}$ is the predicted number of speakers with the corresponding means and covariances $\mathbf{m}_{b|b-1}^{(j)}, \mathbf{P}_{b|b-1}^{(j)}$, comprising both the continuing and new speaker's DOA.

4.3.2. Update step

The updated DOA's intensity, $\lambda_b(\mathbf{x})$, also known as the posterior, is inferred from the measurements, taking into account the predicted DOA intensity $\lambda_{b|b-1}(\mathbf{x})$. As described in [17], the updated intensity is a summation of two mixtures. The first is considered as the miss-detection of any predicted intensity $\lambda_{b|b-1}(\mathbf{x})$ component, and the second is an updated version of the predicted intensity $\lambda_{b|b-1}(\mathbf{x})$ with measurements:

$$\lambda_b(\mathbf{x}) = (1 - P_D)\lambda_{b|b-1}(\mathbf{x}) + \sum_{\theta \in \Omega_b} \lambda_{D,b}(\mathbf{x}, \theta), \tag{17}$$

where

$$\lambda_{D,b}(\mathbf{x}, \theta) = \sum_{i=1}^{J_{b|b-1}} w_b^i(\theta) \mathcal{N}(\mathbf{x}; \mathbf{m}_{b|b}^{(i)}(\theta), \mathbf{P}_{b|b}^{(i)}). \tag{18}$$

The DOA measurements, Ω_b, are taken into account in (18), and each $\theta \in \Omega_b$ is associated with the predicted Gaussian's mean components of a MoG with $J_{b|b-1}$ components with following weight value:

$$w_b^i(\theta) = \frac{P_D w_{b|b-1}^{(i)} q_b^{(i)}(\theta)}{\kappa(\theta) + P_D \sum_{l=1}^{J_{b|b-1}} w_{b|b-1}^{(l)} q_b^{(l)}(\theta)}. \tag{19}$$

where $q_b^{(i)}(\theta)$ is attributed to a single measurement p.d.f., according to the predicted DOA intensity with detector uncertainty:

$$q_b^{(i)}(\theta) = \mathcal{N}(\theta; \mathbf{H}\mathbf{m}_{b|b-1}^{(i)}, \sigma_r^2 + \mathbf{H}\mathbf{P}_{b|b-1}^{(i)}\mathbf{H}^\top). \quad (20)$$

Note that in the denominator of (19), the noise intensity is taken into account, given by $\kappa(\theta) = \rho_k U(\theta)$, where ρ_k is the average number of false detection due to reverberation and $U(\theta)$ is the uniform density over DOA space.

The MoG means and covariances, $\mathbf{m}_{b|b}^{(i)}, \mathbf{P}_{b|b}^{(i)}$, respectively, are calculated with the Kalman filters equations and formulated in [17] (see Eqs. (34-36)).

Note that PHD $\sum_{\theta\in\Omega_b} \lambda_{D,b}(\mathbf{x},\theta)$ has a MoG form. By heuristic pruning of very close associated MoG terms and by neglecting MoG terms with negligible weights, we can ensure the tractability of this recursion. Otherwise, an exponential growth in the number of MoG components may occur, leading to untraceable problems.

The posterior intensity, $\lambda_b(\mathbf{x})$, is a summation of two MoG's, which can be approximated by a MoG. Hence, for each batch, the highest MoG components present the most probable concurrently active speakers, and the corresponding mean value indicates the speaker's DOA.

5. EXPERIMENTAL STUDY

5.1. Simulation setup

The proposed localization algorithm was evaluated using time-varying simulations representing dynamic scenarios. The dimensions of the acoustic enclosure are $6 \times 5 \times 3$ m and the reverberation time $T_{60} = 600$ ms. The signals were captured by a linear array comprising eight microphones with inter-distances of $[3, 3, 3, 8, 3, 3, 3]$ cm, positioned in the center of the room and sampled at 16 kHz with STFT frame length of 1024 with 50% overlap and additive sensor noise with an SNR of 20 dB.

The data in the simulation comprises 15 speakers from the TIMIT database [18]. The utterances were concatenated to a 20 sec long section, and only 2 speakers were randomly chosen for each simulation. We used the signal generator [19] to simulate maneuvering sources. The speakers were first located in an initial DOA with respect to the array center and began to move back and forth along an arc of a circle with a radius of 1.5 m from the array center. The period of the trajectory was randomly selected between 1 and 2.5 sec. We distinguish between two cases by setting different initial DOA locations. The speakers' initial location was $\pm 40^\circ$ for the first case. This scenario is designated "distant". In the second case, the speakers' initial location is set to be $\pm 15^\circ$. This scenario is designated "near". For each one of the cases, 40 Monte Carlo (MC) trials were generated, and the root mean square errors (RMSE) between the true DOAs and the estimated DOAs is presented. The baseline algorithm [7] denoted as $\hat{\theta}_{WH-10/20}$ operates in batch mode with two batch values, 10 and 20, corresponding to STFT windows representing $\sim$ 300 ms or $\sim$ 600 ms time frame, respectively. The proposed algorithm is denoted $\hat{\theta}_{\mathrm{PHD}}$.

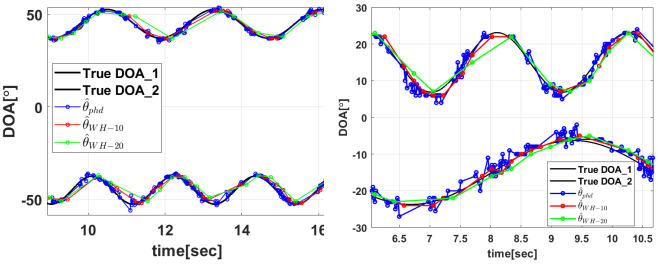
The threshold is set to achieve optimal RMSE for each MC trial for the baseline algorithms. For the proposed algorithm, the threshold was set once for all MC trials and chosen as the minimum value required for algorithm operation.

Table 1: RMSE of the DOA estimation.

Scenario	$\hat{\theta}_{\mathrm{PHD}}$	$\hat{\theta}_{WH-10}$	$\hat{\theta}_{WH-20}$
"distant"	0.89°	1.4°	2.2°
"near"	1.3°	1.3°	2.4°

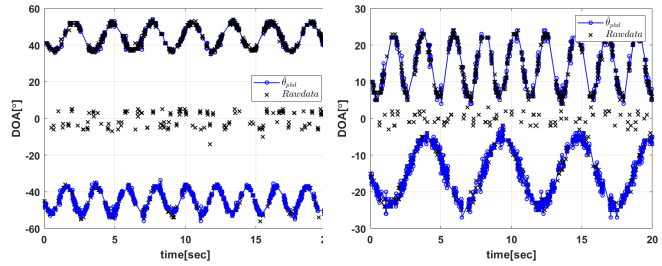
5.2. Simulation results

An illustration of one sample of the "distant" case is presented in Fig. 1(a), demonstrating a clear advantage of the proposed tracking algorithm. The PHD filter is operating in each STFT time frame, yielding higher tracking resolution as compared with the baseline algorithms, $\hat{\theta}_{WH-10/20}$, which average on 10 and 20 STFT frames. Since the DOA location may change during the group of frames, the DOA estimation may suffer from large errors. In Table 1, we can see that the PHD achieves an improvement of more than 30% in the RMSE with respect to the baseline algorithm. In Fig. 2, the raw data and $\hat{\theta}_{\mathrm{PHD}}$ output illustrate that the GM-PHD clearly ignores false detections due to the pruning procedure, which neglects MoG terms with small mixture weights. We set an optimal threshold level for the $\hat{\theta}_{WH-10/20}$ estimators due to its high sensitivity to this level. Since $\hat{\theta}_{\mathrm{PHD}}$ estimator can handle false detections as described above, it exhibits greater robustness across various threshold levels. Smaller DOA distances, such as in the "near" case, yield noisier DOA measurements, resulting in degradation in the RMSE performance of $\hat{\theta}_{\mathrm{PHD}}$ to the level of the baseline estimators. An illustration of one sample is presented in Fig. 1(b). The intermittent nature of the speech utterances clearly demonstrates the power of the proposed probabilistic framework, which enables the sources to appear/disappear at each batch of a new source.



(a) "distant" (b) "near"

Fig. 1: Estimated DOA trajectories.



(a) "distant" (b) "near"

Fig. 2: GM-PHD raw data and estimate trajectory.

5.3. Conclusions

We introduced a GM-PHD DOA tracking algorithm for multiple speakers in a reverberant environment. The proposed algorithm leverages speech sparsity in the STFT domain to estimate a candidate DOA for a single time frame together with a confidence measure. The candidate DOAs are fed into the GM-PHD filter, which manages the tracking mechanism and estimates speakers' DOAs for each batch. The results of our experimental study demonstrate a 30% improvement in tracking accuracy for the proposed algorithm. Recent research on state space models incorporates deep neural networks (DNNs) [20]. Such models can be adopted to extend our work.

6. REFERENCES

[1] C. Knapp and G. 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] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, *Microphone arrays: signal processing techniques and applications*, chapter Robust Localization in Reverberant Rooms, pp. 157–180, Springer Verlag, 2001.

[3] H. Ye and R. 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] B. Ottersten, M. Viberg, P. Stoica, and A. Nehorai, *Radar Array Processing*, chapter Exact and large sample ML techniques for parameter estimation and detection in array processing, Springer Verlag, 1993.

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

[6] N. Madhu and R. Martin, "A scalable framework for multiple speaker localization and tracking," in *International Workshop for Acoustic Echo Cancellation and Noise Control (IWAENC), Seattle, WA*, 2008.

[7] Elior Hadad and Sharon Gannot, "Multi-speaker direction of arrival estimation using SRP-PHAT algorithm with a weighted histogram," in *IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, 2018.

[8] S. Delikaris-Manias, D. Pavlidi, V. Pulkki, and A. Mouchtaris, "3D localization of multiple audio sources utilizing 2D DOA histograms," in *The 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1473–1477.

[9] Ayal Schwartz, Elior Hadad, Sharon Gannot, and Shlomo E Chazan, "Array configuration mismatch in deep doa estimation: Towards robust training," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023.

[10] Daniel Fejgin, Elior Hadad, Sharon Gannot, Zbyněk Koldovsky, and Simon Doclo, "Comparison of frequency-fusion mechanisms for binaural direction-of-arrival estimation for multiple speakers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 731–735.

[11] Elior Hadad and Sharon Gannot, "Maximum likelihood multi-speaker direction of arrival estimation utilizing a weighted histogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 586–590.

[12] M.I. Mandel, D.P.W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," *Advances in Neural Information Processing Systems*, vol. 19, pp. 953, 2007.

[13] Ofer Schwartz and Sharon Gannot, "Speaker tracking using recursive EM algorithms," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 2, pp. 392–402, 2014.

[14] D Michael Titterington, "Recursive parameter estimation using incomplete data," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 46, no. 2, pp. 257–267, 1984.

[15] Olivier Cappé and Eric Moulines, "On-line expectation–maximization algorithm for latent data models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 71, no. 3, pp. 593–613, 2009.

[16] R.P.S. Mahler, "Multitarget bayes filtering via first-order multitarget moments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1152–1178, 2003.

[17] B.-N. Vo and W.-K. Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4091–4104, 2006.

[18] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J. G Fiscus, and D.S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus," *NASA STI/Recon Technical Report N*, vol. 93, pp. 27403, 1993.

[19] Emanuël AP Habets, Israel Cohen, and Sharon 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, 2008.

[20] Tri Dao and Albert Gu, "Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality," *arXiv preprint arXiv:2405.21060*, 2024.