Audio-Radar SMC-PHD Filtering for Indoor Multiple Speaker Tracking

Yi Zhou
Institute of Deep Perception Technology
JITRI
Wuxi, China
zhouyi1023@tju.edu.cn

Miguel López-Benítez
Department of EEE
[1]University of Liverpool
Liverpool, UK
[2]Nebrija University, Spain
mlopben@liverpool.ac.uk

Limin Yu*
School of Advanced Technology
Xi'an Jiaotong-Liverpool University
Suzhou, China
limin.yu@xjtlu.edu.cn

Yutao Yue*
Thrust of Artificial Intelligence/ Intelligent Transportation
HKUST (GZ)
Guangzhou, China
ytyue@ustc.edu

***Abstract*—High-resolution millimetre-wave (mmWave) radar sensors have become increasingly popular in consumer markets. This study addresses the challenge of tracking multiple active speakers in indoor environments using high-resolution radar and microphone array. Through our experiments, we have observed that the Sequential Monte Carlo Probability Hypothesis Density (SMC-PHD) filter, when given point cloud data from a high-resolution radar as input, can provide promising tracking performance. In this work, we add another modality, i.e., audio, to the radar SMC-PHD filtering framework for the active speaker tracking task. Specifically, we use the audio Direction of Arrival (DoA) to guide the particle birth and relocation process in the SMC-PHD filtering framework. Furthermore, we propose a likelihood function that jointly considers the spatial and angular estimation from radar and audio. Experimental results on the RAV4D dataset demonstrate that our audio-radar SMC-PHD filtering approach produces reliable trajectories, especially in the challenging cases such as varying numbers of speakers.**

Keywords—PHD filtering, audio-radar fusion, object tracking

I. INTRODUCTION

Multiple speaker tracking is widely used in applications such as video conferencing and human-computer interfaces. By accurately tracking the speaker's location, we can implement advanced audio enhancement techniques like beamforming and source separation, significantly improving the overall audio quality. While audio-visual tracking is effective in many scenarios, it has inherent limitations, including the absence of reliable distance information, privacy concerns, and reduced functionality in challenging lighting conditions. In response to these challenges, our work focuses on the development of an indoor multiple speaker tracking system that leverages millimeter-wave radar and a microphone array. Millimeter-wave radars have gained widespread use [1] as perception sensors due to their mature manufacturing, compact size, and affordability. The growing demands for advanced autonomous driving systems has led to a reduction in the cost of high-resolution radar sensors, making them practical for use in the consumer market, particularly for people monitoring and tracking applications. Our proposed system is specifically designed for deployment in meeting scenarios that exhibit low illumination and require stringent privacy protection measures.

Radar detections often suffer from low signal-to-noise ratios (SNR) and lack of semantics, leading to missed detections and false alarms in cluttered scenarios. To model these unreliable detections, Bayesian tracking frameworks with random finite sets (RFS) [2] are employed. In this framework, detections are represented as a random finite set described by a multiple object probability density function (pdf). The multi-object pdf is a non-negative function defined over the set cardinality and the spatial positions of these detections. Consequently, it captures both the distribution of the set cardinality, which represents the number of detections, and the distribution of the individual detection given the cardinality. To alleviate the computational burden associated with integrating over sets in the Chapman-Kolmogorov equation, the first-order approximation of the multi-object pdf, known as the Probability Hypothesis Density (PHD) [2], is utilized. The practical implementation of the PHD filter includes methods like the GM-PHD filter [3], which models density functions as Gaussian distributions, and the SMC-PHD filter [5], which employs particles to represent the distributions. To address the challenge of unstable cardinality estimation, the CPHD filter [4] has been proposed, which propagates the cardinality distribution in addition to the PHD. Recent advancements have been made in trajectory PHD/CPHD filtering [5], where sets of trajectories are modeled using a Poisson multi-trajectory density.

Audio signal is used for audio-visual tracking of speakers. Kılıc *et al.* [6] propose an audio-visual tracking framework which use the DoA of the audio sources to guide the particle propagation in the prediction step and to weight the particle in the measurement step of the particle filter. In their following work[7], they propose an SMC-PHD filtering for audio-visual tracking, where the DoA is used to propagate the born particles and re-allocate the surviving and spawned particles. Our proposed fusion framework of radar and audio is mainly inspired by this work.

In our study, we introduce an innovative audio-radar tracking framework that leverages DoA information estimated from a microphone array to enhance radar SMC-PHD filtering. Both radar and audio data suffer from low spatial resolution and low SNR in complex indoor environment. Our experimental results demonstrate the effectiveness of our proposed method, particularly in achieving smoother trajectories when tracking multiple people in complex indoor environments. Moreover, the proposed framework can reliably estimate both the number

This work received financial support from Jiangsu Industrial Technology Research Institute (JITRI) and Wuxi National Hi-Tech District (WND).

of speakers and their positions, even in the presence of frequent occlusions.

The remainder of this work is organized as follows: Section II provides a comprehensive introduction to our proposed framework for audio-radar SMC-PHD filtering. Section III is dedicated to a thorough analysis of the experimental results obtained from the dataset. Finally, Section IV summarises the main findings and discusses directions for future research.

II. AUDIO-RADAR SMC-PHD FILTERING

In this section, we propose an audio radar SMC-PHD filtering framework for multi-speaker tracking. We define the state of particle as x_k^j, a four-dimensional vector of position and velocity in XY plane. The subscript k indicates the frame number, and the superscript j indicates the particle index. The particles can be categorized into three sets: the survival object set S_k, the spawn object set B_k, and the born object set Γ_k. The born particles are mainly used to detect the new speaker who entering the field of view or reappear from occlusion. At each update step, a proportion of λ_S particles from the total particles are duplicated as spawn particles, while a proportion of p_S particles from the last frame are randomly selected as surviving particles. The surviving particles are then propagated using the linear motion model. The measurements denoted by z_k^i consists of two parts, one is for the 2D location measured by radar denoted as p, and the other is for the angles θ measured by radar and microphone array respectively.

As shown in Fig 2.1, the particle propagation process consists of three steps: particle birth, particle relocation and particle resampling.

The particle birth process is an optional process only when new speakers are detected entering the Field of View (FoV) or reappearing after occlusion. The audio DoA can provide prior knowledge about the number of speakers. Therefore, we can infer the appearance of new speakers if the number of current DoA lines is greater than the estimated number of speakers in the last frame. The number of new speakers, denoted as λ_k, is determined by the difference between them. When new speakers are detected, we need to uniformly sample some born particles around the DoA line, as shown in Fig 2.1 (a). Assuming that we generate ρ particles for each new detection, the initial weights for these particles are given by

$$\omega_{k|k-1} = \frac{1}{\rho\lambda_k} \tag{2-1}$$

In the particle relocation process, we can adjust the particle positions towards the corresponding DoA lines to improve particle efficiency. For each particle, we calculate the perpendicular distance to each DoA line. Ideally, particles should be assigned to the nearest DoA line. However, the number of DoA lines may not match the number of speakers, especially in the case of missed detections or false alarms. To address this, we define a distance threshold to prevent incorrect assignment of particles to DoA lines. If the nearest distance d_k^j falls within the threshold, the particle will be relocated to approach the nearest DoA line as

$$x_k^j = x_k^j - k_d d_k^j \xi_k \tag{2-2}$$

where ξ_k represents the vector of the audio DoA, and k_d is a scaling factor. Fig. 1 (b) illustrates the process of particle relocation. The red and blue crosses represent the original particles and the relocated particles, respectively. It is clear that all the particles move towards the DoA line after relocation.

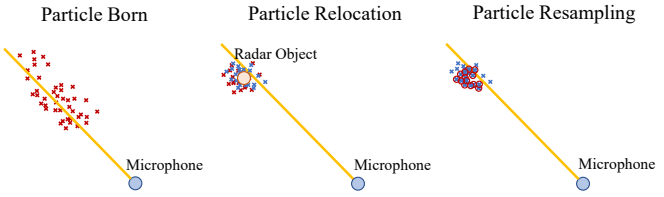


Fig. 1. Particle propagation process: (a) particle birth (b) particle relocation (c) particle resampling

After updating the locations of the particles, we need to determine their weights in the Bayesian recursion framework. In the PHD filtering, the survival set, and the spawn set are modelled as multi-Bernoulli RFS, and the born set is modelled as a Poisson RFS. The motion model can be modelled as a multi-Bernoulli process. Therefore, we can derive the corresponding PHD of the multi-object posterior and propagate the density instead of the full RFS posterior to avoid numerical intractability in the set integral. Accordingly, the prediction weight for a particle is calculated as

$$\omega_{k|k-1}^j = (p_S + \lambda_S)\omega_{k-1|k-1}^j \tag{2-3}$$

In the update step, suppose there are L_k particles and L_z measurements denoted by z_k^i at time step k.The weight is updated according to

$$w_{k|k}^j = \left(1 - p_D + \sum_{i=1}^{L_z} \frac{p_D h_k\left(z_k^i|x_k^j\right)}{\kappa + \sum_{j=1}^{L_{k-1}+J_k} p_D h_k\left(z_k^i|x_k^j\right) w_{k|k-1}^j}\right) w_{k|k-1}^j \tag{2-4}$$

where p_D, h_k and κ are the detection probability, the likelihood, and the density of clutter.

For the likelihood h_k, as shown in Fig 2.2 (a), we consider jointly the spatial localization error estimated by the radar and the fused angular error estimated by both the radar and the microphone array. The equation is given by

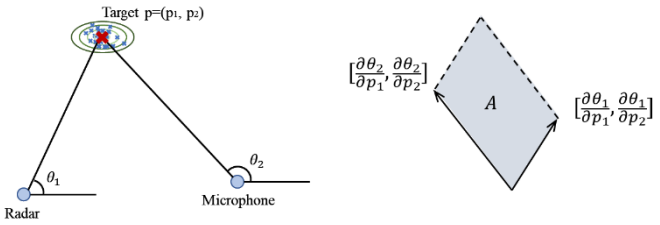
$$h_k = L_p \cdot A_\theta \tag{2-5}$$

where L_p represents the spatial likelihood function calculated based on the fitted two-dimensional Gaussian distributions that are fitted from the radar point detections.

The term A_θ, which serves as an angular penalty factor [8], is defined as:

$$A_\theta = \left| \frac{\partial\theta_1}{\partial p_1}\frac{\partial\theta_2}{\partial p_2} - \frac{\partial\theta_2}{\partial p_1}\frac{\partial\theta_1}{\partial p_2} \right| \quad (2\text{-}6)$$

As shown in Fig. 2 (b), this term corresponds to the area of the parallelogram spanned by two angle derivative vectors from radar detections and the microphone array.



(a) (b)

Fig. 2. Angular penalized likelihood: (a) spatial likelihood (b) angular penalty factor

After the update step, the cardinality, which represents the number of speakers, is estimated by

$$N_k = \sum_{j=1}^{L_z} w_{k|k}^{j} \quad (2\text{-}7)$$

Because the PHD filter tracks the distribution of multiple objects rather than individual objects themselves, we need to apply a k-means clustering to the particle sets and use the centers as the predicted state of the tracked speakers.

One of the key challenges in SMC methods is sample impoverishment [9]. As the number of iterations increases, the variance of the particles grows, and only a small subset of the total particles effectively represent the distribution while the others are distributed in low density areas. To alleviate this problem, a resampling step is essential after each iteration. Assuming that there are N_k estimated speakers, then we sample and select $L_k = \rho N_k$ particles from the original particle set. We adopt the multinomial resampling strategy [9], where the particles are selected according to the multinomial distribution as

$$x_k^* = x\left(F^{-1}(u_k)\right) \text{ with } i \text{ s.t. } u_k \in \left[\sum_{s=1}^{i-1} w_s, \sum_{s=1}^{i} w_s\right) \quad (2\text{-}8)$$

where u_k is one of L_k the pre-generated uniform random numbers and F^{-1} is the inverse of the cumulative distribution function associated with the normalized weights w_s of the particle.

Fig 2.1 (c) illustrates the particle resampling process. The blue crosses represent the original particles, while the red circles represent the resampled particles. During this process, particles with low weights are discarded, and those with higher weights are chosen. This helps to improve the efficiency of the particles.

The entire audio-radar SMC-PHD filtering framework is summarized in the following pseudo-code.

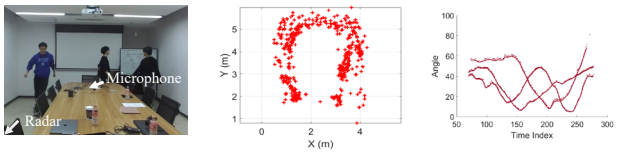
ALGORITHM 1: AUDIO-RADAR SMC-PHD FILTERING

```
for frame k = 1, 2, … do
    Select L_{k-1} particles as survival and spawn particles
    Propagate the survival particles using the motion model
    if DoA exists then
        for Each survival particle do
            Calculate the distance to each DoA line
            Relocate particles towards the nearest DoA line
        end for
    end if
    Estimate the number of speakers N_s
    if N_s > N_{k-1} then   # New speaker is detected
        Generate ρλ_k new-born particles around the DoA lines
        Update weights ω_k^i for born particles
    end if
    Update weights ω_k^i for survived and spawn particles
    for Each particle do
        Calculate the distance to each detection
        Select the nearest distance
        Estimate the likelihood
    end for
    Update all particle weights ω_k^i
    Estimate the cardinality of the set
    K-means clustering to estimate the speaker position
    Resample L_k = ρN_k particles
end for
```

III. EXPERIMENTAL RESULTS

A. Datasets and Performance Metrics

The RAV4D dataset [10] is a recently proposed multimodal dataset for indoor multi-person tracking that provides data from radar, microphone arrays and stereo cameras. It aims to improve tracking performance by exploiting the synergistic and complementary capabilities of these different sensing modalities. As shown in Fig. 3, we used the processed 2D radar point cloud and audio DoA as input to our tracking algorithm. The dataset provides different sequences with varying numbers of people and moving trajectories. Specifically, we select the one containing three people with frequent trajectory crossing as a visual example to test our tracking framework.



(a) Dataset Scenario (b) Radar Point Cloud (c) Audio DoA

Fig.3. The RAV4D dataset

The RAV4D dataset is originally collected for the evaluation of people tracking, so the trajectory annotation is only done for the

segments where all people speak aloud. Therefore, the number of audio sources is mostly constant, and the audio modality mainly acts as a complementary sensor to improve the tracking accuracy. To specifically test the speaker tracking performance, we further define a narrowed FoV as the upper half of the room where $y \in [385,650]$, so that the number of speakers can vary.

For the evaluation of tracking performance, we consider both multiple object tracking accuracy (MOTA) and multiple object tracking precision (MOTP) [11], as well as the generalized optimal sub-pattern assignment metric (GOSPA) [12], which specifically considers the cardinality error and can better evaluate the speaker tracking problem.

B. Results and Discussions

First, we evaluate the tracking performance using the RAV4D dataset. We compare the radar EKF, the radar SMC-PHD filtering and the audio radar SMC-PHD filtering. From Table I, we can find that the SMC-PHD filtering framework achieves an improvement in MOTA and a marginal performance gain with respect to MOTP than the EKF framework. The increase in MOTA indicates smaller tracking errors in terms of false positives, missed targets or identity switches. The small difference in MOTP suggests that while tracking accuracy has improved, localization accuracy has not changed significantly for the single sensor case.

By incorporating the audio source into the tracking framework, we can see that the audio radar SMC-PHD filtering provides the significant improvement in both MOTA and MOTP metrics. This indicates that the fusion of audio and radar data in the SMC-PHD filtering framework improves both the accuracy and precision of multi-object tracking.

TABLE I PERFORMANCE COMPARISION

Methods	MOTA (%)	MOTP (%)
Radar EKF	77.3	82.6
Radar SMC-PHD	79.5	82.7
Audio-Radar SMC-PHD	87.3	86.5

As the evaluation metrics take into account many aspects of tracking, the performance increase cannot be understood directly. Therefore, we visually assess the smoothness of the predicted trajectory, as shown in Figure 4. We can see that by incorporating the angle estimation, the trajectory tends to be smoother, thus reducing the localization error.

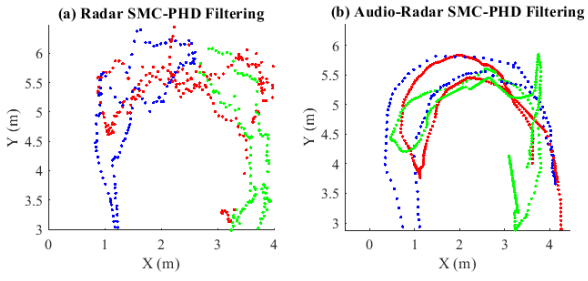


Fig. 4. Trajectory comparison

In the next test, we examine the tracking performance with a narrowed FoV. By defining a narrow FoV, the speakers may frequently enter or leave the FoV, causing the trajectory to be fragmented and the number of speakers to vary over time. Therefore, tracking could be difficult in this case. In Fig. 5, we show the frame-by-frame GOSPA error by radar SMC-PHD filtering and our proposed audio-radar SMC-PHD filtering. From the figure, we can see that incorporating the audio modality into the tracking framework significantly improves the tracking performance by enabling a lower cardinality estimation error and the smaller localization error. From the lower valley of the error curve, we can see that the localization error is smaller due to the high resolution of the radar sensor and the cardinality error is the largest contributor to the GOSPA error.

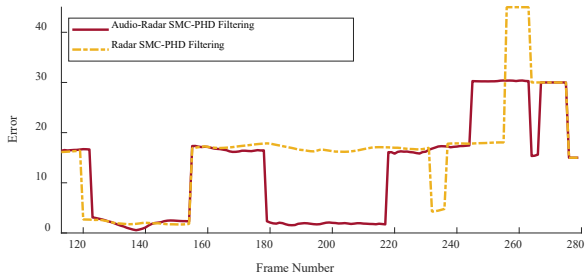


Fig. 5. Performance comparison with respect to GOSPA

We then specifically show the cardinality error of the two methods in Fig. 6. We can see that the audio radar method can give a more consistent estimate of cardinality. When using radar alone, it is possible to lose some pedestrians due to weak reflections from the human body, especially when the speaker is far away. In some extreme cases, such as perpendicular motion, the radar sensor may only return a few points due to possible occlusion and limited ability to detect low angular velocity motion. Therefore, relying on tracking alone cannot fully solve the detection failure problem. In contrast, by introducing the audio DoA into the tracking framework, the cardinality estimation can be significantly improved. In addition, we observe some delay when people enter or leave the FoV, which may be caused by the inaccurate angular estimation.

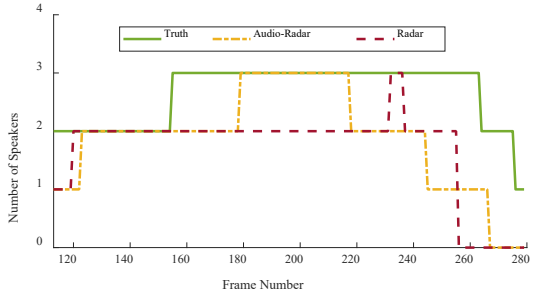


Fig. 6. Number of speakers estimation

IV. Conclusion

In this paper, we propose a novel audio radar tracking framework. The framework is based on SMC-PHD filtering and uses audio DoA to guide the particle birth location and relocation. Furthermore, we propose a likelihood that jointly considers the spatial distribution of the radar point cloud and the DoA estimated by the radar and microphone array. We test our audio-radar tracking algorithm on the RAV4D dataset and achieve better performance than radar-only tracking in terms of MOTA, MOTP and GOSPA metrics. Compared to the Radar-SMC-PHD filtering, the proposed Audio-Radar-SMC-PHD filtering can reliably estimate both the number of speakers and positions, especially in challenging scenarios such as the case of multiple persons and frequent occlusions. In future work, we aim to further address potential modality absences and develop a robust tracking framework.

References

[1] Y. Zhou, L. Liu, H. Zhao, M. López-Benítez, L. Yu, and Y. Yue, "Towards Deep Radar Perception for Autonomous Driving: Datasets, Methods, and Challenges," *Sensors*, vol. 22, no. 11, p. 4208, May 2022.

[2] R. P. S. Mahler, Statistical Multisource-multitarget Information Fusion. Artech House Publishers, 2007.

[3] D. E. Clark, K. Panta and B. -n. Vo, "The GM-PHD Filter Multiple Target Tracker," *2006 9th International Conference on Information Fusion*, Florence, Italy, 2006, pp. 1-8.

[4] R. P. S. Mahler, B. -T. Vo and B. -N. Vo, "CPHD Filtering With Unknown Clutter Rate and Detection Profile," in *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3497-3513, Aug. 2011.

[5] Á. F. García-Fernández and L. Svensson, "Trajectory PHD and CPHD Filters," in *IEEE Transactions on Signal Processing*, vol. 67, no. 22, pp. 5702-5714, 15 Nov.15, 2019.

[6] V. Kılıç, M. Barnard, W. Wang and J. Kittler, "Audio Assisted Robust Visual Tracking With Adaptive Particle Filtering," in *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 186-200, Feb. 2015.

[7] V. Kılıç, M. Barnard, W. Wang, A. Hilton and J. Kittler, "Mean-Shift and Sparse Sampling-Based SMC-PHD Filtering for Audio Informed Visual Speaker Tracking," in *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2417-2431, Dec. 2016.

[8] Z. Wang, J. -A. Luo and X. -P. Zhang, "A Novel Location-Penalized Maximum Likelihood Estimator for Bearing-Only Target Localization," in *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6166-6181, Dec. 2012.

[9] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later." Handbook of nonlinear filtering, pp.656-704, 2009.

[10] Y. Zhou, N. Song, J. Ma, K.L.Man, L. López-Benítez, L. Yu, and Y. Yue,, "RAV4D: A Radar-Audio-Visual Dataset for Indoor Multi-Person Tracking," *2024 IEEE Radar Conference (RadarConf24)*, Denver, CO, USA, 2024, pp. 1-6

[11] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.

[12] A. S. Rahmathullah, Á. F. García-Fernández and L. Svensson, "Generalized optimal sub-pattern assignment metric," *2017 20th International Conference on Information Fusion (Fusion)*, Xi'an, China, 2017, pp. 1-8