2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

I-ASM: Iterative Acoustic Scene Mapping for Enhanced Robot Auditory Perception in Complex Indoor Environments

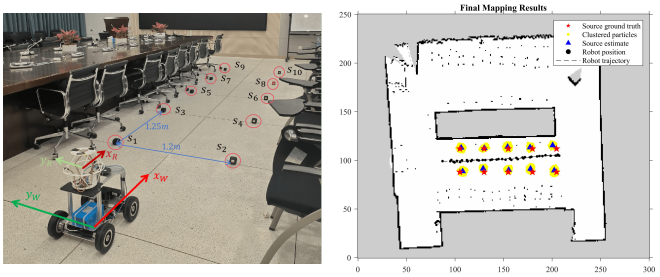
Linya Fu, Yuanzheng He, Jiang Wang, Xu Qiao, and He Kong

***Abstract*— This paper addresses the challenge of acoustic scene mapping (ASM) in complex indoor environments with multiple sound sources. Unlike existing methods that rely on prior data association or SLAM frameworks, we propose a novel particle filter-based iterative framework, termed I-ASM, for ASM using a mobile robot equipped with a microphone array and LiDAR. I-ASM harnesses an innovative "implicit association" to align sound sources with Direction of Arrival (DoA) observations without requiring explicit pairing, thereby streamlining the mapping process. Given inputs including an occupancy map, DoA estimates from various robot positions, and corresponding robot pose data, I-ASM performs multi-source mapping through an iterative cycle of "Filtering-Clustering-Implicit Associating". The proposed framework has been tested in real-world scenarios with up to 10 concurrent sound sources, demonstrating its robustness against missing and false DoA estimates while achieving high-quality ASM results. To benefit the community, we open-source all the codes and data at `https://github.com/AISLAB-sustech/Acoustic-Scene-Mapping`**

I. Introduction

Acoustic scene mapping (ASM) aims at mapping spatial positions of sound sources in the environment [1], and is essential for many robot audition applications [2]-[3]. In microphone array-based robot audition systems [4]-[7], the compact spacing between microphones typically only allows for direction of arrival (DoA) estimation when performing sound source localization (SSL) tasks, lacking the distance information [8]-[9]. Hence, ASM using mobile robots is challenging, especially when there are multiple sound sources.

Prior research can be categorized into three branches. The first category is based on the auditory evidence grid [10], where the environment is divided into uniform grids, and DoA estimates are converted into likelihood representations, using Bayesian updates to create a probability distribution map of sound sources in the environment. Improved methods in [11]-[13] employ acoustic ray casting to create a probabilistic sound map of geometric structures in the environment but are limited by the requirement that all sound sources must be laser-detected. The second category adopts simultaneous localization and mapping (SLAM) frameworks [1], [14]-[15] to jointly estimate the location of sound sources and the robot. However, the multipath effects and reverberation inherent in sound signals could lead to notable inaccuracies when used alone for SLAM. The third category hinges on data association [16]-[17], which transforms the complex multi-source mapping problem into multiple single-source mapping problems by associating observations to the corresponding source, rendering triangulation methods viable. However, in real-world applications, data association suffers from challenges such as missing detections, false alarms, the indeterminate number of sources, etc [18].



(a) (b)

Fig. 1: (a) Experimental scene with 10 sound sources and (b) corresponding map result at 0.05m/pixel resolution

In this paper, we introduce a novel particle filter-based iterative ASM framework (I-ASM) for complex indoor scenarios with multiple sound sources. As shown in Fig. 1, I-ASM effectively maps a scene with closely spaced sources by integrating an occupancy map, DoA estimates, and corresponding robot poses. Different from conventional multi-target tracking methods that require explicit data association (necessitating the cumbersome prior matching of new observations to existing tracks), I-ASM employs an iterative "Filtering-Clustering-Implicit Associating" strategy to associate DoA observations to the estimated sound source locations automatically during the mapping process, enhancing mapping efficiency.

Our contributions include: 1) The development of a novel particle filter-based ASM framework tailored for scenarios with multiple sound sources; 2) The introduction of "implicit associating" for streamlined, automated matching of DoA measurements to sound sources; 3) The enhancement of robustness against inaccuracies in DoA estimates. I-ASM is applicable to any mobile robot with a microphone array and LiDAR, and its effectiveness has been validated through practical experiments. Recent works can deal with only a few (usually between 1-3) sound sources [19]-[21], while in our experiment, I-ASM is able to handle more complex scenarios with 10 concurrent sources and maintain desirable accuracy.

This work was supported by the Science, Technology, and Innovation Commission of Shenzhen Municipality, China, under Grant No. ZDSYS20220330161800001, the Shenzhen Science and Technology Program under Grant No. KQTD20221101093557010, and the National Natural Science Foundation of China under Grant No. 62350055. The authors are with the Shenzhen Key Laboratory of Control Theory and Intelligent Systems, Southern University of Science and Technology, Shenzhen 518055, China. Emails:[12232297;12132259;12132297;12232290]@mail.sustech.edu.cn; kongh@sustech.edu.cn.

II. Preliminaries

We consider a two-dimensional indoor scenario with H static sound sources and a mobile robot. The robot is equipped with a microphone array and a LiDAR, navigating within the environment. As shown in Fig. 2, the world coordinate system $\{\boldsymbol{W}\} = (x_W, y_W)$ and the robot coordinate system $\{\boldsymbol{R}\} = (x_R, y_R)$ are defined, with $\{\boldsymbol{R}\}$ centered at the microphone array's centroid and its x_R axis aligned with the robot's linear velocity. Time is discretized into steps $k = 0, 1, 2, \ldots, K$. At time step k, the robot's pose is represented by its 2D Cartesian position and yaw angle with respect to (w.r.t.) $\{\boldsymbol{W}\}$:

$$\boldsymbol{r}_k = [x_{r,k}, y_{r,k}, \theta_{r,k}]. \tag{1}$$

The set of sound source positions is denoted by:

$$\boldsymbol{\mathcal{S}} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \cdots, \boldsymbol{s}_H\}, \tag{2}$$

with each position

$$\boldsymbol{s}_h = (x_{s,h}, y_{s,h}) \quad h = 1, 2, \cdots, H. \tag{3}$$

As to be introduced in Section IV, the DoA estimates given by the SSL algorithm are prioritized based on peak angular spectrum intensities, typically corresponding to stronger sound source signals. Since the actual number of sound sources H is unknown, we only consider the first N DoA estimates (azimuth only), denoted as:

$$\boldsymbol{z}_k^R = [\phi_{1,k}^R, \phi_{2,k}^R, \cdots, \phi_{N,k}^R]. \tag{4}$$

These DoA estimates are transformed into the world coordinate system to avoid discontinuities in subsequent processes:

$$\boldsymbol{z}_k^W = [\phi_{1,k}^W, \phi_{2,k}^W, \cdots, \phi_{N,k}^W], \tag{5}$$

where

$$\phi_{n,k}^W = \Theta(\phi_{n,k}^R + \theta_{r,k}) \in [-\pi, \pi]. \quad n = 1, 2, \cdots, N. \tag{6}$$

Note that $\boldsymbol{z}_k^W$ at each time step results in potential incomplete observations due to the above selection strategy. Nonetheless, our proposed algorithm demonstrates robustness against such missing detections, which will be elaborated in Section III and IV.

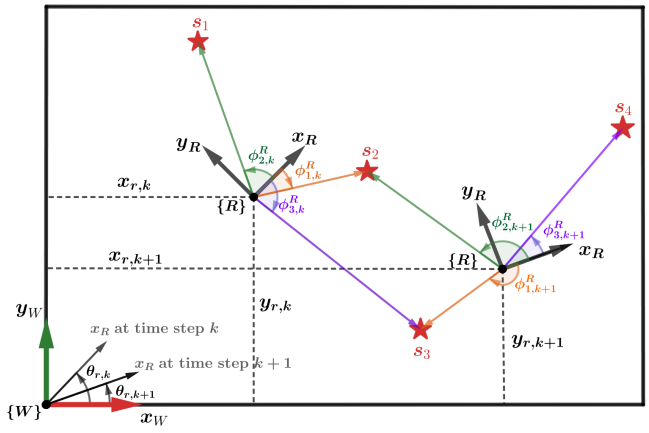


Fig. 2: Schematic diagram of coordinate systems

III. The Proposed Framework

A. System Overview

Our proposed I-ASM framework, as shown in Fig. 3, begins with the robot using SLAM to create a 2-D occupancy grid map for self-localization. The robot then explores the environment, capturing audio data at an observation point at time step k. The SSL algorithm generates N azimuth estimates, which are recorded in the k-th row of the DoA Estimates Table $\boldsymbol{\Phi}$, i.e., $\boldsymbol{\Phi}_k = \boldsymbol{z}_k^W$. Simultaneously, the estimated robot pose is logged in the k-th row of the Pose Estimates Table $\boldsymbol{\Omega}$, i.e., $\boldsymbol{\Omega}_k = \boldsymbol{r}_k$. The above process will be repeated as the robot moves to the next observation point.

During data collection, the I-ASM algorithm uses the occupancy map, $\boldsymbol{\Omega}$, and $\boldsymbol{\Phi}$ as inputs. To avoid prior data association, I-ASM sequentially estimates sound sources through an iterative "Filtering-Clustering-Implicit Associating" cycle. Firstly, the particle filtering step narrows down likely source locations, followed by DBSCAN clustering for refinement. The implicit observation association step then updates $\boldsymbol{\Phi}$ according to clustering results, which, along with $\boldsymbol{\Omega}$, guides the search for subsequent sources. The cycle concludes when clustering fails, yielding the final mapping results:

$$\hat{\boldsymbol{\mathcal{S}}} = [\hat{\boldsymbol{s}}_1, \hat{\boldsymbol{s}}_2, \cdots, \hat{\boldsymbol{s}}_i, \cdots, \hat{\boldsymbol{s}}_I], \quad i = 1, 2, \cdots, I, \tag{7}$$

where $\hat{\boldsymbol{s}}_i$ is the estimated location of the i-th source, and I is the total number of sources estimated.

B. Particle Filtering

I-ASM first utilizes particle filtering [22] to infer the probability distribution of sound source positions. At time step k, the particle set $\boldsymbol{\mathcal{P}}_k$ is defined as:

$$\boldsymbol{\mathcal{P}}_k = \{\boldsymbol{p}_k^{[1]}, \boldsymbol{p}_k^{[2]}, \cdots, \boldsymbol{p}_k^{[m]}, \cdots, \boldsymbol{p}_k^{[M]}\}, \tag{8}$$

where $\boldsymbol{p}_k^{[m]} = (x_{p,k}^{[m]}, y_{p,k}^{[m]})$ is the state of the m-th particle, representing a potential sound source position w.r.t. $\{\boldsymbol{W}\}$.

1) Particle Initialization: Initially, the particle set $\boldsymbol{\mathcal{P}}_0$ is generated by randomly selecting M particles from all grid cells of the occupancy map:

$$M = \lfloor n/D \rfloor, \tag{9}$$

where n is the count of grid cells, D a positive constant influencing particle density, and $\lfloor \cdot \rfloor$ the floor function. Lower D values increase particle density, enhancing accuracy but raising computation complexity, and vice versa. All particles are initialized with equal weights:

$$w_0^{[m]} = 1/M, \quad m = 1, 2, \cdots, M. \tag{10}$$

2) State Update: We update the particle states using a random walk model to reflect the uncertainty in static sound source systems. Specifically, we apply small Gaussian noise to the current states to project them to the next time step:

$$\boldsymbol{p}_{k+1}^{[m]} = \boldsymbol{p}_k^{[m]} + \boldsymbol{v}_k \tag{11}$$

where $\boldsymbol{v}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$ with $\boldsymbol{Q} = \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \sigma_p^2 \end{bmatrix}$.

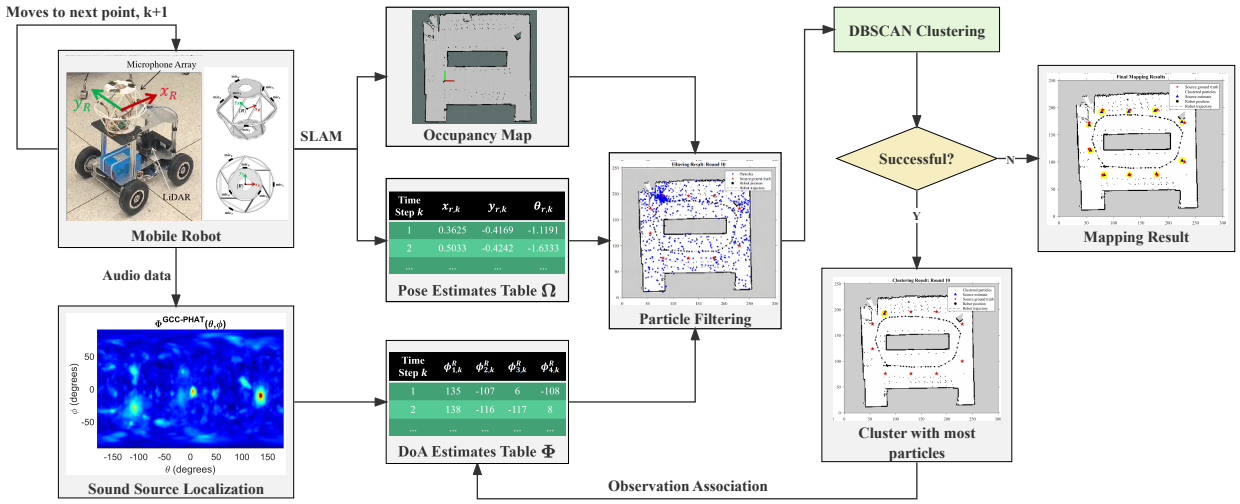


Fig. 3: Overview of the proposed I-ASM framework

3) Weight Update: Particle weights are updated based on the transformed DoA estimates $\boldsymbol{z}_k^W$ and the robot pose $\boldsymbol{r}_k$.

Particle's assumed azimuth angle $\theta_k^{[m]}$ is calculated by:

$$\theta_k^{[m]} = \tan^{-1}\left(\frac{y_k^{[m]} - y_{r,k}}{x_k^{[m]} - x_{r,k}}\right), \quad m = 1, 2, \cdots, M. \tag{12}$$

The minimum angular difference $\Delta\theta^{[m]}$ between $\theta_k^{[m]}$ and DoA estimate in $\boldsymbol{z}_k^W$ is obtained by:

$$\Delta\theta^{[m]} = \min_{n=1,2,\cdots,N}\{|\theta_k^{[m]} - \phi_{n,k}^W|\} \in [0, \pi], \tag{13}$$

where n represents the observation index. This allows us to find the observation line closest to the particle (see Fig. 4).

Assuming $\Delta\theta^{[m]} \sim \mathcal{N}(0, \sigma^2)$, we use the corresponding probability density function (PDF) to update particle weights:

$$f(\Delta\theta^{[m]}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-{\Delta\theta^{[m]}}^2}{2\sigma^2}\right). \tag{14}$$

Moreover, to prevent excessively low weights from potential outliers or noise, a threshold parameter $\lambda = f(2\sigma)$ is introduced, and (14) can be improved to:

$$L_1(\Delta\theta^{[m]}) = \begin{cases} f(\Delta\theta^{[m]}), & \text{if } \Delta\theta^{[m]} \le 2\sigma \\ \lambda, & \text{if } \Delta\theta^{[m]} > 2\sigma \end{cases} \tag{15}$$

To mitigate the risk of "ghost sources" arising from bearing-only DoA measurements [18], we introduce a distance penalty term that accounts for the microphone array's limited pickup range:

$$L_2(d^{[m]}) = \exp\left(-d^{[m]}/\beta\right), \tag{16}$$

where β is the decay factor, and $d^{[m]}$ is the Euclidean distance between the particle and the robot:

$$d^{[m]} = \sqrt{(x_k^{[m]} - x_{r,k})^2 + (y_k^{[m]} - y_{r,k})^2}\quad . \tag{17}$$

Therefore, the likelihood function $L(\Delta\theta^{[m]}, d^{[m]})$ incorporates both angular deviation and distance, ensuring that particles falling on the observation line are weighted appropriately based on their proximity to the robot:

$$L(\Delta\theta^{[m]}, d^{[m]}) = \begin{cases} L_1 \cdot L_2, & \text{if } \Delta\theta^{[m]} \le 2\sigma \\ L_1, & \text{if } \Delta\theta^{[m]} > 2\sigma \end{cases} \tag{18}$$

Particle weights are updated using the likelihood function:

$$w_k^{[m]} = L(\Delta\theta^{[m]}, d^{[m]}) \cdot w_{k-1}^{[m]}. \tag{19}$$

4) Adaptive Resampling: We employ an adaptive resampling strategy that activates only when particle weight disparity falls below a resampling threshold. The particle weight disparity is quantified by:

$$N_{eff} = \frac{1}{\sum_{m=1}^{M}(w_k^{[m]})^2} \in [1, M]. \tag{20}$$

The resampling threshold $T_{resample}$ is defined as:

$$T_{resample} = M/C, \tag{21}$$

where C is a constant and empirically set to 10. This strategy allows particles to accumulate weights from multiple observations, avoiding early resampling that might hinder mapping accuracy. The low variance sampling [22] technique is employed to retain particles with high weights.

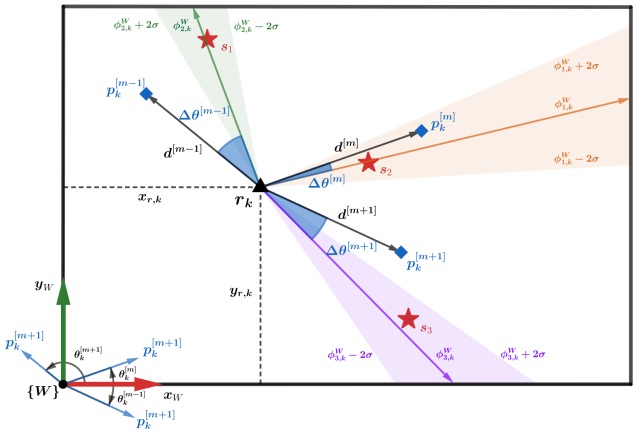


Fig. 4: Minimum angular difference schematic

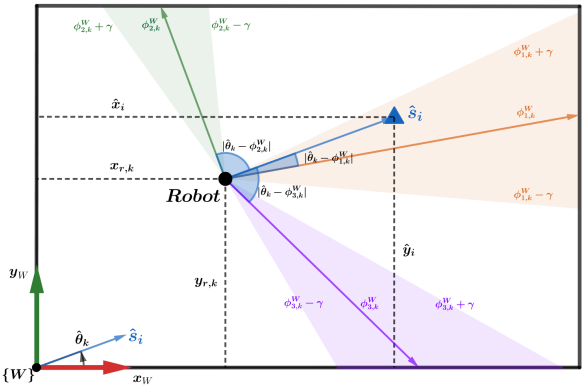


Fig. 5: Illustrative example of observation association

C. DBSCAN Clustering

After particle filtering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [23] is applied to identify clusters of closely converged particles in $\boldsymbol{\mathcal{P}}_K$, distinguishing them from noise. Within the resulting cluster set $\boldsymbol{\mathcal{C}}$, we concentrate on the cluster with the highest particle count, $\boldsymbol{\mathcal{C}}_{max}$, and calculate its centroid $\hat{\boldsymbol{s}}_i$ as the estimated sound source position for the current i-th filtering round:

$$\hat{\boldsymbol{s}}_i = \frac{1}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}} \boldsymbol{p}_k^{[m]} = (\hat{x}_i, \hat{y}_i), \quad \text{where } \boldsymbol{p}_k^{[m]} \in \boldsymbol{\mathcal{C}}_{max}, \quad (22)$$

which will be added to the source mapping result set $\hat{\boldsymbol{\mathcal{S}}}$.

The termination condition for the "Filtering-Clustering-Implicit Associating" loop is met when no valid clusters exist beyond the noise category in the current DoA Estimates Table $\boldsymbol{\Phi}$, indicating all sound sources have been located. Furthermore, by disregarding noise particle points and focusing on $\boldsymbol{\mathcal{C}}_{max}$, employing its centroid for estimation provides higher accuracy than the weighted sum of particle filtering results, as detailed in Section IV.

D. Implicit Observation Association

After estimating the sound source location $\hat{\boldsymbol{s}}_i$, we calculate the corresponding DoA for each robot pose $\boldsymbol{r}_k$ in the Pose Estimates Table $\boldsymbol{\Omega}$ using:

$$\hat{\theta}_k = \tan^{-1}\left(\frac{\hat{y}_i - y_{r,k}}{\hat{x}_i - x_{r,k}}\right), \quad k = 1, 2, \cdots, K. \quad (23)$$

We then identify the observed DoA value $\phi_{n,k}^W$ in the DoA Estimates Table $\boldsymbol{\Phi}$ that is closest to $\hat{\theta}_k$, with:

$$n^* = \underset{n \in \{1,2,\cdots,N\}}{\arg\min} |\hat{\theta}_k - \phi_{n,k}^W|. \quad (24)$$

If the minimum angular deviation $|\hat{\theta}_k - \phi_{n^*,k}^W|$ falls within the associated range $\gamma = 3\sigma$, the observation $\phi_{n,k}^W$ is associated with the estimated source $\hat{\boldsymbol{s}}_i$ and removed from $\boldsymbol{\Phi}$. The updated $\boldsymbol{\Phi}$ is then fed back into the particle filter module for subsequent estimation rounds. For instance, as illustrated in Fig. 5, among all observations $\phi_{1,k}^W$, $\phi_{2,k}^W$, $\phi_{3,k}^W$ at time step k, observation $\phi_{1,k}^W$ is the closest to the current source estimate $\hat{\boldsymbol{s}}_i$, i.e., $n^* = 1$. Meanwhile, $\hat{\boldsymbol{s}}_i$ also falls within the association interval $[\phi_{1,k}^W - \gamma, \phi_{1,k}^W + \gamma]$. Therefore, observation $\phi_{1,k}^W$ is associated with $\hat{\boldsymbol{s}}_i$, indicating a match between the observation and the estimated source.

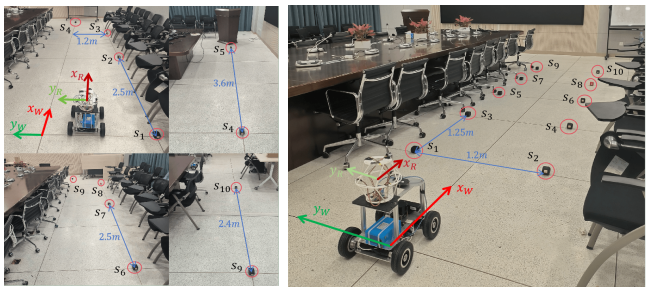
Algorithm 1: Proposed I-ASM algorithm

Input: Occupancy map, $\boldsymbol{\Omega}$, $\boldsymbol{\Phi}$
Output: Source mapping result $\hat{\boldsymbol{\mathcal{S}}}$

1 Initialize $\hat{\boldsymbol{\mathcal{S}}} = \emptyset$, M (9), $i = 0$;
2 **while** *TRUE* **do**
  // Filtering
3 Initialize $\boldsymbol{\mathcal{P}}$, $w^{[m]}$ (10);
4 **for** $stepCount = 0$ **to** $2K - 1$ **do**
5 $k = mod(stepCount, K) + 1$;
6 Update $\boldsymbol{\mathcal{P}}$ (11);
7 **for** $m = 1$ **to** M **do**
8 Calculate $\theta_k^{[m]}$ (12), $d^{[m]}$ (17), $\Delta\theta^{[m]}$ (13);
9 Calculate $L(\Delta\theta^{[m]}, d^{[m]})$ (18);
10 $w^{[m]} = L \cdot w^{[m]}$ (19);
11 **end**
12 **if** $N_{eff} < T_{resample}$ *(20) (21)* **then**
13 Resample $\boldsymbol{\mathcal{P}}$ [22];
14 **end**
15 $\lambda_{\text{max}} = \max(\text{eig}(\text{cov}(\boldsymbol{\mathcal{P}})))$;
16 **if** $\lambda_{max} < \lambda_{thresh}$ **then**
17 break;
18 **end**
19 **end**
  // Clustering
20 $\boldsymbol{\mathcal{C}}$ = DBSCAN($\boldsymbol{\mathcal{P}}$, NOT Noise);
21 **if** $\boldsymbol{\mathcal{C}} \neq \emptyset$ **then**
22 $i = i + 1$;
23 $\hat{\boldsymbol{s}}_i = mean(\boldsymbol{\mathcal{C}}_{max})$ (22);
24 Add $\hat{\boldsymbol{s}}_i$ to $\hat{\boldsymbol{\mathcal{S}}}$;
  // Implicit Associating
25 **for** $k = 1$ **to** K **do**
26 Calculate $\hat{\theta}_k$ (23), n^* (24);
27 **if** $|\hat{\theta}_k - \phi_{n^*,k}^W| < \gamma$ **then**
28 Remove $\boldsymbol{\Phi}(k, n^*)$ from $\boldsymbol{\Phi}$;
29 **end**
30 **end**
31 **end**
32 **else**
33 break;
34 **end**
35 **end**
36 **return** $\hat{\boldsymbol{\mathcal{S}}}$;

E. Algorithm

The I-ASM algorithm, as outlined in Algorithm 1, employs a dual traversal of trajectory and observation points (line 4) to enhance the convergence of particles, thereby mitigating clustering failures due to limited observations. The algorithm also features an early exit mechanism based on the maximum eigenvalue λ_{max} of the covariance matrix (line 15), which signifies the widest particle spread direction. If λ_{max} falls below the threshold λ_{thresh}, the algorithm assumes convergence and moves directly to the clustering step.



(a) Arrangement I (b) Arrangement II

Fig. 6: Sound source arrangements

IV. Experiments

A. Experimental Setup

As illustrated in Fig. 3, the mobile robot in our experiment is equipped with a 6-channel spherical microphone array (0.235m diameter, 48kHz sampling rate) and a 2D LiDAR. Tests are carried out in an 11m $\times$10m conference room with an occupancy map generated by *GMapping* [24] in ROS at $r = 0.05$m resolution. Ambient noise is 29dB, rising to 31dB with the LiDAR active. The environment is populated with $H = 10$ sound sources (pre-recorded voices of 5 men and 5 women) played on loudspeakers, with sound levels from 57dB to 70dB at 1m. Two source arrangements (I and II) are tested, as shown in Fig. 6. Arrangement I offers a sparse distribution with a 2.4m minimum spacing, while II provides a compact setup with a 1.25m maximum spacing.

The robot stops at each observation point at time step k to publish audio data (1.5-3 seconds) via a ROS topic. This data is subscribed by a MATLAB node for SSL using the *Multi-channel BSS Locate*[1] toolbox, which applies angular spectrum-based methods to convert signals into time-frequency representations and rank potential source locations by energy density. Two classic SSL methods, GCC-PHAT (Generalized Cross Correlation with Phase Transform) and MVDR (Minimum Variance Distortionless Response), are employed for DoA estimation. Experiment parameters include: $D = 9$, $\sigma = 5°$, $\beta = 50$, $\sigma_p = 0.01$m, $\lambda_{\text{tresh}} = 0.1\text{m}^2$, DBSCAN parameters $\epsilon = 0.1$m and *MinPts* $= 0.1M$, with $N = 3$ and 4 for Arrangement I and II, respectively.

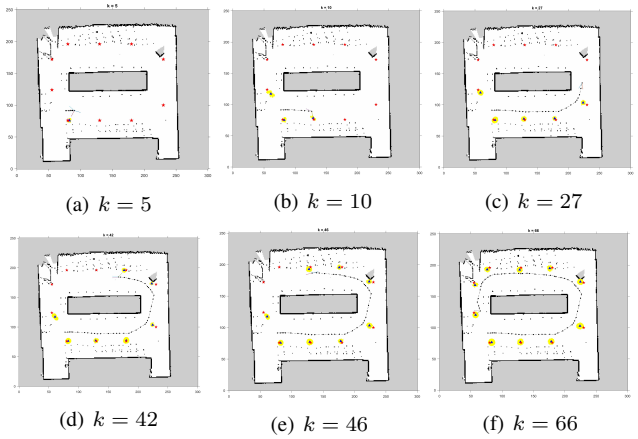
B. Performance Metrics

We use the Optimal Subpattern Assignment (OSPA) distance [25] to quantify the discrepancy between the true source set $\boldsymbol{\mathcal{S}}$ and the estimated set $\hat{\boldsymbol{\mathcal{S}}}$. For the case $I \leq H$, the OSPA distance is defined as:

$$d_p^c(\boldsymbol{\mathcal{S}}, \hat{\boldsymbol{\mathcal{S}}}) = \left[\frac{1}{H} \left(\min_{\pi \in \Pi_{\text{H}}} \sum_{i=1}^{I} \left(d^c(\boldsymbol{s}_i, \hat{\boldsymbol{s}}_{\pi(i)}) \right)^p + (H - I) c^p \right) \right]^{\frac{1}{p}},$$

where Π_{H} represents all permutations of length I, with elements taken from $\{1, 2, \cdots, H\}$; function $d^{(c)}(\boldsymbol{s}_i, \hat{\boldsymbol{s}}_{\pi(i)}) = \min(c, \|\boldsymbol{s}_i - \hat{\boldsymbol{s}}_{\pi(i)}\|^p)$ is the distance between true and estimated position with cutoff value c and order p. For the case $I > H$, $d_p^c(\boldsymbol{\mathcal{S}}, \hat{\boldsymbol{\mathcal{S}}}) = d_p^c(\hat{\boldsymbol{\mathcal{S}}}, \boldsymbol{\mathcal{S}})$. Following [25], we set $p = 1$ and $c = 1$m to balance localization and cardinality errors.

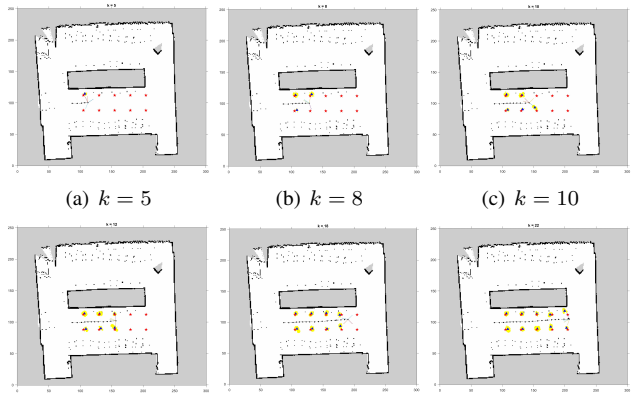
[1] https://gitlab.inria.fr/bass-db/mbss_locate



(a) $k = 5$ (b) $k = 10$ (c) $k = 27$

(d) $k = 42$ (e) $k = 46$ (f) $k = 66$

Fig. 7: Mapping result for Arrangement I (MVDR)



(a) $k = 5$ (b) $k = 8$ (c) $k = 10$

(d) $k = 12$ (e) $k = 18$ (f) $k = 22$

Fig. 8: Mapping result for Arrangement II (GCC-PHAT)

C. Experiment Results and Discussions

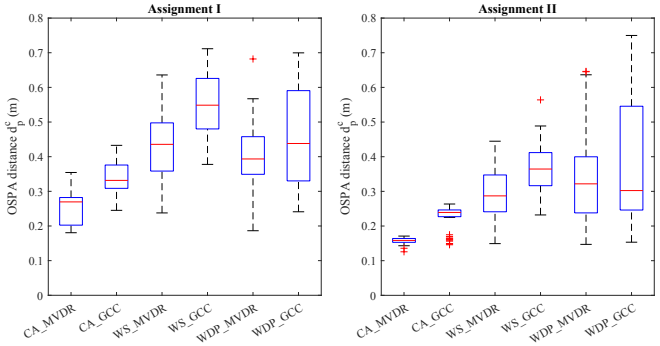
Figures 7 and 8 illustrate the mapping results during the robot's exploration, where Arrangement I has $K = 71$ and Arrangement II has $K = 40$ observation points visited by the robot. Considering the Monte Carlo nature of the algorithm (which includes random initialization and resampling of particles), we ran it 50 times on data collected from both arrangements to ensure a robust evaluation. To further underscore the benefits of integrating DBSCAN clustering and distance penalties in (16), we examined three algorithm versions: Cluster Average (CA), Weighted Sums (WS), and Without Distance Penalty (WDP). CA refers to the full I-ASM, while WS skips clustering and uses the weighted sum of particle filtering results directly, and WDP excludes distance penalties from the likelihood function.

Table I summarizes the OSPA distances' averages and medians under different conditions, alongside comparative boxplots depicted in Fig. 9. It is clear that the CA version outperformed WS and WDP across all tests, indicating the importance of employing DBSCAN clustering and distance penalties for mapping accuracy. Moreover, Arrangement II showed a better OSPA performance, probably attributed to the closer proximity of sources and the unobstructed paths to the robot, which allowed for more accurate DoA estimates.

TABLE I: Evaluation results for 50 Monte Carlo runs

	Source Arrangement	d_p^c average (m)		d_p^c median (m)	
		MVDR	GCC-PHAT	MVDR	GCC-PHAT
I	**CA (Full I-ASM)**	**0.249***	**0.338**	**0.270**	**0.332**
	WS	0.434	0.553	0.436	0.549
	WDP	0.403	0.456	0.393	0.438
II	**CA (Full I-ASM)**	**0.157**	**0.225**	**0.158**	**0.239**
	WS	0.289	0.363	0.287	0.364
	WDP	0.335	0.376	0.322	0.302

* Bold font indicates better performance across three algorithm versions.



(a) Assignment I (b) Assignment II

Fig. 9: Boxplots of OSPA distances under different experimental conditions

Conversely, the scattered source layout in Arrangement I increased the likelihood of false DoA detections, adversely affecting mapping accuracy. Furthermore, as for the SSL algorithm, MVDR outperformed GCC-PHAT across all scenarios, suggesting better adaptability for ASM.

Time efficiency is influenced by the number of observation points, audio recording duration, selected SSL methods, and the number of particles. It is crucial to make a trade-off between time efficiency and mapping accuracy based on the requirements of practical applications.

V. Conclusions

In this paper, we introduce I-ASM, a particle filter-based iterative framework for indoor acoustic scene mapping with multiple sound sources. Requiring no prior data association, I-ASM employs a sequential estimation strategy through an iterative "Filtering-Clustering-Implicit Associating" cycle, successfully achieving ASM of multi-sound source positions. Real-world experiments validate its effectiveness in two challenging scenarios of sound source layouts, each with 10 concurrent sound sources. Overall, I-ASM delivers precise mapping and exhibits resilience to missing and false DoA estimates. Building on these results, future work will focus on extending I-ASM's adaptability to dynamic environments and refining its accuracy under more complex scenarios.

References

[1] C. Evers and P. A. Naylor, Acoustic SLAM, *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, Vol. 26, No. 9, pp. 1484-1498, 2018.

[2] L. Wang and A. Cavallaro, Deep learning assisted sound source localization from a flying drone, *IEEE Sensors Journal*, Vol. 22, No. 21, pp. 20828-20838, 2022.

[3] P. P. Rao and A. R. Chowdhury, Learning to listen and move: An implementation of audio-aware mobile robot navigation in complex indoor environment, *Proc. of the IEEE ICRA*, pp. 3699-3705, 2022.

[4] D. Su, H. Kong, S. Sukkarieh, and S. Huang, Necessary and sufficient conditions for observability of SLAM-based TDOA sensor array calibration and source localization, *IEEE Trans. on Robotics*, Vol. 37, No. 5, pp. 1451-1468, 2021.

[5] J. Wang, Y. He, D. Su, K. Itoyama, K. Nakadai, J. Wu, S. Huang, Y. Li, and H. Kong, SLAM-based joint calibration of multiple asynchronous microphone arrays and sound source localization, *IEEE Trans. on Robotics*, DOI: 10.1109/TRO.2024.3410456, 2024.

[6] C. Zhang, J. Wang, and H. Kong, Asynchronous microphone array calibration using hybrid TDOA information, Accepted and to appear, *Proc. of the IEEE/RSJ IROS*, 2024.

[7] X. Li, H. Deng, J. Wang, L. Fu, and H. Kong, Information-aware joint calibration of microphone array and sound source localization, Accepted and to appear, *Proc. of the Int. Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2024.

[8] C. Rascon and I. Meza, Localization of sound sources in robotics: A review, *Rob. Auton. Syst.*, Vol. 96, pp. 184-210, 2017.

[9] I. An, G. An, T. Kim, and S. Yoon, Microphone pair training for robust sound source localization with diverse array configurations, *IEEE Robot. Autom. Lett.*, Vol. 9, No. 1, pp. 319-326, 2023.

[10] E. Martinson and A. Schultz, Auditory evidence grids, *Proc. of the IEEE/RSJ IROS*, pp. 1139-1144, 2006.

[11] N. Kallakuri, J. Even, Y. Morales, C. Ishi, and N. Hagita, Probabilistic approach for building auditory maps with a mobile microphone array, *Proc. of the IEEE ICRA*, pp. 2270-2275, 2013.

[12] J. Even, J. Furrer, Y. Morales, C. T. Ishi, and N. Hagita, Probabilistic 3-D mapping of sound-emitting structures based on acoustic ray casting, *IEEE Trans. on Robotics*, Vol. 33, No. 2, pp. 333-345, 2017.

[13] D. Su, K. Nakamura, K. Nakadai, and J. V. Miro, Robust sound source mapping using three-layered selective audio rays for mobile robots, *Proc. of the IEEE/RSJ IROS*, pp. 2771-2777, 2016.

[14] J. S. Hu, C. Y. Chan, C. K. Wang, M. T. Lee, and C. Y. Kuo, Simultaneous localization of a mobile robot and multiple sound sources using a microphone array, *Advanced Robotics*, Vol. 25, pp.135-152, 2011.

[15] C. Schymura and D. Kolossa, Potential-field-based active exploration for acoustic simultaneous localization and mapping, *Proc. of the IEEE ICASSP*, pp. 76-80, 2018.

[16] M. Wakabayashi, H. G. Okuno, and M. Kumon, Multiple sound source position estimation by drone audition based on data association between sound source localization and identification, *IEEE Robot. Autom. Lett.*, Vol. 5, No. 2, pp. 782-789, 2020.

[17] Y. Sasaki, R. Tanabe, and H. Takernura, Online spatial sound perception using microphone array on mobile robot, *Proc. of the IEEE/RSJ IROS*, pp. 2478-2484, 2018.

[18] X. Dang, Q. Cheng, and H. Zhu, Indoor multiple sound source localization via multi-dimensional assignment data association, *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, Vol. 27, No. 12, pp. 1944-1956, 2019.

[19] S. Michaud, S. Faucher, F. Grondin, J.-S. Lauzon, M. Labbé, D. Létourneau, F. Ferland, and F. Michaud, 3D localization of a sound source using mobile microphone arrays referenced by SLAM, *Proc. of the IEEE/RSJ IROS*, pp. 10402-10407, 2020.

[20] Z. Wang, W. Zou, H. Su, Y. Guo, and D. Li, Multiple sound source localization exploiting robot motion and approaching control, *IEEE Trans. on Instrumentation and Measurement*, Vol. 72, pp. 1-16, 2023.

[21] I. An, Y. Kwon, and S. Yoon, Diffraction-and reflection-aware multiple sound source localization, *IEEE Trans. on Robotics*, Vol. 38, No. 3, pp. 1925-1944, 2021.

[22] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, MIT Press, 2005.

[23] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, DBSCAN revisited, revisited: why and how you should (still) use DBSCAN, *ACM Trans. on Database Systems*, Vol. 42, No. 3, pp. 1-21, 2017.

[24] G. Grisetti, C. Stachniss, and W. Burgard, Improved techniques for grid mapping with Rao-Blackwellized particle filters, *IEEE Trans. on Robotics*, Vol. 23, No. 1, pp. 34-46, 2007.

[25] D. Schuhmacher, B. -T. Vo, and B. -N. Vo, A consistent metric for performance evaluation of multi-object filters, *IEEE Trans. on Signal Processing*, Vol. 56, No. 8, pp. 3447-3457, 2008.