Three-Dimensional Gradient-Based Tracking of Multiple Sound Sources

Shaoheng Xu*, Wei-Ting Lai*, Yile (Angela) Zhang*, Jihui (Aimee) Zhang†*,
Amy Bastine*, Prasanga N. Samarasinghe*, and Thushara D. Abhayapala*
* The Australian National University, Australia
† The University of Queensland, Australia

***Abstract*—This paper presents a gradient-based algorithm for tracking multiple sound sources in three-dimensional (3-D) space under reverberant conditions. The proposed method, called 3-D Gradient-based Multiple-source Tracking (3D-GMT), operates directly in the continuous spatial domain and estimates source trajectories via a frequency-constrained loss function. Unlike conventional grid-based approaches that suffer from quantization errors and require post-processing filters, 3D-GMT jointly refines all source locations in parallel, implicitly smoothing trajectories and reducing cumulative error. Simulation experiments in reverberant environments demonstrate the robustness and accuracy of 3D-GMT under two representative motion scenarios: straight-line paths and complex spiral trajectories. The results indicate potential for applying the method to sound source tracking in reverberant environments with smooth, moderate-speed motion.**

I. Introduction

Sound source tracking (SST) plays an important role in many real-world applications. For instance, in robotics, SST enables a robot to localize and follow a person of interest, thereby enhancing other functions such as automatic speech recognition [1]. In audio augmented reality (AAR), accurate tracking of users is crucial for maintaining spatial consistency and improving the perceived realism of applications like smart conferencing systems [2].

Current research remains predominantly focused on static acoustic scenes (i.e., localization tasks), with only a few studies that focus on tracking moving sources. This is evident in the localization tasks of the LOCATA challenge [3] and the sound event localization and detection task of the DCASE challenge [4]. Although LOCATA evaluates performance for multiple moving sources, its results are reported primarily in terms of azimuth accuracy and for two speakers. Their baseline method used MUltiple SIgnal Classification (MUSIC) for source azimuth estimates which are used by the Kalman filter to update the source position predictions [3]. Various microphone arrays are used for tracking evaluation, including robot head, hearing aids, Eigenmike and the DICIT array.

Single moving source tracking has been explored using various combinations of localization and tracking techniques [5]–[10]. MUSIC-derived source position estimates have been combined with Kalman filtering through generalized eigenvalue decomposition [5]. Time difference of arrival estimates from generalized cross-correlation are used in the adaptive distributed particle filters [8] and conventional particle filters [9]. Steered-response-power beamforming feeds either a particle filter [6] or a diagonally unloaded beamformer paired with a Kalman filter [7], while ambisonic intensity histograms underpin the TRAMP tracker [11]. Probabilistic-data-association updates coupled to a cubature Kalman filter have also been explored, and evaluates source motion on a two-dimensional (2-D) plane [10].

Multi-source tracking has also been addressed. The TRAMP algorithm [11] is applicable to both single- and multi-source tracking. The multi-source tracking algorithm proposed in [12] estimates the source azimuth, elevation, and their angular velocities similarly using MUSIC as localization technique, and combined with sequential-Monte-Carlo probability-hypothesis-density (SMC-PHD) filtering that explicitly models clutter and missed detections. The random-finite-set SMC method of [13] also considers 2-D source localization of two simultaneous speakers, with a simulated straight-line motion using the image-source [14] model. They placed eight omnidirectional microphones near the wall which encloses the region of interest, and one microphone at the room center.

Existing approaches remain limited in two key respects. First, relatively few works have specifically addressed three-dimensional (3-D) localization and tracking of multiple moving sources in reverberant environments [15]. Second, most tracking methods often include a post-processing filter to stabilize the tracking trajectories, resulting in added complexity. In this paper, our proposed method tracks multiple sound sources in 3-D space and introduces three key contributions: 1) It enables continuous 3-D tracking, addressing the grid mismatch problem in discrete grid-based methods; 2) It jointly optimizes all source locations in parallel, avoiding cumulative errors; 3) It incorporates a frequency-domain constraint that provides implicit smoothing, eliminating the need for post-processing filters.

These contributions lay the groundwork for robust and reproducible multi-source tracking under realistic acoustic conditions. We simulate tracking scenarios in reverberant environments to approximate real-world acoustic conditions. Two types of source trajectories—straight lines and spiral paths—are used to evaluate performance.

II. Problem Formulation

Setup: Consider a 3-D acoustic environment affected by additive noise and reverberation. Suppose there are Q sound sources moving freely within a region of interest (ROI). The

979-8-3315-7206-8/25/$31.00

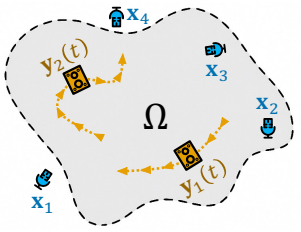

Fig. 1. Illustration of a general sound source tracking setup in a noisy and reverberant 3-D acoustic environment. Q moving sound sources (in orange) travel within the region of interest (ROI) Ω, with their trajectories $\mathbf{y}_q(t)$ shown as dashed orange lines. Triangles along the paths indicate movement direction. M microphones (in blue), fixed at positions $\mathbf{x}_m$, are spatially distributed within the environment.

position of the q-th source at continuous time τ is denoted by

$$\mathbf{y}_q(\tau) \triangleq \big(x_q(\tau),\, y_q(\tau),\, z_q(\tau)\big), \quad q = 1, 2, \ldots, Q,$$

defined with respect to a global origin O. Additionally, M microphones are fixed and spatially distributed at arbitrary locations within the environment, located at

$$\mathbf{x}_m \triangleq \big(x_m,\, y_m,\, z_m\big), \quad m = 1, 2, \ldots, M.$$

An illustration of this setup is shown in Fig. 1.

For simplicity, we assume that the number of active sources Q is known and that all microphone signals are time-synchronized.

Signal Model: To analyze signals in the time-frequency domain, the time axis is discretized into frames indexed by $t = 1, 2, \ldots, T$, where T is the total number of frames.

At each time frame t, let the q-th source emit a narrowband complex signal $w_q(t,k)$ at angular wavenumber $k = 2\pi f/c$, where f is the frequency and c is the speed of sound. The signal received at the m-th microphone is modeled as

$$P_m(t,k) = \sum_{q=1}^{Q} G\big(k, \mathbf{x}_m, \mathbf{y}_q(t)\big) \cdot w_q(t,k) + \varepsilon_m(t,k), \quad (1)$$

where $G(k, \mathbf{x}_m, \mathbf{y}_q(t))$ models the direct-path propagation from the q-th source to the m-th microphone, while $\varepsilon_m(t,k)$ represents additive noise, including both reverberation and system noise. Specifically, $G\big(k, \mathbf{x}_m, \mathbf{y}_q(t)\big)$ is assumed to follow the free-space point-source Green's function (normalized to the origin) [16], [17], given by:

$$G\big(k, \mathbf{x}_m, \mathbf{y}_q(t)\big) = \|\mathbf{y}_q(t)\|_2 \cdot e^{-ik\|\mathbf{y}_q(t)\|_2} \cdot \frac{e^{ik\|\mathbf{x}_m - \mathbf{y}_q(t)\|_2}}{4\pi\|\mathbf{x}_m - \mathbf{y}_q(t)\|_2}. \quad (2)$$

Stacking the microphone signals into a vector yields:

$$\mathbf{p}(t,k) = \big[P_1(t,k) \quad P_2(t,k) \quad \cdots \quad P_M(t,k)\big]^\top \in \mathbb{C}^{M\times 1}, \quad (3)$$

where $(\cdot)^\top$ denotes transpose operation.

Estimated Trajectories: Let $\hat{\mathbf{y}}_q(t) \triangleq \big(\hat{x}_q(t), \hat{y}_q(t), \hat{z}_q(t)\big)$ denote the estimated position of source q at time t, which is expected to be close to the true position $\mathbf{y}_q(t)$. The corresponding estimated source weight at frequency k is denoted by $\hat{w}_q(t,k)$, and the stacked vector of estimates is:

$$\hat{\mathbf{w}}(t,k) = \big[\hat{w}_1(t,k) \quad \hat{w}_2(t,k) \quad \cdots \quad \hat{w}_Q(t,k)\big]^\top \in \mathbb{C}^{Q\times 1}. \quad (4)$$

Using (2), the free-field transfer functions from the estimated source positions to the microphones can be assembled into a dictionary matrix $\hat{\mathbf{G}}(t,k) \in \mathbb{C}^{M\times Q}$. The estimated sound field generated at all microphones is then:

$$\hat{\mathbf{p}}(t,k) = \hat{\mathbf{G}}(t,k) \cdot \hat{\mathbf{w}}(t,k). \quad (5)$$

Objective: Unlike sound source localization (SSL), sound source tracking requires incorporating spatial information from past observations via spatiotemporal models of the source dynamics to obtain smooth and coherent trajectories [3].

We formulate the tracking problem as estimating the source positions $\hat{\mathbf{y}}_q(t)$ and signals $\hat{w}_q(t,k)$ such that the mismatch between the observed and estimated sound fields is minimized. Specifically:

$$\begin{aligned} &\min_{\hat{\mathbf{y}}(t),\, \hat{\mathbf{w}}(t,k)} \|\mathbf{p}(t,k) - \hat{\mathbf{p}}(t,k)\|_2^2, \\ &\text{subject to} \quad \hat{\mathbf{y}}_q(t) = f\big(\hat{\mathbf{y}}_q(t-1),\, \mathbf{p}(t,k)\big), \\ &\qquad\qquad\ \ \hat{\mathbf{y}}_q(t) \approx \mathbf{y}_q(t), \end{aligned} \quad (6)$$

where the update function $f(\cdot)$ models the temporal dynamics based on the previous state and current observation. Although the ground truth $\mathbf{y}_q(t)$ is not directly accessible, the formulation ensures that $\hat{\mathbf{y}}_q(t)$ closely approximates it through minimization of the observed signal error.

Initialization: Since the focus of this work is on tracking, we assume that coarse initial estimates of source positions are available from an existing SSL method. Examples include Sparse Bayesian Learning (SBL) [18], Parametric Dictionary Learning [19], Time Difference of Arrival (TDOA)-based approaches [20], and Group Orthogonal Matching Pursuit (G-COMP) [21]. The initial position estimates are denoted as

$$\hat{\mathbf{y}}(t{=}0) = \big[\hat{\mathbf{y}}_1(t{=}0) \quad \hat{\mathbf{y}}_2(t{=}0) \quad \cdots \quad \hat{\mathbf{y}}_Q(t{=}0)\big]^\top \in \mathbb{R}^{Q\times 3}. \quad (7)$$

After obtaining the initial coarse location estimates, we compute the initial signal weights for each frequency k using the following least-squares formulation:

$$\hat{\mathbf{w}}(t,k) = \hat{\mathbf{G}}(t,k)^\dagger \cdot \mathbf{p}(t,k), \quad (8)$$

where $(\cdot)^\dagger$ denotes the Moore–Penrose pseudo-inverse. By substituting $t = 0$ into (8), we obtain the initial estimates $\hat{\mathbf{w}}(t{=}0,k)$ associated with the initial estimated locations $\hat{\mathbf{y}}(t{=}0)$. These initial estimates serve as inputs to the tracking algorithm described in Section III.

III. Proposed Method

We propose a gradient descent-based framework for continuously tracking multiple sound sources in the 3-D spatially continuous domain. The proposed algorithm, called *3-D Gradient-based Multiple-source Tracking* (3D-GMT), is inspired by the Point Neuron Learning (PNL) method [22], originally developed for narrowband sound field modeling.

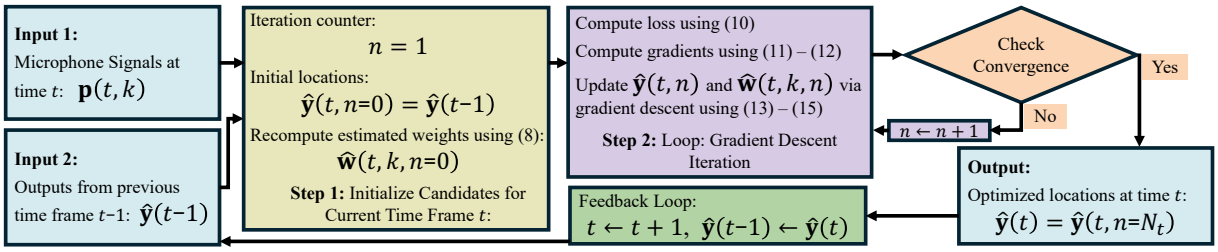

Fig. 2. Flowchart of the proposed 3D-GMT.

In [22], estimated source positions (termed point neurons) are iteratively updated via gradient descent to refine both spatial locations and signal weights. Building on this idea, we extend the approach to broadband scenarios and formulate 3D-GMT for online wideband sound source tracking, where the gradient of the loss with respect to estimated positions drives the sources to follow their true trajectories.

A. Iterative Online Tracking Framework

The overall pipeline of 3D-GMT is illustrated in Fig. 2. As discussed in Section II, the algorithm begins with initial estimates of source locations $\hat{\mathbf{y}}(t{=}0)$ and corresponding weights $\hat{\mathbf{w}}(t{=}0,k)$ for each frequency k, initialized using (8).

For each subsequent time frame $t = 1, \ldots, T$, the algorithm initializes the estimated positions using the final estimates from the previous frame $t{-}1$:

$$\hat{\mathbf{y}}(t, n{=}0) = \hat{\mathbf{y}}(t{-}1), \tag{9}$$

where n is the iteration index. The weights $\hat{\mathbf{w}}(t,k,n{=}0)$ for each frequency are re-initialized using (8).

3D-GMT then performs gradient descent to iteratively refine the Q estimated sources, adjusting their positions to better match the observed sound field $\mathbf{p}(t,k)$. After N_t iterations, the updated positions $\hat{\mathbf{y}}(t, n{=}N_t)$ are used as the final estimates for time frame t. This process continues until $t = T$, yielding the estimated trajectories $\hat{\mathbf{y}}(1{:}T)$ for all sources.

Details of the gradient descent update process are provided in the following Section III-B.

B. Location Refinement via Gradient Descent

Loss Function: At each time frame t and frequency k, the loss function at iteration n is defined as the squared ℓ_2-norm between the observed and estimated microphone signals [22]:

$$\mathcal{L}(t,k,n) = \|\mathbf{p}(t,k) - \hat{\mathbf{p}}(t,k)\|_2^2 . \tag{10}$$

Gradient Computation: The gradients of $\mathcal{L}(t,k,n)$ with respect to the signal weights and spatial coordinates of the q-th estimated source are:

$$\begin{aligned}\nabla_{\hat{w}_q}(t,k,n) &= \frac{\partial \mathcal{L}(t,k,n)}{\partial \hat{w}_q(t,k,n)^*} \\ &= \sum_{m=1}^{M} \Big(\hat{P}_m(t,k,n) - P_m(t,k)\Big) \\ &\quad \times \frac{D_q(t,n)}{D_q^m(t,n)} \times e^{-ik\left(D_q^m(t,n) - D_q(t,n)\right)},\end{aligned} \tag{11}$$

$$\begin{aligned}\nabla_{\hat{\theta}_q}(t,k,n) &= \frac{\partial \mathcal{L}(t,k,n)}{\partial \hat{\theta}_q(t,k,n)} \\ &= \sum_{m=1}^{M} 2\,\mathbb{R}e\Bigg\{ \big(\hat{P}_m(t,k,n) - P_m(t,k)\big)^* \\ &\quad \times \hat{w}_q(t,k,n) \times \frac{D_q(t,n)}{D_q^m(t,n)} e^{ik\left(D_q^m(t,n) - D_q(t,n)\right)} \\ &\quad \times \left(\frac{-(ikD_q(t,n)-1)}{D_q(t,n)^2} \hat{\theta}_q(t,k,n) \right. \\ &\quad \left. + \frac{(ikD_q^m(t,n)-1)}{D_q^m(t,n)^2} \big(\hat{\theta}_q(t,k,n) - \theta_m\big) \right) \Bigg\},\end{aligned} \tag{12}$$

where $\theta \in \{x, y, z\}$ denotes a spatial coordinate axis, and $\hat{\theta}_q$ and θ_m represent the θ-coordinate of the q-th estimated source and the m-th microphone, respectively. Here, $D_q(t,n) = \|\hat{\mathbf{y}}_q(t,n)\|_2$ is the distance from the q-th estimated source to the origin, and $D_q^m(t,n) = \|\hat{\mathbf{y}}_q(t,n) - \mathbf{x}_m\|_2$ is the distance between the q-th estimated source and the m-th microphone.

For broadband sources composed of K frequency components, we follow the simultaneous optimization principle [23] and compute a weighted sum of per-frequency gradients:

$$\nabla_{\hat{\theta}_q}(t,n) = \sum_{k=1}^{K} \alpha(k) \cdot \nabla_{\hat{\theta}_q}(t,k,n), \quad \sum_{k=1}^{K} \alpha(k) = 1. \tag{13}$$

Parameter Updates: The update rules for spatial coordinates and signal weights are:

$$\hat{\theta}_q(t, n{+}1) = \hat{\theta}_q(t,n) - \eta_\theta \cdot \nabla_{\hat{\theta}_q}(t,n), \tag{14}$$

$$\hat{w}_q(t,k,n{+}1) = \hat{w}_q(t,k,n) - \eta_w \cdot \nabla_{\hat{w}_q}(t,k,n), \tag{15}$$

where η_θ and η_w are learning rates for coordinate and weight updates, respectively. Please note that the estimated source weights are still updated independently for each frequency following (15).

We also define the total loss across all frequencies:

$$\mathcal{L}_{\text{sum}}(t,n) = \sum_{k=1}^{K} \mathcal{L}(t,k,n). \tag{16}$$

Output: The gradient descent loop at time frame t terminates when $|\mathcal{L}_{\text{sum}}(t,n) - \mathcal{L}_{\text{sum}}(t,n{-}1)| \le \zeta$, where ζ is a convergence threshold. The final estimated source positions are:

$$\hat{\mathbf{y}}(t) = \hat{\mathbf{y}}(t, n{=}N_t) \triangleq \Big[\hat{\theta}_1(t,n), \ldots, \hat{\theta}_Q(t,n)\Big]^\top . \tag{17}$$

C. Novelty of 3D-GMT

The proposed 3D-GMT method introduces several innovations:

1) **Continuous 3-D tracking**: Unlike grid-based methods that suffer from quantization errors and grid mismatch [24], [25], 3D-GMT tracks sound sources in a continuous spatial domain.

2) **Parallel and temporal optimization**: 3D-GMT updates all Q estimated sources in parallel at each time frame, mitigating error accumulation across sources. Additionally, each time frame uses results from the previous frame as initialization, enabling online, frame-by-frame tracking without requiring access to future frames or the entire dataset.

3) **Implicit smoothing via frequency constraints**: Due to the oscillatory behavior of exponential terms in (11) and (12), the loss function (10) is highly non-convex with multiple local minima. These properties naturally limit the maximum displacement of estimated sources in each time frame—much like traversing a narrow valley where steep terrain constrains movement—thereby preventing large updates that could lead to tracking errors. This behavior acts as an implicit regularization mechanism, reducing the need for post-processing filters such as multiple model Kalman filtering (MMKF) [15] and mitigating error propagation over time.

IV. Simulation Validation

This section presents a simulation-based evaluation of the proposed 3D-GMT algorithm. Its performance is compared against three baseline tracking methods introduced in Section IV-C, with results and analysis detailed in Section IV-D.

A. Simulation Setup

We simulate a rectangular room of size $5\,\mathrm{m} \times 4.435\,\mathrm{m} \times 3\,\mathrm{m}$ with reverberation, as illustrated in Fig. 3. The setup includes a 4×2 wall-mounted microphone array on one wall, a second identical array on the opposite wall, and a 2×2 microphone array mounted at the center of the ceiling. In total, 20 microphones are deployed, representing a feasible configuration for evaluating the proposed tracking method in a realistic acoustic environment. The signal-to-noise ratio (SNR) at each microphone is set to 30 dB.

The ROI is defined as a $3.5\,\mathrm{m} \times 2.935\,\mathrm{m} \times 2\,\mathrm{m}$ volume centered within the room. We simulate $Q = 3$ moving sound sources constrained within the ROI. Two different scenarios are considered: in the first, sources follow straight-line trajectories; in the second, they follow distinct spiral paths.

To simplify the analysis, all simulations are performed in the frequency domain. Along the trajectory of each source, we sample $T = 50$ spatial points, each corresponding to a time frame in the STFT domain.

At each time frame t, we select 7 frequency components: $[300, 400, 500, 750, 900, 1000, 1200]$ Hz. For each frequency k, the corresponding signal weights $\mathbf{w}(t, k)$ are randomly sampled from a uniform distribution within $[0.5, 1]$.

The image-source method [14] is used to compute the acoustic transfer function from each source to each microphone at every time frame t, according to the model in (2). All wall reflection coefficients are set to 0.7, and the reflection order is set to 15, resulting in a simulated reverberation time (T60) of approximately 0.176 seconds.

Finally, we compute the microphone observation vector $\mathbf{p}(t, k)$ using (1) and (3), which serves as input to the tracking algorithms under evaluation.

For initialization, we adopt the G-COMP [21] algorithm, as it offers a good balance between speed and accuracy for coarse localization. G-COMP is a fast but coarse on-grid method, making it well-suited for providing initial estimates to the continuous refinement process of the proposed 3D-GMT. Specifically, we construct a $40 \times 40 \times 40$ uniform grid covering the entire ROI and set the sparsity level to 3 to select three coarse initial positions corresponding to the three true sound sources. These estimates are then passed to 3D-GMT and refined frame by frame through iterative tracking. The empirically chosen ζ for 3D-GMT is 1×10^{-5}, with larger values yielding faster convergence at minor accuracy loss.

B. Evaluation Metrics

To assess the accuracy of sound source tracking, we define an evaluation metric that computes the average tracking error across all sound sources at each time frame t, separately for the x-, y-, and z-axes. The tracking error $\mathbf{TE}_\theta(t)$ for each axis is defined as:

$$\mathbf{TE}_\theta(t) = \frac{1}{Q} \sum_{q=1}^{Q} \left\| \hat{\theta}_q(t) - \theta_q(t) \right\|_2 . \tag{18}$$

C. Comparison Methods

We compare the proposed 3D-GMT algorithm with three baseline methods: a TDOA-based method with multi-model Kalman filtering (TDOA-MMKF) [15], a G-COMP-based method with MMKF (GCOMP-MMKF), and an SBL-based method with MMKF (SBL-MMKF). Two additional MUSIC-based baseline methods [5], [12] were also evaluated; however, their performance was consistently lower than that of TDOA-MMKF in the current simulation setup. Therefore, their results are omitted from this paper for brevity.

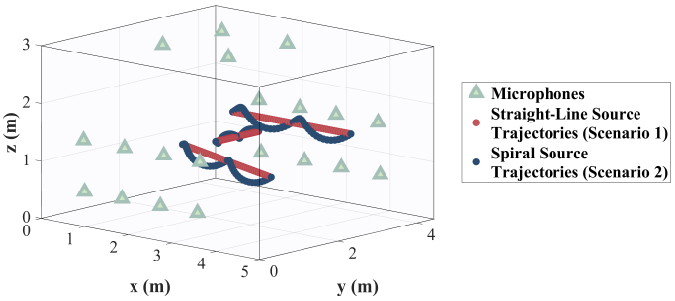

Fig. 3. Simulation environment setup with three moving sound sources. Green triangles represent microphone positions. Red circles indicate the true sound source trajectories in the straight-line scenario, while dark blue circles represent the trajectories in the spiral-path scenario, both plotted across discrete time frames.

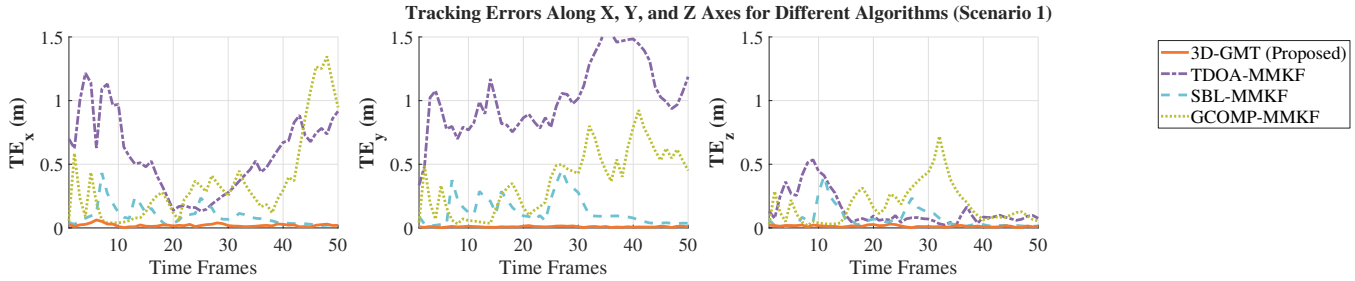

Fig. 4. Tracking errors $\mathbf{TE}_\theta(t)$ along the x-, y-, and z-axes for different algorithms in Scenario1 (see Fig.3). The orange solid line represents the proposed 3D-GMT; the purple dash-dot line corresponds to TDOA-MMKF; the light blue dashed line to SBL-MMKF; and the green dotted line to GCOMP-MMKF.

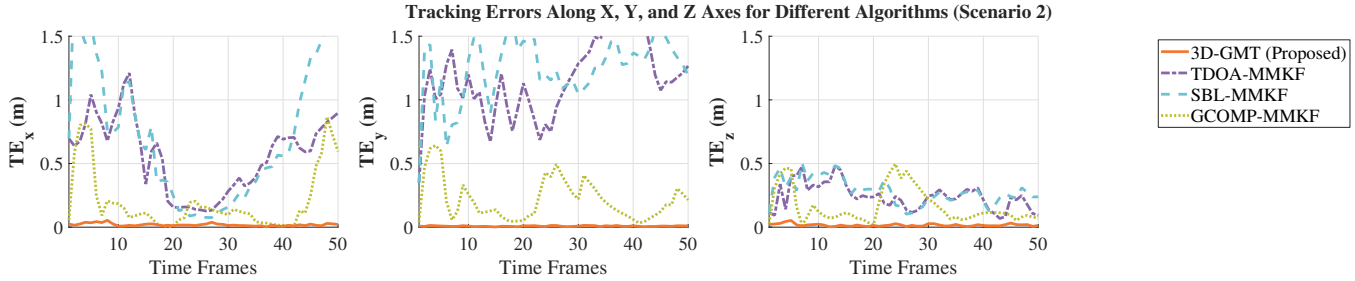

Fig. 5. Tracking errors $\mathbf{TE}_\theta(t)$ along the x-, y-, and z-axes for different algorithms in Scenario2 (see Fig.3). The orange solid line represents the proposed 3D-GMT; the purple dash-dot line corresponds to TDOA-MMKF; the light blue dashed line to SBL-MMKF; and the green dotted line to GCOMP-MMKF.

In TDOA-MMKF, TDOA estimates are used to extract noisy observations at each time frame. These observations are then smoothed using a MMKF, which integrates predictions from three motion models: stationary, constant velocity, and constant acceleration. GCOMP-MMKF and SBL-MMKF adopt the same tracking framework but replace the TDOA-based observation step with G-COMP [21], a greedy source localization method, and SBL [18], a probabilistic approach, respectively.

For all MMKF-based methods, the time interval between two consecutive observations, d_t, is set to 0.2 seconds. The observation noise covariance is set to 0.01, and the process (system) noise variance is set to 10^{-3}.

D. Results and Discussion

Figures 4 and 5 show the tracking errors along the x-, y-, and z-axes for two source trajectories: Scenario 1 (straight-line path) and Scenario 2 (spiral path), respectively.

In the simpler Scenario 1 (Fig. 4), 3D-GMT consistently outperforms all baseline methods. Its tracking errors along all three axes remain below 0.06 m, with most frames exhibiting errors close to 0.01 m. This demonstrates the algorithm's high accuracy and robustness. In contrast, the errors of the other algorithms are significantly larger across all dimensions.

Scenario 2 presents a more complex and challenging condition, as the spiral trajectory involves continuous changes in both velocity and acceleration directions. Despite this, as shown in Fig. 5, 3D-GMT maintains near-zero tracking errors across all three axes. Other methods, particularly along the x- and y-axes, exhibit noticeably larger errors. Notably, while SBL-MMKF performs reasonably well in Scenario 1, its performance degrades substantially in Scenario 2, indicating a lack of robustness to complex motion patterns.

In both scenarios, $\mathbf{TE}_x(t)$ and $\mathbf{TE}_y(t)$ are higher than $\mathbf{TE}_z(t)$ for the comparison methods, likely due to fewer microphones and greater source motion along the x- and y-axes. In contrast, 3D-GMT maintains low errors across all directions, highlighting its robustness in 3D tracking.

Overall, the results demonstrate the accuracy and robustness of 3D-GMT for 3-D source tracking. By estimating trajectories directly in the continuous spatial domain, 3D-GMT avoids grid mismatch errors that affect grid-based methods such as TDOA-MMKF, GCOMP-MMKF, and SBL-MMKF. Furthermore, the oscillatory nature of the exponential terms in the loss function introduces a natural smoothing effect, guiding the solution toward local minima and preventing abrupt deviations. As a result, 3D-GMT does not require additional post-processing filters such as MMKF [15].

V. Conclusion

We proposed a novel algorithm for multiple sound source tracking, named 3D-GMT. Operating in the continuous 3-D spatial domain, it jointly optimizes source locations in parallel, thereby mitigating grid mismatch present in previous methods and avoiding error accumulation across sources. Moreover, 3D-GMT uses the oscillatory nature of its loss function as an implicit smoothing mechanism, eliminating the need for post-processing filters. Simulation results confirm its robustness and accuracy in 3-D tracking under reverberant conditions and with complex source trajectories. Future work will focus on: 1) Generalization to an unknown number of sources; 2) Integrating re-identification algorithms for closely spaced

sources; and 3) Experimental validation in real-world environments, including investigations of microphone number, array geometry, discontinuous source signals such as speech, and various RT60 conditions.

REFERENCES

[1] J. M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auton. Syst.*, vol. 55, no. 3, pp. 216–228, 2007.

[2] A. Privitera, F. Fontana, and M. Geronazzo, "On the effect of user tracking on perceived source positions in mobile audio augmented reality," in *Proc. Biannu. Conf. Ital. SIGCHI Chapter*, 2023, pp. 1–9.

[3] C. Evers, H. W. Löllmann, H. Mellmann, *et al.*, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1620–1643, 2020.

[4] D. Diaz-Guerra, A. Politis, P. Sudarsanam, *et al.*, "Baseline models and evaluation of sound event localization and detection with distance estimation in dcase 2024 challenge," in *Workshop Detect. Classif. Acoust. Scenes Events*, DCASE, 2024, pp. 41–45.

[5] K. Nakadai, K. Itoyama, K. Hoshiba, and H. G. Okuno, "MUSIC-based sound source localization and tracking for tasks 1 and 3," in *Proc. LOCATA Challenge Workshop-Satell. Event IWAENC*, 2018, pp. 1–5.

[6] R. Lebarbenchon, E. Camberlein, D. Di Carlo, C. Gaultier, A. Deleforge, and N. Bertin, "Evaluation of an open-source implementation of the srp-phat algorithm within the 2018 locata challenge," in *Proc. LOCATA Challenge Workshop-Satell. Event IWAENC*, 2018.

[7] D. Salvati, C. Drioli, and G. L. Foresti, "Localization and tracking of an acoustic source using a diagonal unloading beamforming and a kalman filter," *arXiv preprint arXiv:1812.01521*, 2018.

[8] Y. Jing, Z. Li, and C. Liu, "Acoustic source tracking based on adaptive distributed particle filter in distributed microphone networks," *Signal Process.*, vol. 154, pp. 375–386, 2019.

[9] X. Qian, A. Cavallaro, A. Brutti, and M. Omologo, "LOCATA challenge: Speaker localization with a planar array," *CoRR*, vol. abs/1901.08983, 2019.

[10] Y. Chen, Y. Cao, and R. Wang, "Acoustic source tracking based on probabilistic data association and distributed cubature kalman filtering in acoustic sensor networks," *Sensors*, vol. 22, no. 19, p. 7160, 2022.

[11] S. Kitić and A. Guérin, "Tramp: Tracking by a real-time ambisonic-based particle filter," in *IEEE-AASP Challenge Acoust. Source Localizat. Track.*, 2018.

[12] Y. Liu, W. Wang, and V. Kilic, "Intensity particle flow smc-phd filter for audio speaker tracking," *arXiv preprint arXiv:1812.01570*, 2018.

[13] Y. Guo, H. Zhu, and X. Dang, "Tracking multiple acoustic sources by adaptive fusion of tdoas across microphone pairs," *Digital Signal Process.*, vol. 106, p. 102 853, 2020.

[14] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.

[15] Z. Liang, X. Ma, and X. Dai, "Robust tracking of moving sound source using multiple model Kalman filter," *Appl. Acoust.*, vol. 69, no. 12, pp. 1350–1355, 2008.

[16] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Elsevier, 1999.

[17] L. I. Birnie, T. D. Abhayapala, V. Tourbabin, and P. N. Samarasinghe, "Mixed source sound field translation for virtual binaural application with perceptual validation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1188–1203, 2021.

[18] A. Xenaki, J. Bünsow Boldt, and M. G. Christensen, "Sound source localization and speech enhancement with sparse Bayesian learning beamforming," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. 3912–3921, 2018.

[19] K. You, W. Guo, T. Peng, Y. Liu, P. Zuo, and W. Wang, "Parametric sparse bayesian dictionary learning for multiple sources localization with propagation parameters uncertainty," *IEEE Trans. Signal Process.*, vol. 68, pp. 4194–4209, 2020.

[20] H. Do and H. F. Silverman, "A method for locating multiple sources from a frame of a large-aperture microphone array data without tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 301–304.

[21] S. Xu, J. Zhang, T. D. Abhayapala, A. Bastine, and P. N. Samarasinghe, "Iterative and complex orthogonal matching pursuit for broadband sparse sound field reconstruction," in *Int. Workshop Acoust. Signal Enhanc.*, 2024, pp. 195–199.

[22] H. Bi and T. D. Abhayapala, "Point neuron learning: A new physics-informed neural network architecture," *EURASIP J. Audio Speech Music Process.*, vol. 2024, no. 1, p. 56, 2024.

[23] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.

[24] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2182–2195, 2011.

[25] Y. Yang, Z. Chu, Y. Yang, and S. Yin, "Two-dimensional Newtonized orthogonal matching pursuit compressive beamforming," *J. Acoust. Soc. Amer.*, vol. 148, no. 3, pp. 1337–1348, Sep. 2020.