INTENSITY PARTICLE FLOW SMC-PHD FILTER FOR AUDIO SPEAKER TRACKING

Yang Liu, Wenwu Wang

University of Surrey
Centre for Vision, Speech and Signal Processing
Guildford, GU2 7XH, U.K
[yangliu, w.wang]@surrey.ac.uk

Volkan Kilic

Izmir Katip Celebi University
Department of Electrical and Electronics Engineering
35620 Cigli-Izmir, Turkey
volkan.kilic@ikc.edu.tr

ABSTRACT

Non-zero diffusion particle flow Sequential Monte Carlo probability hypothesis density (NPF-SMC-PHD) filtering has been recently introduced for multi-speaker tracking. However, the NPF does not consider the missing detection which plays a key role in estimation of the number of speakers with their states. To address this limitation, we propose to use intensity particle flow (IPF) in NPF-SMC-PHD filter. The proposed method, IPF-SMC-PHD, considers the clutter intensity and detection probability while no data association algorithms are used for the calculation of particle flow. Experiments on the LOCATA (acoustic source Localization and Tracking) dataset with the sequences of task 4 show that our proposed IPF-SMC-PHD filter improves the tracking performance in terms of estimation accuracy as compared to its baseline counterparts.

Index Terms— LOCATA, SMC-PHD, particle flow.

1. INTRODUCTION

The problem of acoustic source localization and tracking in an enclosed space has attracted an increased amount of attention in the last decade due to its potential applications such as personal assistants [1], advanced computer interfaces [2], hearing aids [3] and speech recognition [4]. To address this problem, several methods, such as direction of arrival (DOA) [5], generalized cross-correlation (GCC) phase transform (PHAT) [6], steered response power (SRP) PHAT, beam steering [7], and time delay of arrival (TDOA) estimates [8], have been proposed. The trajectories of the speakers can be extracted using estimated positions by aforementioned methods. However, these trajectories may involve the random errors, false returns from background clutters, and detection loss [9]. To overcome these issues, filters are used to smooth the estimated trajectories. Representative filters include Kalman [10] and particle [11] filters employed in tracking of a single moving sound source.

To track multiple moving sources, the unknown and variable number of sources need to be handled for reliable tracking. Therefore, PHD filter [12] and its extension such as cardinalized PHD filter [13] are elegant solutions for multiple source tracking. The Gaussian mixture (GM) [14] and sequential Monte Carlo (SMC) [15] are the implementations to obtain practical solutions of the PHD filter. [15]. However, it suffers from the weight degeneracy problem [16]. To address this problem, particle flow is proposed [16], which migrates particles from the prior distribution to the posterior distribution based on a homotopy function defined for particle flow. In the literature, particle flow is categorized into five main classes: incompressible particle flow [16], zero diffusion particle flow (ZPF) [17], Coulombs law particle flow [18], zero-curvature particle flow [19] and non-zero diffusion particle flow (NPF) [20]. Recently, ZPF-SMC-PHD and NPF-SMC-PHD filters are used to track multi-speakers based on the audio-visual information [4, 21].

For acoustic source tracking, the filters are mostly conducted with simulated data [22]. For the objective benchmarking of state-of-the-art algorithms on real-world data, the LOCATA dataset under the IEEE AASP Challenge is released [23]. The dataset comprises six tasks ranging from the tracking of a single static sound source to the tracking of multiple moving speakers. It contains real-world audio recordings obtained by DICIT array, Eigenmike array, Robot head and Hearing aids in an enclosed acoustic environment. The sound sources are represented by moving human talkers or static loudspeakers.

In this paper, we propose a new algorithm for multi-speaker tracking, namely IPF-SMC-PHD filter for the task 4 of the LOCATA dataset. This task covers the multiple moving talkers using a static microphone array. The proposed method considers the clutter intensity and detection probability while no data association algorithms are used for the calculation of particle flow. The DOA lines are employed as the measurements of the IPF-SMC-PHD filter for multi-speaker tracking under challenging conditions such as occlusion. The speaker identity is estimated using the target position under the assumption that it is not changed abruptly in subsequent frames. Our methods are tested on all sub-arrays of task 4.

The reminder of this paper is organized as follows: the next section introduces the NPF-SMC-PHD filter. Section III describes our proposed IPF-SMC-PHD filtering algorithm. In Section IV, experiments on the LOCATA dataset are presented to show the performance of the proposed IPF-SMC-PHD algorithm as compared with the baseline algorithms.

2. PROBLEM STATEMENT AND BACKGROUND

This section describes our problem formulation and the NPF-SMC-PHD filter. For the LOCATA challenge, we assume that the target dynamics and measurements are described as:

$$\tilde{\boldsymbol{m}}_k = \boldsymbol{f}_{\tilde{\boldsymbol{m}}}\left(\tilde{\boldsymbol{m}}_{k-1}, \boldsymbol{\tau}_k\right), \tag{1}$$

$$\boldsymbol{z}_k = \boldsymbol{f}_{\boldsymbol{z}}\left(\tilde{\boldsymbol{m}}_k, \varsigma_k\right) \tag{2}$$

where $\tilde{\boldsymbol{m}}_k \in \mathbb{R}^4$ is the target state vector in time k, defined as $\tilde{\boldsymbol{m}}_k = [x_k, y_k, \dot{x}_k, \dot{y}_k]^T$, which consists of the source azimuth x, elevation y and the angular velocity $(\dot{x}_k, \dot{y}_k)$. $\tilde{}$ is used to distinguish the target state from the particle state used later. Let $\boldsymbol{Z}_k$ denote the set of DOA calculated by Multiple Signal Classification in time k. $\boldsymbol{Z}_k = \{\boldsymbol{z}_k^1, \boldsymbol{z}_k^2, ..., \boldsymbol{z}_k^{R_k}\}$ where R_k is the number of measurements at time k. The measurement $\boldsymbol{z}_k^r$ is a noisy version of the position (x_k, y_k), where r is the index of the measurement. $\boldsymbol{\tau}_k$ and ς_k are system excitation and measurement noise terms, respectively. $\boldsymbol{f}_{\tilde{\boldsymbol{m}}}$ is a transition model and $\boldsymbol{f}_{\boldsymbol{z}}$ is a measurement model.

In the NPF-SMC-PHD filter [21], target PHD is approximated by N_{k-1} survival particles $\{\boldsymbol{m}_{k-1}^i\}_{i=1}^{N_{k-1}}$ and their weights $\{\omega_{k-1}^i\}_{i=1}^{N_{k-1}}$ at time $k-1$. In the prediction step, the particle set is obtained by the proposal distribution q_k,

$$\boldsymbol{m}_{k|k-1}^i \sim q_k(\cdot|\boldsymbol{m}_{k-1}^i, \boldsymbol{Z}_k) \tag{3}$$

The proposal weights are

$$\omega_{k|k-1}^i = q_s \omega_{k-1}^i \tag{4}$$

where q_s is the surviving possibility. N_B born particles are sampled by the importance function p_k,

$$\boldsymbol{m}_{k|k-1}^i \sim p_k(\cdot|\boldsymbol{Z}_k) \tag{5}$$

The born particle weights are

$$\omega_{k|k-1}^i = \frac{\gamma_k(\boldsymbol{m}_{k|k-1}^i)}{N_B p_k(\boldsymbol{m}_{k|k-1}^i|\boldsymbol{Z}_k)} \quad i = N_{k-1}+1, ..., N_{k-1}+N_B \tag{6}$$

where $\gamma_k(\cdot)$ is the born possibility. N_{k-1} is the number of surviving particles at time $k-1$.

After predicting particles, a particle flow mitigates particle states via the Ito stochastic differential equation [24]:

$$\triangle\boldsymbol{m}_{k|k-1}^i = \boldsymbol{f}_k^i(\boldsymbol{m}_{k|k-1}^i, \lambda)\triangle\lambda + \upsilon_k^i \boldsymbol{w}_k^i \tag{7}$$

where $\boldsymbol{f}_k^i \in \mathbb{R}^4$ is the particle flow vector which moves the particle $\boldsymbol{m}_{k|k-1}^i$ with the distance $\triangle\boldsymbol{m}_{k|k-1}^i$ at λ. $\boldsymbol{w}_k^i \in \mathbb{R}^4$ is the Wiener process with the diffusion coefficient υ_k^i, λ, called the synthetic time, takes values from $[0, \triangle\lambda, 2\triangle\lambda, \cdots, N_\lambda\triangle\lambda]$ and $N_\lambda\triangle\lambda = 1$. In NPF [20], $\boldsymbol{f}_k^i \in \mathbb{R}^4$ is given by,

$$\boldsymbol{f}_k^i = -[-(\boldsymbol{P}_{k-1}^i)^{-1} + \lambda\boldsymbol{\nabla}^2 \ln h_k^i]^{-1}(\boldsymbol{\nabla} \ln h_k^i) \tag{8}$$

where $\boldsymbol{P}_{k-1}^i$ is the covariance matrix of $\boldsymbol{m}_{k-1}^i$. $\boldsymbol{\nabla}$ is the spatial vector differentiation operator $\frac{\partial}{\partial \boldsymbol{m}_{k-1}^i}$. The likelihood h_k^i is given by

$$h_k^i = \mathcal{N}(\boldsymbol{m}_{k|k-1}^i | \underset{\boldsymbol{z}_k^r}{\operatorname{argmin}} \left\| \boldsymbol{z}_k^r - \boldsymbol{m}_k^i \right\|, \boldsymbol{R}) \tag{9}$$

where $\boldsymbol{R}$ is the covariance matrix of the measurement noise. $\|\cdot\|$ is the l_2 norm. Then each particle state is updated as

$$\boldsymbol{m}_{k|k-1}^i \Leftarrow \boldsymbol{m}_{k|k-1}^i + \triangle\boldsymbol{m}_{k|k-1}^i \tag{10}$$

The weights are calculated as

$$\omega_k^i = \left[1 - p_{D,k} + \sum_{r=1}^{R_k} \frac{p_{D,k} h_k^i}{\kappa_k + \sum_{i=1}^{N_k} p_{D,k} h_k^i \omega_{k|k-1}^i}\right] \omega_{k|k-1}^i \tag{11}$$

where $p_{D,k}^i$ and κ_k are the abbreviations of $p_{D,k}^i(\boldsymbol{m}_k^i|\boldsymbol{m}_{k-1}^1, ..., \boldsymbol{m}_{k-1}^{N_{k-1}})$ and $\kappa_k(\boldsymbol{Z}_k)$, respectively. $p_{D,k}^i$ is the detection probability. κ_k is the intensity function of clutter at time k. The number of targets is estimated as the sum of the weights. The states and weights of the targets $\{\tilde{\boldsymbol{m}}_k^j, \tilde{\omega}_k^j\}_{j=1}^{\tilde{N}_k}$ can be calculated using e.g. K-means clustering method [25] or multi-expected a posterior (MEAP) [26].

Algorithm 1 NPF-SMC-PHD Filter

```
Input: ${\boldsymbol{m}_{k-1}^i, \omega_{k-1}^i\}_{i=1}^{N_{k-1}}$, $\{\boldsymbol{P}_{k-1}^i\}_{i=1}^{N_{k-1}}$ and $\boldsymbol{Z}_k$.
Output: $\{\tilde{\boldsymbol{m}}_k^j, \tilde{\omega}_k^j\}_{j=1}^{\tilde{N}_k}$, $\{\boldsymbol{P}_k^i\}_{i=1}^{N_k}$ and $\{\boldsymbol{m}_k^i, \omega_k^i\}_{i=1}^{N_k}$.
    Initialize: $k$, $q_k$, $p_s$, $\phi_k$, $p_k$, $\kappa_k$, $P_{D,k}$, $\triangle\lambda$, $N_\lambda$, $\upsilon_k^i$, $\boldsymbol{w}_k^i$ and $N_B$.
 1: Run:
 2: Propagate the particle states $\{\boldsymbol{m}_{k|k-1}^i\}_{i=1}^{N_{k-1}}$ by Eq. (3).
 3: Calculate the particle weights $\{\omega_{k|k-1}^i\}_{i=1}^{N_{k-1}}$ by Eq. (4).
 4: Sample $N_B$ born particles $\{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=N_{k-1}+1}^{N_{k-1}+N_B}$ uniformly around each measurement by Eq. (5) and Eq. (6).
 5: Combine all the particles: $\{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=1}^{N_k}$ = $\{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=1}^{N_{k-1}} \cup \{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=N_{k-1}+1}^{N_{k-1}+N_B}$.
 6: for $\lambda \in [0, \triangle\lambda, 2\triangle\lambda, \cdots, N_\lambda\triangle\lambda]$ do
 7:     Calculate the likelihood $h_k^i$ by Eq. (9).
 8:     Calculate particle flow $\boldsymbol{f}_k^i$ by Eq. (8).
 9:     Calculate $\triangle\boldsymbol{m}_{k|k-1}^i$ by Eq. (7).
10:     Update each particle state by Eq. (10).
11: end for
12: Update the particle weights $\{\omega_{k|k-1}^i\}_{i=1}^{N_k}$ to obtain $\{\omega_k^i\}_{i=1}^{N_k}$ by Eq. (11) and calculate $\tilde{N}_k = \sum_{i=1}^{N_k} \omega_k^i$.
13: Set $\{\boldsymbol{m}_k^i\}_{i=1}^{N_k}$ as $\{\boldsymbol{m}_{k|k-1}^i\}_{i=1}^{N_k}$.
14: Cluster particles and get $\{\tilde{\boldsymbol{m}}_k^j, \tilde{\omega}_k^j\}_{j=1}^{\tilde{N}_k}$ by the K-means method or MEAP
15: Calculate $\{\boldsymbol{P}_k^i\}_{i=1}^{N_k}$ by Kalman filter or clustered particle group.
16: if ESS < $N_k/2$ then
17:     Resample $\{\boldsymbol{m}_k^i, \omega_k^i\}_{i=1}^{N_k}$.
18: end if
```

Finally, resampling is performed when the effective sample size (ESS) [27] is smaller than half number of particles. In the resampling step, we can obtain $\{\boldsymbol{m}_k^i, \omega_k^i\}_{i=1}^{N_k}$, where $\{\omega_k^i\}_{i=1}^{N_k} = 1/N_k$.

The NPF has mitigated the weight degeneracy problem in the SMC-PHD filter under the assumption that all targets are on the scene (visually) or active (continuously talking) during tracking. However, the LOCATA includes the practical challenges of data processing of conversational speech, such as natural speech inactivity during sentences, sporadic utterances and dialogues between multiple talkers. Therefore, the clutter intensity and detection probability should be considered for multi-speaker tracking.

3. IPF-SMC-PHD FILTER

To address the limitations of the NPF-SMC-PHD filter, we propose the IPF-SMC-PHD filter for the task 4 of the LOCATA challenge. The measurements of the IPF-SMC-PHD filter is given by the MUSIC, which is the baseline method of the LOCATA challenge. In this section, the IPF and identification of the speaker are discussed.

3.1. Intensity particle flow

The IPF is used to replace the NPF, lines 6-11 of the Algorithm 1. For decreasing the computational cost, we only update the survival particles by the IPF, since the born particles are created as the measurements. After the prediction step, the particle set is shown as $\{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=1}^{N_{k-1}}$. Based on the intensity function [28], the

Algorithm 2 Intensity particle flow step in the IPF-SMC-PHD filter

```
Input: ${\boldsymbol{m}^i_{k|k-1}, \omega^i_{k|k-1}, \boldsymbol{P}^i_{k|k-1}}_{i=1}^{N_{k-1}}$ and $\boldsymbol{Z}_k$.
Output: ${\boldsymbol{m}^i_k, \omega^i_k}_{i=1}^{N_{k-1}}$.
    Initialize: $\boldsymbol{R}$, $p_{D,k}$, $\triangle\lambda$, $\kappa_k$, $\upsilon^i_k$, $\partial_{\boldsymbol{m}}$ and $\boldsymbol{w}^i_k$.
for $\lambda \in [0, \triangle\lambda, 2\triangle\lambda, \cdots, N_\lambda \triangle\lambda]$ do
    for Each surviving particles do
        Calculate the likelihood density $\boldsymbol{h}^{i,r}_k$ based on the probability density function of the Gaussian distribution $\mathcal{N}(\boldsymbol{z}^r_k, \boldsymbol{R})$.
        Calculate $\boldsymbol{\nabla} h^{i,r}_k$ ,$\boldsymbol{\nabla}(\boldsymbol{\nabla} h^{i,r}_k)$ by Eq. (15) and Eq. (16).
        Calculate particle flow by Eq. (12).
        Calculate $\triangle\boldsymbol{m}^i_{k|k-1}$ by Eq. (7).
        if $\triangle\boldsymbol{m}^i_{k|k-1} < \partial_{\boldsymbol{m}}$ then
            Stop calculating the particle flow for this surviving particle.
        end if
        Update each particle state by Eq. (10).
        Update the weights of the particles ${\omega^i_{k|k-1}}_{i=1}^{N_{k-1}}$ to obtain ${\omega^i_k}_{i=1}^{N_{k-1}}$ by Eq. (11).
    end for
end for
Set ${\boldsymbol{m}^i_k}_{i=1}^{N_{k-1}}$ as ${\boldsymbol{m}^i_{k|k-1}}_{i=1}^{N_{k-1}}$.
```

particle flow can be calculated according to

$$\boldsymbol{f}^i_k = -[\sum_{r=1}^{R_k} \frac{\lambda p_{D,k} \boldsymbol{\nabla}(\boldsymbol{\nabla} h^{i,r}_k)}{G^r_k} + \boldsymbol{\nabla}(\boldsymbol{\nabla} \ln(\omega^i_{k|k-1}))]^{-1} \cdot \sum_{r=1}^{R_k} \frac{p_{D,k} \boldsymbol{\nabla} h^{i,r}_k}{G^r_k} \tag{12}$$

where

$$G^r_k = \kappa_k + \sum_{i=N_{k-1}+1}^{N_{k-1}+N_B} S^{i,r}_k + \sum_{i=1}^{N_{k-1}} h^{i,r}_k \omega^i_{k|k-1} \tag{13}$$

$$S^{i,r}_k = \gamma_k(\boldsymbol{m}^{i,r}_{k|k-1}) * \max(0, 1 - \sum_{i=1}^{N_{k-1}} h^{i,r}_k \omega^{i,r}_{k|k-1}) \tag{14}$$

where $S^{i,r}_k$ is the birth intensity function for the i-th particle and the r-th DOA line at k. $\boldsymbol{\nabla}(\boldsymbol{\nabla} \ln(\omega^i_{k|k-1}))$ is independent of the particle state and a constant for the particle flow. If we assume that likelihood model is Gaussian, the particle flow in Eq. (12) may be derived analytically for particle motion. The differentiation of the likelihood $h^{i,r}_k$ is calculated as follows:

$$\boldsymbol{\nabla} h^{i,r}_k = -h^{i,r}_k \boldsymbol{R}^{-1}(\boldsymbol{f_z}(\tilde{\boldsymbol{m}}_k, \varsigma_k) - \boldsymbol{z}^r_k) \tag{15}$$

$$\begin{aligned}\boldsymbol{\nabla}(\boldsymbol{\nabla} h^{i,r}_k) =& h^{i,r}_k [\boldsymbol{R}^{-1}(\boldsymbol{f_z}(\tilde{\boldsymbol{m}}_k, \varsigma_k) - \boldsymbol{z}^r_k) \\ & (\boldsymbol{f_z}(\tilde{\boldsymbol{m}}_k, \varsigma_k) - \boldsymbol{z}^r_k)^{-1} \boldsymbol{R} - \boldsymbol{R}^{-1}]\end{aligned} \tag{16}$$

With the increment of λ, the rate of change of $\triangle\boldsymbol{m}^i_{k|k-1}$ may decrease. If $\triangle\boldsymbol{m}^i_{k|k-1}$ is smaller than the sensor resolution $\partial_{\boldsymbol{m}}$, $\boldsymbol{m}^i_{k|k-1}$ is invariant based on Eq (10) after $\triangle\boldsymbol{m}^i_{k|k-1}$ is added to $\boldsymbol{m}^i_{k|k-1}$, which is inefficient and wasteful. So if $\triangle\boldsymbol{m}^i_{k|k-1} < \partial_{\boldsymbol{m}}$, the particle flow step would be ignored. The pseudo code of IPF in the IPF-SMC-PHD filter is shown in Algorithm 2.

Algorithm 3 Identification step in the IPF-SMC-PHD filter

```
Input: ${\hat{\boldsymbol{m}}^j_{k-1}}_{j=1}^{\hat{N}_m}$ and ${\tilde{\boldsymbol{m}}^j_k}_{j=1}^{\tilde{N}_k}$
Output: ${\hat{\boldsymbol{m}}^j_k}_{j=1}^{\hat{N}_m}$
    Initialize: $\boldsymbol{d}$ and $\hat{N}_m$.
if $\tilde{N}_k = \hat{N}_m$ then
    for $j \in [1, .., \tilde{N}_k]$ do
        $\hat{\boldsymbol{m}}^j_k = \underset{\tilde{\boldsymbol{m}}^j_k}{\text{argmin}} \left\| \hat{\boldsymbol{m}}^j_{k-1} - \tilde{\boldsymbol{m}}^j_k \right\|$
    end for
end if
if $\tilde{N}_k > \hat{N}_m$ then
    for $j \in [1, .., \hat{N}_m]$ do
        $\hat{\boldsymbol{m}}^j_k = \underset{\tilde{\boldsymbol{m}}^j_k}{\text{argmin}} \left\| \hat{\boldsymbol{m}}^j_{k-1} - \tilde{\boldsymbol{m}}^j_k \right\|$
    end for
end if
if $\tilde{N}_k < \hat{N}_m$ then
    for $j \in [1, .., \tilde{N}_k]$ do
        $\hat{\boldsymbol{m}}^j_k = \underset{\tilde{\boldsymbol{m}}^j_k}{\text{argmin}} \left\| \hat{\boldsymbol{m}}^j_{k-1} - \tilde{\boldsymbol{m}}^j_k \right\|$
        if $\left\| \hat{\boldsymbol{m}}^j_{k-1} - \hat{\boldsymbol{m}}^j_k \right\| > \boldsymbol{d}$ then
            $\hat{\boldsymbol{m}}^j_k = \boldsymbol{f}_{\tilde{\boldsymbol{m}}} \left( \hat{\boldsymbol{m}}^j_{k-1}, \boldsymbol{\tau}_k \right)$
        end if
    end for
end if
```

3.2. Identification of the speaker

As all estimated positions must be associated with an identity (ID) in the LOCATA challenge, the estimates resulting from the IPF-SMC-PHD filter should consider false tracks, missing tracks, broken tracks and track swaps. However, the PHD filter does not consider the identity of speakers. An assistant identifier should be added. Since the number of speakers is not known, the identification problem is normally solved by the Blind Source Separation (BSS) method. However, the BSS has a high computational complexity. As the IPF-SMC-PHD filter can provide the estimate of the speaker state, in our proposed method, the speaker identity is estimated by the speaker states under the assumption that it is not changing abruptly in subsequent frames. Although the number of speakers at each frame $\{\tilde{N}_k\}_{i=1}^k$ has been estimated at the line 12 of the Algorithm 1, the estimated number is varying due to the noise and undetected DOA lines. For smoothing the trajectory of speakers, we assume the mean number of speakers $\hat{N}_m$ is given by:

$$\hat{N}_m = \frac{\sum_{i=1}^k N_k}{k} \tag{17}$$

For each frame, if the number of the estimated speakers N_k is larger than $\hat{N}_m$ at frame k, it may imply that the noises are estimated as the speakers. To detect the noise, the distance from the estimated speaker state at k and the speaker state $k-1$ is considered. As we assume the positions are not changed abruptly in subsequent frames, the estimated speaker state at k with less distance to the state at $k-1$ is considered as the speaker state at k, where $j \in 1, ..., \hat{N}_m$. If the number of the estimated speakers N_k is less than $\hat{N}_m$ at frame k, it may imply the miss detection of speakers. The undetected speaker states are updated by the velocity as Eq. (1). If the number of the estimated speakers N_k is equal to $\hat{N}_m$ at

Table 1: The index of used microphones and the number of the subspaces for the DICIT array, Eigenmike array, Robot head and Hearing aids.

Array	Index	Number
DICIT	5,6,7,9,10	1,2
Eigenmike	1,...,32	1,2,3,4,5
Robot head	1,...,12	1,2,3,4
Hearing aids	1,2,3,4	3,4

frame k. The identify of the speaker is given based on the distance from the estimated speaker state to the speaker state at last frame. The pseudo code of identification step in the IPF-SMC-PHD filter is shown in Algorithm 3, where $\{\hat{\boldsymbol{m}}_{k-1}^{j}\}_{j=1}^{\hat{N}_m}$ is the set of the speaker states which is ordered by ID, for example, $\hat{\boldsymbol{m}}_{k-1}^{1}$ means the state of the first speaker.

4. EXPERIMENTAL RESULTS

In this section, the proposed algorithm is compared with its baseline counterparts including the NPF-SMC-PHD [21], SMC-PHD algorithms [29] and the baseline MUSIC of the LOCATA dataset [23]. The parameters of the PHD filter and particle flow filters are set as in [29] and [4]. The number of particles per speaker is 50 and the particles are spread randomly in the tracking area. The experiments are run in Matlab on Windows 7 with Intel i7 (3.2 GHz).

The LOCATA dataset consists of sequences where multiple speakers may speak or walk. Those actions are recorded by four circular eight-element microphone arrays at 48 kHz. Although the baseline MUSIC method is provided by the LOCATA challenge, the MUSIC only considers one speaker. So we consider more signal subspaces to calculate the DOA lines than the baseline MUSIC. The parameters of the microphone arrays are shown in Table 1, which are chosen based on the ground truth dataset of the Task 1 and Task 2.

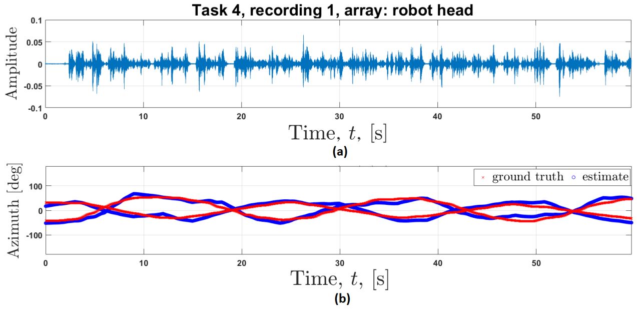

Figure 1: The audio signal is illustrated in (a), and (b) shows the speaker state estimated by the IPF-SMC-PHD filter and ground truth speaker state with the Robot head array on recording 1 of the evaluation data of the task 4.

Due to the space limitation, we only show the tracking result on recording 1 on the robot head. Figure 1a shows the signal representation of recording 1 of task 4. Speaker states are indicated with blue and red line in Figure 1b, respectively for the IPF-SMC-PHD and ground truth. Here, we performed down-sampling to the plots for better visualization. At the beginning of the recording, the speakers are silent and the estimates are calculated when the speakers start to talk. Although the filter can detect the occlusions, the error increases when the occlusions happens.

Table 2: The OSPA for the IPF-SMC-PHD, NPF-SMC-PHD, ZPF-SMC-PHD filters and MUSIC algorithm in terms of the OSPA error on the Locata task 4.

Array	Recording	IPF	NPF	SMC	MUSIC
Robot head	1	**1.084**	1.178	1.247	1.875
	2	**1.079**	1.165	1.242	1.753
	3	**1.093**	1.205	1.253	1.897
DICIT	1	**4.826**	5.893	7.089	10.357
	2	**4.543**	5.407	6.580	10.182
	3	**5.405**	6.777	7.860	11.057
Hearing aids	1	**4.833**	5.894	7.091	10.360
	2	**4.591**	5.603	6.736	9.848
	3	**5.310**	6.507	7.895	11.490
Eigenmike	1	**1.465**	1.559	1.568	2.288
	2	**1.295**	1.461	1.616	2.212
	3	**1.399**	1.503	1.656	2.429
Average OSPA		**3.077**	3.679	4.319	6.312

The Optimal Sub-pattern Assignment (OSPA) for trackers [30], which gives a combined score for the estimation performance in the number of sources and their positions, is used to evaluate the tracking accuracy. Table 2 reports the average OSPA over 10 random tests. With the contribution of the IPF, 16% reduction in tracking error has been achieved as compared with the NPF-SMC-PHD filter. In addition, the IPF-SMC-PHD filter also improves the estimation accuracy by 29% and 51% over the SMC-PHD and baseline MUSIC method, respectively. However, the running time of IPF (about 10s/frame) is three times and ten times of the SMC-PHD filter (about 3s/frame) and the MUSIC method (about 1s/frame), respectively.

5. CONCLUSION

We have presented a novel IPF-SMC-PHD filter for audio multi-speaker tracking by smoothly migrating the particles. The proposed algorithm has been tested on the task 4 of the LOCATA dataset. The experimental results show that the proposed filter offers a higher tracking accuracy than the baseline methods with a higher computational cost.

6. ACKNOWLEDGMENT

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1), the BBC as part of the BBC Audio Research Partnership, the China Scholarship Council (CSC), and in part by the EPSRC grant EP/K014307/2.

7. REFERENCES

[1] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti, "Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 38–51, Mar. 2005.

[2] H.-S. Yeo, B.-G. Lee, and H. Lim, "Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware," *Multimedia Tools and Applications*, vol. 74, no. 8, pp. 2687–2715, 2015.

[3] H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, *et al.*, "Multicenter evaluation of signal enhancement algorithms for hearing aids," *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1491–1505, 2010.

[4] Y. Liu, W. Wang, J. Chambers, V. Kilic, and A. Hilton, "Particle flow SMC-PHD filter for audio-visual multi-speaker tracking," in *Proc. IEEE Intl. Conf. Latent Variable Analysis and Signal Separation*, Mar. 2017, pp. 344–353.

[5] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos, "Estimation of direction of arrival using information theory," *IEEE Signal Processing Letters*, vol. 12, no. 8, pp. 561–564, 2005.

[6] B. Qin, H. Zhang, Q. Fu, and Y. Yan, "Subsample time delay estimation via improved GCC PHAT algorithm," in *Proc. IEEE Intl. Conf. on Signal Processing*, 2008, pp. 2579–2582.

[7] A. Johansson and S. Nordholm, "Robust acoustic direction of arrival estimation using Root-SRP-PHAT, a realtime implementation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 4, 2005, pp. iv–933.

[8] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Trans. Signal Processing*, vol. 54, no. 9, pp. 3291–3304, 2006.

[9] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 186–200, Feb. 2015.

[10] S. S. Haykin *et al.*, *Kalman Filtering and Neural Networks*. Wiley Online Library, 2001.

[11] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund, "Particle filters for positioning, navigation, and tracking," *IEEE Trans. on signal processing*, vol. 50, no. 2, pp. 425–437, 2002.

[12] R. P. Mahler, "A theoretical foundation for the stein-winter probability hypothesis density (PHD) multitarget tracking approach," DTIC Document, Tech. Rep., 2000.

[13] R. Mahler, "PHD filters of higher order in target number," *IEEE Trans. Aerospace and Electronic Systems*, vol. 43, no. 4, 2007.

[14] B.-N. Vo and M. Wing-Kin, "The Gaussian mixture probability hypothesis density filter," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4091–4104, Oct. 2006.

[15] B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo methods for multitarget filtering with random finite sets," *IEEE Trans. Aerospace and Electronic Systems*, vol. 41, no. 4, pp. 1224–1245, 2005.

[16] F. Daum and J. Huang, "Nonlinear filters with log-homotopy," in *Proc. IEEE Int. Conf. Information Processing of Small Targets*. International Society for Optics and Photonics, 2007, pp. 669 918–669 918.

[17] P. Bunch and S. Godsill, "Approximations of the optimal importance density using Gaussian particle flow importance sampling," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 748–762, 2016.

[18] F. Daum, J. Huang, and A. Noushin, "Coulomb's law particle flow for nonlinear filters," in *Proc. SPIE Conf. Signal and Data Processing*, O. E. Drummond, Ed. International Society for Optics and Photonics, Aug. 2011, pp. 1–15.

[19] F. Daum and J. Huang, "Zero curvature particle flow for nonlinear filters," in *Proc. SPIE Symposium on Signal and Data Processing of Small Targets*. International Society for Optics and Photonics, Apr. 2013, pp. 83 930A–83 930A–11.

[20] ——, "Particle flow with non-zero diffusion for nonlinear filters," in *Proc. SPIE Conf. Signal Processing, Sensor Fusion, and Target Recognition*, 7697, Ed., vol. 04, 2013, pp. 87 450P–87 450P–13.

[21] Y. Liu, A. Hilton, J. Chambers, Y. Zhao, and W. Wang, "Non-zero diffusion particle flow SMC-PHD filter for audio-Visual multi-speaker tracking," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2018.

[22] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The locata challenge data corpus for acoustic source localization and tracking," in *IEEE Sensor Array and Multi-channel Signal Processing Workshop*, 2018.

[23] ——, "The locata challenge data corpus for acoustic source localization and tracking."

[24] F. Daum, J. Huang, and A. Noushin, "Exact particle flow for nonlinear filters," in *Proc. SPIE Conf. Signal Processing Sensor Fusion, Target Recognition*. International Society for Optics and Photonics, Apr. 2010, pp. 769 704–1–769 704–19.

[25] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. the Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[26] T. Li, S. Sun, M. Bolić, and J. M. Corchado, "Algorithm design for parallel implementation of the SMC-PHD filter," *Signal Processing*, vol. 119, pp. 115–127, 2016.

[27] A. Kong, J. S. Liu, and W. H. Wong, "Sequential imputations and Bayesian missing data problems," *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 278–288, 1994.

[28] B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo implementation of the PHD filter for multi-target tracking," in *Proc. IEEE Int. Conf. Information Fusion*, July 2003, pp. 792–799.

[29] V. Kilic, M. Barnard, W. Wang, A. Hilton, and J. Kittler, "Mean-shift and sparse sampling based SMC-PHD filtering for audio informed visual speaker tracking," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2417–2431, 2016.

[30] B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Trans. Signal Processing*, vol. 59, no. 7, pp. 3452–3457, 2011.