Audio-Visual Particle Flow SMC-PHD Filtering for Multi-Speaker Tracking

Yang Liu, *Student Member, IEEE,* Volkan Kılıç, Jian Guan, Wenwu Wang, *Senior Member, IEEE,*

***Abstract*—Sequential Monte Carlo probability hypothesis density (SMC-PHD) filtering is a popular method used recently for audio-visual (AV) multi-speaker tracking. However, due to the weight degeneracy problem, the posterior distribution can be represented poorly by the estimated probability, when only a few particles are present around the peak of the likelihood density function. To address this issue, we propose a new framework where particle flow (PF) is used to migrate particles smoothly from the prior to the posterior probability density. We consider both zero and non-zero diffusion particle flows (ZPF/NPF), and developed two new algorithms, AV-ZPF-SMC-PHD and AV-NPF-SMC-PHD, where the speaker states from the previous frames are also considered for particle relocation. The proposed algorithms are compared systematically with several baseline tracking methods using the AV16.3, AVDIAR and CLEAR datasets, and are shown to offer improved tracking accuracy and average effective sample size (ESS).**

***Index Terms*—Audio-Visual Tracking, Sequential Monte Carlo, PHD filter, Particle Flow, Optimal Proposal Distribution.**

I. Introduction

MULTI-SPEAKER tracking has drawn increasing attention in applications, such as security surveillance [1], and human-computer interaction [2]. However, the measurements used in multi-speaker tracking, either audio [3] or visual [4], often contain noise, clutter, and missing data [5], [6]. For example, in visual tracking, the tracking result is often affected by occlusions and the limited field of view of cameras [4], while in audio tracking, speakers are not always detectable when strong background noise and room reverberations are present in the measurements or when the speakers are silent.

To address this problem, different modalities can be exploited jointly for their complementarity. For example, speakers can be tracked using audio information, if they are visually occluded, likewise, they can be tracked with visual information, when the audio information becomes unreliable, e.g. due to the presence of acoustic noise. Other modalities, such as thermal vision and laser rangefinders, could also be considered, however, we will focus on audio-visual sensors, as in [7], due to their widespread use, low cost, and easy installation [8]. More specifically, in our work, we consider the visual measurements obtained by the CAMShift [9] or a face detector [10], and the audio measurements, such as the direction of arrival (DOA) of the sources [11].

To fuse the audio-visual data, the Bayesian inference framework is a popular choice, which provides an intuitive way for the estimation of speaker states [12] from the measurements. Early methods include the Kalman filter (KF) [13], extended Kalman filter (EKF) [14] and particle filter [15], which can be used for a fixed and known number of speakers, while more recent methods include random finite sets (RFS) [16], Gaussian mixture (GM) PHD filter [6], sequential Monte Carlo (SMC) PHD filter [17], cardinalized PHD filter [18], and generalized labeled multi-Bernoulli (GLMB) RFS [19], which can be employed to track an unknown and time-variant number of speakers. The SMC-PHD filter uses a set of random particles to estimate the posterior density, as a result, it often suffers from the weight degeneracy problem [20], i.e. the weights of most particles will become negligible, while only a few remain significant, during the iterations in the particle propagation process.

To address the issue, several ideas have been developed. The SMC methods exploit the most recent measurements or the unscented transformation to approximate the optimal proposal distribution and to minimize the variance of the importance weights, as in e.g. the auxiliary particle filter [21], unscented particle filter [22], auxiliary SMC-PHD filter [23] and unscented auxiliary cardinalized PHD filter [24]. Another idea is based on the Markov Chain Monte Carlo method [25], as performed in the well-known resample-move algorithm [26], where the particles are drawn to represent independent samples from the target posterior. Finally, an idea based on bridging densities [27], [28], [29] has also been developed for approaching the true posterior density from the tractable prior density. This method offers theoretical elegance and promising performance, but involves complicated approximation of the optimal bridging densities.

In this paper, we propose a new method to address the weight degeneracy issue, based on particle flow [30], [31], [32]. Different from the above-mentioned techniques, such as the popular particle re-sampling technique, our method aims to improve the effective sample size with a particle relocation strategy designed using particle flow. More specifically, the particles are migrated from the prior to the posterior distribution, using a homotopy function which defines the flow in synthetic time and incorporated for particle update at each time frame [33].

According to the different assumptions employed for solving the homotopy function, particle flow can be divided into five classes: incompressible particle flow [34], zero diffusion

Y. Liu and W. Wang are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, U.K. E-mails: {yangliu, w.wang}@surrey.ac.uk.

V. Kılıç is with Department of Electrical and Electronics Engineering, Izmir Katip Celebi University, 35620 Cigli-Izmir, Turkey. E-mail: volkan.kilic@ikc.edu.tr.

J. Guan is with College of Computer Science and Technology, Harbin Engineering University, Harbin, 150001, China. E-mail: j.guan@hrbeu.edu.cn.

particle flow (ZPF) [32], Coulomb's law particle flow [35], zero-curvature particle flow [36] and non-zero diffusion particle flow (NPF) [37]. In this work, we consider ZPF and NPF. The ZPF is easy to implement [38] and widely used [39], [40], [41]. NPF does not assume any particular form of the prior density and likelihood function [37]. NPF offers slightly better performance than ZPF. Note that, we do not consider the other types of particle flows for the following reasons. Incompressible particle flow equation only works on some special prior densities, such as Gaussian density [42]. Coulomb's law and zero-curvature particle flows are sensitive to initialization of the state vector [36].

Zero particle flow has been used previously to improve the particle filter by reducing the number of particles in the particle flow particle filter (PFPF) [43] and the δ-GLMB particle filter [44], and to improve the Gaussian mixture PHD filter in [29] for an unknown number of targets. However, there are several main differences between our proposed methods and these methods. First, particle flow is used to improve the SMC-PHD filter in our method, mainly for particle relocation and weight update in order to mitigate the weight degeneracy issue. Second, only ZPF has been considered in the previous methods, while we have also developed a new method using NPF. Third, the mean and covariance of the particles are estimated by a clustering technique, while in previous methods, these are estimated using EKF [43], Gaussian mixture model [29], or label information [44]. Fourth, our proposed methods offer significant performance improvements and computational advantages compared with these previous methods.

To demonstrate the advantages of the proposed method, we consider a recent baseline i.e. audio-visual SMC-PHD (AV-SMC-PHD) filter introduced in [5], and developed two new algorithms, namely, AV-ZPF-SMC-PHD and AV-NPF-SMC-PHD. We compare systematically the proposed algorithms with several other state-of-the-art methods, such as the sparse-AVMS-SMC-PHD filter in [5], where mean-shift (MS) was used for particle relocation.

Preliminary results of this work were presented in conference papers [20] [7] [45]. This paper provides a comprehensive treatment of the proposed methods together with new improvements and experimental results. First, the direction of arrival (DOA) and color histograms are both used for deriving the particle flow. When speakers are not detected with the DOA information or the color histograms, particle states can still be updated with particle flow. Second, the speaker states and weights in the previous frames are used for relocating particles in terms of DOA, in order to reduce the adverse impact of acoustic noise on particle relocation. Third, we perform extensive experiments on the AV16.3 [46], AVDIAR [47] and CLEAR [48] datasets, and compare the proposed method with several baseline methods including the PF-PF [41], ZPF-GPF-PHD [29], sparse-AVMS-SMC-PHD [5] filters, auxiliary SMC-PHD filter [23], and a deep learning based face detector [10].

This paper is organized as follows. The next section discusses the problems and background. Section III presents the details of the proposed methods. In Section IV, the proposed algorithms are compared with several baseline algorithms based on comprehensive experiments. Finally, Section V concludes the paper.

II. Problem Statement And Background

This section describes the problem formulation, the SMC-PHD filter, the particle flow filter, and a baseline AV-SMC-PHD algorithm. We assume that the speaker dynamics and measurements are described as a Markov state-space signal model:

$$\{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k} = \mathbf{F}_{\tilde{\boldsymbol{m}}}\left(\{\tilde{\boldsymbol{m}}_{k-1}^j\}_{j=1}^{\tilde{N}_{k-1}}, \boldsymbol{\tau}_k\right) \tag{1}$$

$$\boldsymbol{Z}_k = \mathbf{F}_{\boldsymbol{z}}\left(\{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k}, \varsigma_k\right) + \epsilon_k \tag{2}$$

where $\tilde{\boldsymbol{m}}_k^j \in \mathbb{R}^M$ represents the state vector for the jth speaker at time k, and $\tilde{\ }$ is used to distinguish the speaker state from the particle state used later. The state $\tilde{\boldsymbol{m}}_k^j = [x_k^j, y_k^j, \dot{x}_k^j, \dot{y}_k^j]^T$ consists of positions (x_k^j, y_k^j) and velocities $(\dot{x}_k^j, \dot{y}_k^j)$, while the observation is a noisy version of the position. Hence $M = 4$. For 3D calculations, the speaker state is set as $\tilde{\boldsymbol{m}}_k^j = [x_k^j, y_k^j, w_k^j, \dot{x}_k^j, \dot{y}_k^j, \dot{w}_k^j]^T$, where (x_k^j, y_k^j, w_k^j) is the 3D position of the jth speaker and $(\dot{x}_k^j, \dot{y}_k^j, \dot{w}_k^j)$ is the speaker velocity. At time k, $\tilde{\boldsymbol{m}}_k^j$ is approximated by N_k particles $\{\boldsymbol{m}_k^i\}_{i=1}^{N_k}$ with weights $\{\omega_k^i\}_{i=1}^{N_k}$. Let $\boldsymbol{Z}_k$ denote the set of measurements at time k, defined as $[\{\mathring{\boldsymbol{m}}_k^o\}_{o=1}^{O_k}, \{\breve{\boldsymbol{m}}_k^u\}_{u=1}^{U_k}]$ for audio-visual measurements, where O_k and U_k are the number of audio $\mathring{\ }$ and visual $\breve{\ }$ measurements, respectively. $\boldsymbol{\tau}_k$ and ς_k are system excitation and measurement noise terms, respectively. ϵ_k is the clutter term. $\mathbf{F}_{\tilde{\boldsymbol{m}}}$ is the transition model and $\mathbf{F}_{\boldsymbol{z}}$ is the nonlinear measurement model. A list of important notations is given in Table I, where o and u are the indices of audio and visual measurements.

TABLE I: List of important Notations

Notations	Meaning
$\{\boldsymbol{m}_k^i, \omega_k^i\}_{i=1}^{N_k}$	The set of particle states and weights
$\{\tilde{\boldsymbol{m}}_k^j, \tilde{\omega}_k^j\}_{j=1}^{\tilde{N}_k}$	The set of speaker states and weights
$\{\tilde{\boldsymbol{m}}_{k\|k-1}^j\}_{j=1}^{\tilde{N}_{k\|k-1}}$	The set of candidate speaker states
$\{\breve{\boldsymbol{m}}_k^u, \breve{\omega}_k^u\}_{u=1}^{U_k}$	The set of visual measurements and weights
$\{\mathring{\boldsymbol{m}}_k^o, \mathring{\omega}_k^o\}_{o=1}^{O_k}$	The set of audio measurements and weights

A. SMC-PHD filter

In the prediction step [49], the particles are obtained by the proposal distribution $q_k(\boldsymbol{m}_{k|k-1}^i|\boldsymbol{m}_{k-1}^i, \boldsymbol{Z}_k)$, with weights

$$\omega_{k|k-1}^i = \frac{\phi_{k|k-1}(\boldsymbol{m}_{k|k-1}^i|\boldsymbol{m}_{k-1}^i)\omega_{k-1}^i}{q_k(\boldsymbol{m}_{k|k-1}^i|\boldsymbol{m}_{k-1}^i, \boldsymbol{Z}_k)}, i = 1, ..., N_{k-1} \tag{3}$$

where $\phi_{k|k-1}$ is the analogue of the state transition probability. N_B particles are sampled from the new born importance function p_k with weights

$$\omega_{k|k-1}^i = \frac{\gamma_k(\boldsymbol{m}_{k|k-1}^i)}{N_B p_k(\boldsymbol{m}_{k|k-1}^i|\boldsymbol{Z}_k)}, \quad i = N_{k-1}+1, \ldots, N_{k-1}+N_B \tag{4}$$

where γ_k is the probability of the new born speaker, whose integral approximates the average number of speakers in the state space. In the update step, the weights are calculated as

$$\omega_k^i = \left[1 - p_{D,k}^i + \sum_{\boldsymbol{z}_k^r \in \boldsymbol{Z}_k} \frac{p_{D,k}^i h_k^{i,r}}{\kappa_k + \sum_{i=1}^{N_k} p_{D,k}^i h_k^{i,r} \omega_{k|k-1}^i}\right] \omega_{k|k-1}^i \tag{5}$$

in which κ_k denotes the clutter intensity of the rth measurement $\boldsymbol{z}_k^r$ at time k. $p_{D,k}^i$ is the detection probability of the ith particle at time k. κ_k and $p_{D,k}^i$ can be assumed known a priori and constant as in [50]. Alternatively, a model of Beta-Gaussian mixtures can be used to estimate unknown κ_k and $p_{D,k}^i$ as in [51]. $h_k^{i,r}$ is the likelihood of the ith particle for the rth measurement, which can be estimated in terms of their Bhattacharyya distance [52] or Euclidean distance [53] with a Gaussian distribution. The number of speakers $\tilde{N}_k$ is estimated as the sum of the weights. The states and weights of the speakers $\{\tilde{\boldsymbol{m}}_k^j, \tilde{\omega}_k^j\}_{j=1}^{\tilde{N}_k}$ can be calculated using e.g. k-means clustering method [54] or multi-expected a posterior (MEAP) [55]. Finally, resampling is performed when the ESS is smaller than half of the number of particles. Resampling was proposed originally in [56] and popularised in [57]. According to the different strategies taken for selecting the particles, the resampling methods mainly fall into three categories: multinomial resampling [58], residual resampling [58] and systematic resampling [59]. In multinomial resampling, the particles are independently selected and resampled. In residual resampling, the particles to be resampled are selected based on the values of their weights. In systematic resampling, the particles are clustered and the resampled particles are taken from each cluster. This can help reduce the discrepancy among the particles. The systematic resampling method has been used in the baseline SMC-PHD filter and will also be used in our proposed filter, due to its high resampling quality and low computational complexity.

B. Particle flow

In particle flow, a homotopy function is used to estimate the posterior density, as follows [33],

$$\log(\psi_k^i) = \log(g_k^i) + \lambda \log(h_k^i) - \log K_k^i \tag{6}$$

where K_k^i is the normalization constant independent of $\boldsymbol{m}_k^i$, and λ, called pseudo time, is a step size parameter taking values from set $[0, \triangle\lambda, 2\triangle\lambda, \cdots, N_\lambda\triangle\lambda]$ and $N_\lambda\triangle\lambda = 1$, with N_λ being the number of pseudo time steps. In fact, λ can also be variable step sizes as in [41], [32], which is further discussed in Section IV. Eq. (6) represents the evolution of the particles from the prior density g_k^i to the posterior density. When $\lambda = 0$, ψ_k^i represents the prior density g_k^i at time k. When λ is varied to 1, ψ_k^i is translated into the normalized posterior density [30]. In ZPF, the posterior and likelihood function are assumed to be Gaussian and linear, and the flow $\boldsymbol{f}_{k,\lambda}^i$ is derived as,

$$\boldsymbol{f}_{k,\lambda}^i = \frac{d\boldsymbol{m}_{k|k-1}^i}{d\lambda} = \boldsymbol{A}_k^i \boldsymbol{m}_{k|k-1}^i + \boldsymbol{b}_k^i \tag{7}$$

where

$$\boldsymbol{A}_k^i = -\frac{1}{2}\boldsymbol{C}_k^i \left(\lambda \boldsymbol{H}_k^i \boldsymbol{P}_{k|k-1}^i (\boldsymbol{H}_k^i)^T + \boldsymbol{R}\right)^{-1} \boldsymbol{H}_k^i, \tag{8}$$

$$\boldsymbol{b}_k^i = \left(\boldsymbol{I} + 2\lambda \boldsymbol{A}_k^i\right) \left[\left(\boldsymbol{I} + \lambda \boldsymbol{A}_k^i\right) \boldsymbol{C}_k^i \boldsymbol{R}^{-1} \boldsymbol{z}_k^r + \boldsymbol{A}_k^i \bar{\boldsymbol{m}}_{k|k-1}^i\right] \tag{9}$$

$$\boldsymbol{C}_k^i = \boldsymbol{P}_{k|k-1}^i (\boldsymbol{H}_k^i)^T \tag{10}$$

where $\boldsymbol{P}_{k|k-1}^i$ is the covariance matrix of the prediction error for the particle state $\boldsymbol{m}_{k|k-1}^i$, $\bar{\boldsymbol{m}}_{k|k-1}^i$ is the mean of the states over the particle set, $\boldsymbol{R}$ is the covariance matrix of the measurement noise, $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{H}_k^i$ is a Jacobian matrix [60]. The details about ZPF are given in [33]. ZPF has been widely used for its simplicity. For nonlinear problems, this flow can be used to linearize the measurement equation for each particle with an extended Kalman filter. The main critical parameters in ZPF are the number of particles and the tolerance for the step-size selection. As a result, only little effort is required for parameter tuning.

Different from ZPF, the particle flow equation used in NPF is derived by retaining the diffusion term, as detailed in [37], given as follows,

$$\boldsymbol{f}_{k,\lambda}^i = -[\boldsymbol{\nabla}^2 \log \psi_k^i]^{-1} (\boldsymbol{\nabla} \log h_k^i) \tag{11}$$

where

$$\boldsymbol{\nabla}^2 \log \psi_k^i \approx -(\boldsymbol{P}_{k|k-1}^i)^{-1} + \lambda \boldsymbol{\nabla}^2 \log h_k^i \tag{12}$$

where $\boldsymbol{\nabla}$ is the spatial vector differential operator $\frac{\partial}{\partial \boldsymbol{m}_{k|k-1}^i}$. h_k^i is the likelihood of the ith particle at time k. As compared to ZPF, the prior density and likelihood function in NPF need to be sufficiently smooth [61], and the performance NPF tends to be more sensitive to the choice of parameters e.g. the pseudo-time step size. However, once properly tuned, NPF offers a lower computational cost and slightly better performance than ZPF. Therefore, NPF is also implemented in this work. It is worth noting that, other ideas are emerging to address the issues in NPF. For example, stochastic particle flow based on Langevin diffusion, has been proposed in [62], and the Gromovs method has been used to reduce sensitivity to parameter choice of NPF in [63]. These new developments make NPF an increasingly attractive choice.

C. AV-SMC-PHD filter

The SMC-PHD filter was recently used in [5] for multi-speaker tracking, based on the fusion of audio-visual information. The audio information used is the DOA, e.g. the approximate direction of the speakers with respect to a microphone array, which can be detected by e.g. the sam-spare-mean (SSM) method [64], which is a joint detection and localization method, where the space is divided into sectors, with each sector corresponding to a certain direction, and the active sources are searched over these sectors in terms of the sound energy presented. Color histogram has been used as the visual information [5].

To fuse the audio-visual information, the surviving particles are relocated around the DOA lines which are drawn from the centre of the microphone array to points in the image frame estimated by the projection of DOAs to a 2D image plane [65]. The visual information is used to estimate the likelihood function and to update the particle weights by Eq. (5).

More specifically, the distance for particle movement d_k^i is calculated as:

$$d_k^i = \frac{\acute{d}_k^i}{\sum_{i=i}^{N_{k-1}} \acute{d}_k^i} \acute{d}_k^i \tag{13}$$

where $\acute{d}_k^i$ is the perpendicular Euclidean distance from the particles to the DOA line [5], assuming the DOA line is available at time k. Then, the surviving particles $\{\boldsymbol{m}_{k|k-1}^i\}_{i=1}^{N_{k-1}}$ are relocated to near the DOA line [5] in terms of d_k^i as:

$$\boldsymbol{m}_{k|k-1}^i \Leftarrow \boldsymbol{m}_{k|k-1}^i + \boldsymbol{o}_k d_k^i \tag{14}$$

where $\boldsymbol{o}_k = [\cos(\theta_k^o), \sin(\theta_k^o), 0, 0]^T$ and θ_k^o is the corresponding oth DOA angle. The visual information is used to derive the likelihood which is then used to update the weights ω_k^i. The pseudo code of baseline AV-SMC-PHD filter [5] is given in Algorithm 1, where T is the length of a frame.

Algorithm 1 AV-SMC-PHD Filter

```
Input: $\{\boldsymbol{m}_{k-1}^i, \omega_{k-1}^i\}_{i=1}^{N_{k-1}}$, $N_B$, $\boldsymbol{Z}_k$, $k$ and DOA lines.
Output: $\{\tilde{\boldsymbol{m}}_k^j, \tilde{\omega}_k^j\}_{j=1}^{\tilde{N}_k}$, and $\{\boldsymbol{m}_k^i, \omega_k^i\}_{i=1}^{N_k}$.
Initialize: $\boldsymbol{\tau}_k$, $q_k$, $\phi_{k|k-1}$, $p_k$, $\gamma_k$, $\kappa_k$, $P_{D,k}$, $\mathbf{F}_{\tilde{\boldsymbol{m}}}$, $\mathbf{F}_{\boldsymbol{z}}$, $T$ and speaker histograms.
Run:
Step 1: Propagation step
Propagate surviving particles $\{\boldsymbol{m}_{k|k-1}^i\}_{i=1}^{N_{k-1}}$.
Step 2: Particle birth and relocation step
if DOA lines exist then Calculate $\boldsymbol{d}_k$ by Eq. (13).
    Concentrate particles around the DOA line by Eq. (14).
    if new speaker then 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 the
        DOA line Eq. (4).
Calculate $\{\omega_{k|k-1}^i\}_{i=1}^{N_{k-1}}$ by Eq. (3).
Step 3: Prediction step
$\{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=1}^{N_k} = \{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=1}^{N_{k-1}} \cup \{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=N_{k-1}+1}^{N_{k-1}+N_B}$.
Step 4: Update step
Estimate colour likelihood.
(Optional) Update the states and the weights of the particles by the particle flow using Algorithm 3 and 2.
Step 5: Estimation step
Update $\{\omega_{k|k-1}^i\}_{i=1}^{N_k}$ to obtain $\{\omega_k^i\}_{i=1}^{N_k}$ by Eq. (5) and calculate $\tilde{N}_k = \sum_{i=1}^{N_k} \omega_k^i$.
Set $\{\boldsymbol{m}_k^i\}_{i=1}^{N_k}$ as $\{\boldsymbol{m}_{k|k-1}^i\}_{i=1}^{N_k}$.
Get $\{\tilde{\boldsymbol{m}}_k^j, \tilde{\omega}_k^j\}_{j=1}^{\tilde{N}_k}$ by the k-means method or MEAP
if ESS $< N_k/2$ then (Optional) Resample $\{\boldsymbol{m}_k^i, \omega_k^i\}_{i=1}^{N_k}$.
```

The AV-SMC-PHD filter, however, suffers from the weight degeneracy problem, as illustrated in Fig. 1(a), where ten particles are shown in blue solid circles, with their weights indicated by the size of the circles, and the prior density and the likelihood as the red and green solid lines, respectively. At the top of the figure, the propagated particles are given. After the relocation step, most of the particles converge to the area around the peak of the prior density. In the prediction step, the weights of the particles are adjusted. According to the Bayes' theorem, the posterior density is proportional to the multiplication of the prior density with the likelihood density, and its estimation becomes less accurate when no particles are around the posterior density. As a result, only a small number of particles have high weights after the update step.

In addition, the color histograms of the speakers' faces were used as measurements for particle update. The weights of the particles may decrease sharply when the speakers do not face the camera, and this can lead to unreliable tracking results.

Finally, the particles are relocated based on the DOA lines via Eq. (13) and Eq. (14), and thus they could be migrated from the undetected speaker to the clutter or other speakers, when the DOA estimation is corrupted by background clutter.

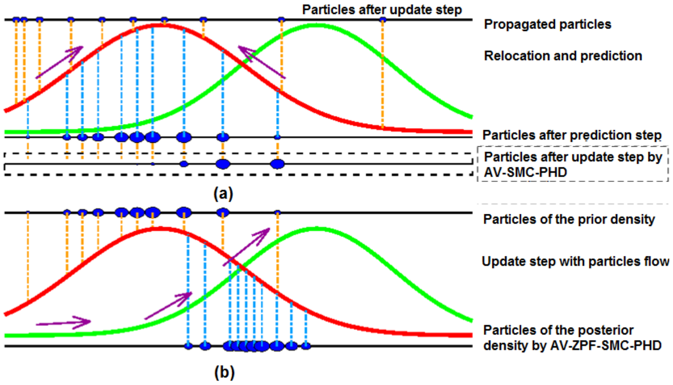

Particles after update step
Propagated particles
Relocation and prediction
Particles after prediction step
Particles after update step by AV-SMC-PHD
(a)
Particles of the prior density
Update step with particles flow
Particles of the posterior density by AV-ZPF-SMC-PHD
(b)

Fig. 1: Illustration of (a) the weight degeneracy problem and (b) the particle flow process. The particles are shown as the blue circles whose sizes indicate their weights. The prior density and likelihood density are represented by the red dashed and green solid lines, respectively.

III. Proposed AV particle flow SMC-PHD Filter

To address the above problems, we propose an AV particle flow SMC-PHD filter, where the particle flow is used for particle migration before the update step of the AV-SMC-PHD filter (i.e. Step 4 of Algorithm 1). Furthermore, the measurements $\boldsymbol{z}_k^r$ used in particle flow Eq. (9) and update step Eq. (5) are replaced by the candidate speaker states, calculated in terms of DOA and color histograms. The speaker states in the previous frames are also used to relocate surviving particles with DOA at Step 2 of Algorithm 1. This allows performance benchmarking with the baseline method in [5], to show the advantage of our proposed method. However, to show the flexibility of the proposed method, we have also considered other type of visual measurements such as those obtained by a state-of-the-art deep learning based face detector, in our experiments.

A. Particle flow for AV-SMC-PHD filter

We consider both ZPF and NPF in our proposed algorithms. As the number of speakers and the labels of the particles are unknown and time-varying, we assume that the number of particle flows is identical to the number of candidate speakers $\tilde{N}_{k|k-1}$. The particles are clustered for the particle flow based on candidate speaker states $\tilde{\boldsymbol{m}}^{j}_{k|k-1}$, as discussed in Section III-B. However, in practice, some particles are created due to clutter and noise. We assume that each flow will only be influenced by the particles in the neighborhood of $\tilde{\boldsymbol{m}}^{j}_{k|k-1}$ within certain distance $\xi_{\boldsymbol{f}}$,

$$\left\| \boldsymbol{m}^{i}_{k|k-1} - \tilde{\boldsymbol{m}}^{j}_{k|k-1} \right\|_2 < \xi_{\boldsymbol{f}} \tag{15}$$

The flow will be applied to the set of particles, $\{\boldsymbol{m}^{i}_{k|k-1}, \omega^{i}_{k|k-1}\}_{i \in \boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^{j}_{k|k-1})}$, where $\boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^{j}_{k|k-1})$ is a subset of $[1, \cdots, N_k]$, determined via (15), in terms of the threshold $\xi_{\boldsymbol{f}}$. A high $\xi_{\boldsymbol{f}}$ may lead to an inaccurate estimation of the variance of the speaker states, while a low $\xi_{\boldsymbol{f}}$ may result in an insufficient number of particles for estimating the variance of the target state. In practice, $\xi_{\boldsymbol{f}}$ is set according to the variance of the measurement noise, estimated empirically in our experiments. The mean state $\bar{\boldsymbol{m}}_{k|k-1}$ of this particle set is given by:

$$\bar{\boldsymbol{m}}_{k|k-1} = \frac{\sum_{i \in \boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^{j}_{k|k-1})} \left(\omega^{i}_{k|k-1} \mathbf{F}_{\boldsymbol{m}} \left(\boldsymbol{m}^{i}_{k-1} \right) \right)}{\sum_{i \in \boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^{j}_{k|k-1})} \omega^{i}_{k|k-1}} \tag{16}$$

The covariance matrix $\boldsymbol{P}_{k|k-1}$ of this particle set is given by:

$$\boldsymbol{P}_{k|k-1} = \frac{\sum_{i \in \boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^{j}_{k|k-1})} \left(\omega^{i}_{k|k-1} \boldsymbol{e}(\boldsymbol{m}^{i}_{k-1}) \boldsymbol{e}(\boldsymbol{m}^{i}_{k-1})^T \right)}{\sum_{i \in \boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^{j}_{k|k-1})} \omega^{i}_{k|k-1}} \tag{17}$$

where

$$\boldsymbol{e}(\boldsymbol{m}^{i}_{k-1}) = \mathbf{F}_{\boldsymbol{m}} \left(\boldsymbol{m}^{i}_{k-1} \right) - \bar{\boldsymbol{m}}_{k|k-1} \tag{18}$$

For the ith particle in the subset $\boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^{j}_{k|k-1})$, $\boldsymbol{P}^{i}_{k|k-1}$ and $\bar{\boldsymbol{m}}^{i}_{k|k-1}$ in Eq. (8) and Eq. (9) are set as $\boldsymbol{P}_{k|k-1}$ and $\bar{\boldsymbol{m}}_{k|k-1}$, respectively.

1) Zero Diffusion Particle Flow: For ZPF, the particle flow is calculated by Eqs. (7)-(10) and the measurement model $\boldsymbol{H}^{i}_{k}$ for the ith particle is calculated as,

$$\boldsymbol{H}^{i}_{k} = \left. \frac{\partial \mathbf{F}_{\boldsymbol{z}}(\boldsymbol{m}^{i}_{k}, \varsigma_k)}{\partial \boldsymbol{m}^{i}_{k}} \right|_{\boldsymbol{m}^{i}_{k|k-1}} \tag{19}$$

In this paper, as the measurement is replaced by the candidate speaker state, $\boldsymbol{H}^{i}_{k} \in \mathbb{R}^{4 \times 4}$ is taken as an identity matrix.

The flow $\boldsymbol{f}^{i}_{k,\lambda}$ of the particle set $\{\boldsymbol{m}^{i}_{k|k-1}\}_{i \in \boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^{j}_{k|k-1})}$ is calculated via Eq. (7) and applied for migrating the particles. $\boldsymbol{A}^{i}_{k}$ and $\boldsymbol{b}^{i}_{k}$ are derived according to Eq. (8) and Eq. (9). Since the particle states have been updated, the weights of $\{\boldsymbol{m}^{i}_{k|k-1}\}_{i \in \boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^{j}_{k|k-1})}$ may have a poor representation of the prior distribution. As the weights are inversely proportional to the proposal distribution as Eq. (3), the weights should be adjusted by:

$$\omega^{i}_{k|k-1} \Leftarrow \frac{q_k(\boldsymbol{m}^{i}_{k|k-1} | \tilde{\boldsymbol{m}}^{j}_{k|k-1}) |\mathrm{det}(\boldsymbol{I} + \Delta\lambda \boldsymbol{A}^{i}_{k}(\lambda))|}{q_k(\boldsymbol{m}^{i}_{k|k-1} + \triangle\lambda \boldsymbol{f}^{i}_{k,\lambda} | \tilde{\boldsymbol{m}}^{j}_{k|k-1})} \omega^{i}_{k|k-1} \tag{20}$$

where the proposal distribution $q_k(\boldsymbol{m}^{i}_{k|k-1} | \tilde{\boldsymbol{m}}^{j}_{k|k-1}) \propto \mathcal{N}(\tilde{\boldsymbol{m}}^{j}_{k|k-1}, \boldsymbol{\Sigma}^2_q)$, $\boldsymbol{\Sigma}_q$ is the covariance of the proposal distribution [66], [67], and det is a determinant. $\triangle\lambda$ is the step of the pseudo time, and the choice of this parameter dictates a trade-off between the computational cost for calculating the particle flow, and the accuracy for estimating the posterior probability. Although det$(\boldsymbol{I} + \Delta\lambda \boldsymbol{A}^{i}_{k}(\lambda))$ is a constant for the particles in the same set $\boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^{j}_{k|k-1})$, it may be different for particles in different set $\boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^{j}_{k|k-1})$ and thus improves the estimate of the particle weights [43].

2) Non-zero Diffusion Particle Flow: For NPF, the particle flow $\boldsymbol{f}^{i}_{k,\lambda}$ is calculated by Eqs. (11)-(12), where $\boldsymbol{\nabla} \log h^{i}_{k}$ and $\boldsymbol{\nabla}^2 \log h^{i}_{k}$ are calculated as

$$\boldsymbol{\nabla} \log h^{i}_{k} = \frac{\boldsymbol{\nabla} h^{i}_{k}}{h^{i}_{k}} \tag{21}$$

$$\boldsymbol{\nabla}^2 \log h^{i}_{k} = \frac{\boldsymbol{\nabla}^2 h^{i}_{k}}{h^{i}_{k}} - (\boldsymbol{\nabla} \log h^{i}_{k})(\boldsymbol{\nabla} \log h^{i}_{k})^T \tag{22}$$

When h^{i}_{k} is Gaussian, we have $\boldsymbol{\nabla} \log h^{i}_{k} = \boldsymbol{P}^{-1}_{k|k-1}(\boldsymbol{m}^{i}_{k|k-1} - \bar{\boldsymbol{m}}_{k|k-1})$ and $\boldsymbol{\nabla}^2 \log h^{i}_{k} = \boldsymbol{P}^{-1}_{k|k-1}$[45]. The weight is adjusted by:

$$\omega^{i}_{k|k-1} \Leftarrow \frac{q_k(\boldsymbol{m}^{i}_{k|k-1} | \tilde{\boldsymbol{m}}^{j}_{k|k-1}) |\mathrm{det}(\boldsymbol{I} + \boldsymbol{\nabla} \boldsymbol{f}^{i}_{k,\lambda})|}{q_k(\boldsymbol{m}^{i}_{k|k-1} + \triangle\lambda \boldsymbol{f}^{i}_{k,\lambda} | \tilde{\boldsymbol{m}}^{j}_{k|k-1})} \omega^{i}_{k|k-1} \tag{23}$$

where

$$\boldsymbol{\nabla} \boldsymbol{f}^{i}_{k,\lambda} = \begin{cases} \frac{\boldsymbol{f}^{i}_{k,\lambda} - \boldsymbol{f}^{i}_{k,\lambda-\triangle\lambda}}{\triangle\lambda \boldsymbol{f}^{i}_{k,\lambda}} & \lambda \neq 0 \\ 0 & \lambda = 0 \end{cases} \tag{24}$$

The particle state $\boldsymbol{m}^{i}_{k|k-1}$ is updated by:

$$\boldsymbol{m}^{i}_{k|k-1} \Leftarrow \boldsymbol{m}^{i}_{k|k-1} + \triangle\lambda \boldsymbol{f}^{i}_{k,\lambda} \tag{25}$$

The pseudo code of particle flow is given in Algorithm 2, where $\{\tilde{\boldsymbol{m}}^{j}_{k|k-1}\}_{j=1}^{\tilde{N}_{k|k-1}}$ is the candidate speaker state which replaces $\boldsymbol{z}^{r}_{k}$ in Eq. (9). After applying Algorithm 2, $\boldsymbol{\omega}^{i}_{k|k-1}$ will be updated to $\boldsymbol{\omega}^{i}_{k}$ by Eq. (5). The particle flow step is also illustrated in Fig. 1(b). The particles are moved towards the peak of the likelihood. As a result, most of the particles are localized between the two peaks of the prior density (red line) and the likelihood density (green line). This means that the shifted particles provide an improved local characterization of the posterior density.

B. Candidate Speaker States

In this subsection, we explain how the candidate speaker states are estimated using audio-visual information e.g. DOA and color histograms. It is worth noting that our proposed tracking framework is flexible and can be adapted easily to accommodate other audio-visual information, such as face detector [10], as considered in our experiments. The DOA information can be obtained by either a circular array (as in our work) or a linear array. As the DOAs are determined by the relative delay between the pairs of microphone signals [5], it only shows the approximate direction θ^{o}_{k} of the sound sources with respect to the microphones. In practice, the

Algorithm 2 Particle flow for the AV-SMC-PHD filter

```
**Input:** $\{\boldsymbol{m}^i_{k|k-1}, \omega^i_{k|k-1}\}_{i=1}^{N_k}$ and $\{\tilde{\boldsymbol{m}}^j_{k|k-1}\}_{j=1}^{\tilde{N}_{k|k-1}}$.
**Output:** $\{\boldsymbol{m}^i_{k|k-1}, \omega^i_{k|k-1}\}_{i=1}^{N_k}$.
**Initialize:** $\xi_{\boldsymbol{f}}$, $\triangle\lambda$, $N_\lambda$, $\boldsymbol{R}$, $\boldsymbol{e}$, $\mathbf{F}_{\boldsymbol{z}}$, $\varsigma_k$ and $\boldsymbol{\Sigma}_q$.
**Run:**
**for** each $\tilde{\boldsymbol{m}}^j_{k|k-1}$ **do**
  Select particle set $\boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^j_{k|k-1})$ according to Eq. (15).
  Calculate $\bar{\boldsymbol{m}}_{k|k-1}$ and $\boldsymbol{P}_{k|k-1}$ by (16) and Eq. (17), respectively.
  Set $\bar{\boldsymbol{m}}^i_{k|k-1} = \bar{\boldsymbol{m}}_{k|k-1}$ and $\boldsymbol{P}^i_{k|k-1} = \boldsymbol{P}_{k|k-1}$.
  **for** $i \in \boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}^j_{k|k-1})$ **do**
    **for** $\lambda \in [0, \triangle\lambda, 2\triangle\lambda, \cdots, N_\lambda\triangle\lambda]$ **do**
      **if** Zero diffusion particle flow **then**
        Evaluate flow $\boldsymbol{f}^i_{k,\lambda}$ by Eqs. (7)-(10).
        Update the particle weights by Eq. (20).
      **if** Non-zero diffusion particle flow **then**
        Evaluate flow $\boldsymbol{f}^i_{k,\lambda}$ by Eqs. (11)-(12).
        Update the particle weights by Eq. (23).
      Update $\boldsymbol{m}^i_{k|k-1}$ by Eq. (25).
```

rectangular coordinate $[x^o_k, y^o_k]$ of $\mathring{\boldsymbol{m}}^o_k$ can be transformed to polar coordinate $[r^o_k, \theta^o_k]$, where r^o_k is the Euclidean distance from the nearby speaker state at the previous frame to the microphone position,

$$r^o_k = \left\| [\tilde{x}^{\hat{j}}_{k-1}, \tilde{y}^{\hat{j}}_{k-1}]^T - [x_{mic}, y_{mic}]^T \right\|_2 \tag{26}$$

where

$$\hat{j} = \underset{j}{\operatorname{argmin}} \left\| \frac{\tilde{y}^j_{k-1} - y_{mic} - (\tilde{x}^j_{k-1} - x_{mic}) \tan\theta^o_k}{\tilde{x}^j_{k-1} - x_{mic} - (\tilde{y}^j_{k-1} - y_{mic}) \tan\theta^o_k} \right\|_1 \tag{27}$$

where $\|.\|_1$ and $\|.\|_2$ are the L_1 and L_2 norm, respectively. $[\tilde{x}^j_{k-1}, \tilde{y}^j_{k-1}]^T$ and $[x_{mic}, y_{mic}]^T$ are the positions of the jth speaker state at the previous frame $\tilde{\boldsymbol{m}}^j_{k-1}$ and the state of microphone array $\boldsymbol{m}_{mic}$, respectively. They both correspond to the center of the microphone array. Here, the aim is to find the index of the speaker, i.e. $\hat{j}$, which gives the minimum tangent value of the angle between the DOA θ^o_k and the direction of each target speaker. Then $\mathring{\boldsymbol{m}}^o_k$ is given by:

$$\mathring{\boldsymbol{m}}^o_k = [r^o_k \cos\theta^o_k + x_{mic}, r^o_k \sin\theta^o_k + y_{mic}, 0, 0]^T \tag{28}$$

The weight $\mathring{\omega}^o_k$ of $\mathring{\boldsymbol{m}}^o_k$ is given by,

$$\mathring{\omega}^o_k = \mathcal{N}(\theta^o_k | \text{arcsin}(\frac{\tilde{y}^j_{k-1} - y_{mic}}{r^o_k}), \sigma^2_o) \tag{29}$$

where σ^2_o is the variance of the DOA angle distribution. The candidate speaker states $\{\tilde{\boldsymbol{m}}^j_{k|k-1}\}_{j=1}^{\tilde{N}_{k|k-1}}$ can be calculated as

$$\tilde{\boldsymbol{m}}^j_{k|k-1} = \begin{cases} \frac{\mathring{\omega}^{\hat{o}}_k \mathring{\boldsymbol{m}}^{\hat{o}}_k + \breve{\omega}^u_k \breve{\boldsymbol{m}}^u_k}{\mathring{\omega}^{\hat{o}}_k + \breve{\omega}^u_k}, & \text{if } d^{u,\hat{o}}_m \le \xi_{\boldsymbol{m}} \\ \breve{\boldsymbol{m}}^u_k, & \text{if } d^{u,\hat{o}}_m > \xi_{\boldsymbol{m}} \end{cases} \tag{30}$$

where

$$d^{u,\hat{o}}_m = \left\| [\mathring{x}^{\hat{o}}_k, \mathring{y}^{\hat{o}}_k]^T - [\breve{x}^u_k, \breve{y}^u_k]^T \right\|_2 \tag{31}$$

where $\breve{\omega}^u_k$ is the weight for $\breve{\boldsymbol{m}}^u_k$, which can be obtained by CAMShift [9] or face detector [10]. $[\breve{x}^u_k, \breve{y}^u_k]^T$ is the position information taken from $\breve{\boldsymbol{m}}^u_k$. As the association between $\mathring{\boldsymbol{m}}^{\hat{o}}_k$ and $\breve{\boldsymbol{m}}^u_k$ is unknown and time-varying, we assume $\tilde{N}_{k|k-1}$ is equal to the number of the speakers detected, and o is the index of the DOA line that is closest to $\breve{\boldsymbol{m}}^u_k$.

$$\hat{o} = \underset{o}{\operatorname{argmin}}\, d^{u,o}_m \tag{32}$$

In practice, $\mathring{\boldsymbol{m}}^{\hat{o}}_k$ and $\breve{\boldsymbol{m}}^u_k$ may represent different speaker states. To address this issue, the distance $d^{u,\hat{o}}_m$ is compared with a threshold value $\xi_{\boldsymbol{m}}$. If $d^{u,\hat{o}}_m \le \xi_{\boldsymbol{m}}$, $\tilde{\boldsymbol{m}}^j_{k|k-1}$ is estimated in terms of the DOA and color histograms. When the speakers go out of the view of the camera or are visually occluded, $\boldsymbol{m}^u_k$ will become inaccurate. In this case, the DOA lines near the speaker will be used to calculate the candidate speaker states which are then used to calculate the likelihood as:

$$h^{i,j}_k \propto \mathcal{N}(\tilde{\boldsymbol{m}}^i_{k|k-1} - \tilde{\boldsymbol{m}}^j_{k|k-1} | \mathbf{0}, \boldsymbol{\Sigma}_h) \tag{33}$$

where $\boldsymbol{\Sigma}_h$ is the covariance of the likelihood and $h^{i,r}_k$ in Eq. (5) is replaced by $h^{i,j}_k$. As a result, the particle weights are likely to retain high values even when the speakers do not face the camera. If $d^{u,\hat{o}}_m > \xi_{\boldsymbol{m}}$, $\tilde{\boldsymbol{m}}^j_{k|k-1}$ is only estimated by the uth color histograms.

The pseudo code for calculating the candidate speaker state is given in Algorithm 3, which, together with Algorithm 2, can be plugged into Step 4 of Algorithm 1.

Algorithm 3 Candidate Speaker States

```
**Input:** DOA lines, reference color histograms, $\{\tilde{\boldsymbol{m}}^j_{k-1}\}_{j=1}^{\tilde{N}_{k-1}}$ and $\{\boldsymbol{m}^i_{k|k-1}, \omega^i_{k|k-1}\}_{i=1}^{N_k}$.
**Output:** $\{\tilde{\boldsymbol{m}}^j_{k|k-1}\}_{j=1}^{\tilde{N}_{k|k-1}}$.
**Initialize:** $\boldsymbol{m}_{mic}$, $\xi_{\boldsymbol{m}}$ and $\sigma_o$.
**Run:**
**for** each DOA line indexed by $o$ **do**
  Calculate $\theta^o_k$ from the DOA line [5].
  Select the nearby speaker $\tilde{\boldsymbol{m}}^{\hat{j}_{k-1}}$ as Eq. (27).
  Calculate $r^o_k$ and $\mathring{\omega}^o_k$ as Eq. (26) and Eq. (29), respectively.
  Calculate $\mathring{\boldsymbol{m}}^o_k$ as Eq. (28).
$j = 0$
**for** each reference histogram indexed by $u$ **do**
  **if** the $u$th reference histogram is detected **then** $j = j + 1$.
    Calculate $\breve{\boldsymbol{m}}^u_k$ and $\breve{\omega}^u_k$ by CAMShift [9].
    Select the nearby $\mathring{\boldsymbol{m}}^{\hat{o}}_k$ as Eq. (32).
    Calculate $\tilde{\boldsymbol{m}}^j_{k|k-1}$ as Eq. (30).
$\tilde{N}_{k|k-1} = j$.
```

C. Relocating particles

Due to the presence of noise and clutter in acoustic measurements, the DOAs estimated are not always reliable. To address this issue, we also consider the speaker state $\{\tilde{\boldsymbol{m}}^j_{k-1}\}_{j=1}^{\tilde{N}_{k-1}}$ at the previous time frame $k-1$.

After calculating d_k^i by using Eq. (13), the movement distance d_k^i is updated as

$$d_k^i \Leftarrow \begin{cases} (1-\tilde{\omega}_k^{\hat{j}})d_k^i + \tilde{\omega}_k^{\hat{j}} \left\| \triangle \tilde{\boldsymbol{m}}_{k|k-1}^i \right\|_2 & d_k^i \leq \xi_d \\ \tilde{\omega}_k^{\hat{j}} \left\| \triangle \tilde{\boldsymbol{m}}_{k|k-1}^i \right\|_2 & d_k^i > \xi_d \end{cases} \tag{34}$$

where

$$\triangle \tilde{\boldsymbol{m}}_{k|k-1}^i = \mathbf{F}_{\tilde{\boldsymbol{m}}}(\tilde{\boldsymbol{m}}_{k-1}^{\hat{j}}, \boldsymbol{\tau}_k) - \boldsymbol{m}_{k|k-1}^i \tag{35}$$

$$\hat{j} = \underset{j}{\operatorname{argmin}} \left\| \mathbf{F}_{\tilde{\boldsymbol{m}}}(\tilde{\boldsymbol{m}}_{k-1}^{j}, \boldsymbol{\tau}_k) - \boldsymbol{m}_{k|k-1}^i \right\|_2 \tag{36}$$

where $\hat{j} \in \{1, ..., \tilde{N}_{k-1}\}$ is the index of the speaker closest to the ith particle. The threshold ξ_d is used to control the movement distance d_k^i, in order to reduce the effect of noise in DOA estimate. When the DOA estimates are noisy, we relocate the particles in terms of the motion model, otherwise, in terms of the DOA and the speaker states in the previous frames. In our work, ξ_d is set empirically according to the variance of the DOA estimates. As the particles are located near the DOA lines, the relocation step is applied only when $U_k \neq \tilde{N}_{k-1}$, where U_k and $\tilde{N}_{k-1}$ are the number of audio measurements at time k and the number of speakers at time $k-1$, respectively.

The DOAs are also applied to detect new speakers, i.e. whether Step 2 in Algorithm 1 is needed. Comparing the number of visual measurements O_k with $\tilde{N}_{k-1}$, the PHD filter is able to identify the appearance and disappearance of speakers. If $O_k = \tilde{N}_{k-1}$, the number of speakers remains unchanged. If $O_k < \tilde{N}_{k-1}$, the speakers may walk out of the camera view, or be occluded by other speakers. If $O_k > \tilde{N}_{k-1}$, new speakers may appear in the scene, and hence new born particles are created.

IV. Experimental Evaluations

The proposed algorithms are evaluated using real AV data. First, we briefly discuss the datasets, baseline algorithms, performance metrics and the parameter set up. Then we show the improvement achieved by the particle flow, the candidate speaker states and the novel localization method. Finally, we compare the proposed methods with several recent baselines.

A. Datasets and Baselines

Several audio-visual datasets are publicly available, such as the AV16.3 [46], AVDIAR [47], AVTRACK-1 [68], AVASM [69], AMI [70], CLEAR [48], MVAD [71] and SPEVI [72]. We have considered our requirements when choosing the datasets. For example, the calibration information should be provided for the projection of the audio information from the physical space to the image plane. In addition, the dataset should contain some challenging situations, e.g, the number of speakers changes and some speakers are occluded. For these reasons, we have chosen AV16.3, AVDIAR and CLEAR datasets in our evaluations.

The AV16.3 [46] consists of real-world data with both audio and video sequences. It provides the calibration information of the cameras to map the audio information from the physical space to the image plane. AV16.3 includes the occlusion as a challenging scenario, and consists of sequences where the speakers are walking and speaking at the same time. The video and audio signals are recorded by three calibrated video cameras at 25 Hz and two circular eight-element microphone arrays at 16 kHz, respectively. The audio and video streams are synchronized before running the algorithms. The size of each image frame is 288×360. All algorithms are tested with all three different camera angles of five sequences: Sequences 1, 24, 25, 30 and 45, which correspond to the cases of one to three speakers and are the most challenging sequences in term of movements of the speakers and occlusions.

Different from the AV16.3 dataset, the speakers in the AVDIAR dataset [47] talk one by one. There are six microphones mounted on Sennheiser Triaxial MKE 2002. Two of them are on the left and right ears and the other four are on each side of the head. However, since the details of the microphone positions are not provided, only the microphones on the left and right ears are considered. Another issue is that the calibration information of the cameras is not available. The AVDIAR provides training data to learn a mapping as in [47]. This dataset includes 23 sequences. Each image frame is of 1920×1200 pixels. The audio and video were recorded at 48 kHz and 25 Hz, respectively, which were synchronized by an external trigger controlled by software. There are 12 different participants and up to 4 people are recorded in each sequence.

AVTRACK-1 [68] and AVASM [69] are provided by the same institution as for AVDIAR. However, they are less challenging than AVDIAR. AMI and MVAD, which are designed for speaker diarization, are not used in our tests since the speakers are mostly static or with small movements. In SPEVI [72], audio signals were recorded with linear microphone arrays. Since the calibration information and training set are not available, this dataset is also not chosen. The CLEAR dataset is chosen for our experiments since it has the largest number of speakers among these datasets. Although our proposed algorithms could be used in other scenarios such as sport-video analysis and smart surveillance systems, due to the lack of suitable datasets, such scenarios are not considered here.

Several baselines are considered for benchmarking our proposed algorithms, including AV-PF-PF [43], AV-ZPF-GPF-PHD [29], SAVMS-SMC-PHD [5], auxiliary SMC-PHD filter [23], baseline AV-SMC-PHD filter [5] and the filter proposed in our previous work [7]. For convenience, the AV-ZPF-SMC-PHD, AV-NPF-SMC-PHD, SAVMS-SMC-PHD, AV-PF-PF, AV-ZPF-GPF-PHD, AV-SMC-PHD and AV auxiliary SMC-PHD filters are abbreviated as ZPF, NPF, SMS, PPF, GPF, SMC and ASMC respectively. The GLMB method [19] was not considered since it was used only for audio tracking, and did not address the weight degeneracy problem.

B. Performance metrics

We use the Optimal Sub-pattern Assignment (OSPA), ESS, and distance between particles and ground truth speak state as performance metrics.

The OSPA [73] is defined as,

$$\mathrm{OSPA}(\{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k}, \{\tilde{\mathfrak{m}}_k^{\tilde{j}}\}_{\tilde{j}=1}^{\tilde{\mathfrak{N}}_k}) = \sqrt[a]{\frac{\min\limits_{\pi \in \Pi_{\tilde{\mathfrak{N}}_k, \tilde{N}_k}} \sum_{j=1}^{\tilde{N}_k} \overline{d}^{(c)}(\tilde{\boldsymbol{m}}_k^j, \tilde{\mathfrak{m}}_k^{\pi(j)})^a + c^a(\tilde{\mathfrak{N}}_k - \tilde{N}_k)}{\tilde{\mathfrak{N}}_k}} \quad (37)$$

where $\{\tilde{\mathfrak{m}}_k^1, ..., \tilde{\mathfrak{m}}_k^{\tilde{\mathfrak{N}}_k}\}$ is the set of ground truth speaker states, and $\{\tilde{\boldsymbol{m}}_k^1, ..., \tilde{\boldsymbol{m}}_k^{\tilde{N}_k}\}$ is the set of the estimated speaker states. $\Pi_{\tilde{\mathfrak{N}}_k, \tilde{N}_k}$ is the set of maps $\pi : 1, ..., \tilde{N}_k \rightarrow 1, ..., \tilde{\mathfrak{N}}_k$. Here the state cardinality estimation $\tilde{N}_k$ may not be the same as the ground truth $\tilde{\mathfrak{N}}_k$. The OSPA error given in Eq. (37) is for $\tilde{N}_k \leq \tilde{\mathfrak{N}}_k$. If $\tilde{\mathfrak{N}}_k < \tilde{N}_k$, then $\mathrm{OSPA}(\{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k}), \{\tilde{\mathfrak{m}}_k^{\tilde{j}}\}_{\tilde{j}=1}^{\tilde{\mathfrak{N}}_k}) = \mathrm{OSPA}(\{\tilde{\mathfrak{m}}_k^{\tilde{j}}\}_{\tilde{j}=1}^{\tilde{\mathfrak{N}}_k}, \{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k}))$. The function $\bar{d}^{(c)}(\cdot)$ is defined as $\min(c, \bar{d}(\cdot))$ where c is the cut-off value, which determines the relative weighting of the penalties for the cardinality and localization errors and a is the metric order which determines the sensitivity to outliers. A lower OSPA implies a better performance.

ESS is applied widely to evaluate the severity of weight degeneracy problem [43], [20], [32], which is given by

$$\mathrm{ESS} = \frac{(\sum_{i=1}^{N_k} \omega_k^i)^2}{\sum_{i=1}^{N_k} (\omega_k^i)^2} \quad (38)$$

When ESS is small, e.g. ESS $< N_k/2$, the resampling step is performed with the uniform weights. When ESS is high, more particles are used to estimate the posterior density with an increased accuracy.

As the label information for each particle is unavailable, the minimal distance $d_m(\boldsymbol{m}_{k|k-1}^i)$ between each particle and speaker is used:

$$d_m(\boldsymbol{m}_{k|k-1}^i) = \underset{\tilde{j} \in \mathfrak{N}_k}{min} \left\| \boldsymbol{m}_{k|k-1}^i - \tilde{\mathfrak{m}}_{k-1}^{\tilde{j}} \right\|_2 \quad (39)$$

C. Parameter settings

In this subsection, we discuss the setting of five important parameters, i.e. pseudo time λ, the number of particles N_k, and three thresholds $\xi_{\boldsymbol{f}}$, $\xi_{\boldsymbol{m}}$, and $\xi_{\boldsymbol{d}}$. We only show the experiments based on ZPF, as these are similar to NPF. Other parameters are given as in the baseline method SMC [5]. The parameters used for detecting the DOAs are set the same as in [74]. The initial distributions of the particles are randomly sampled in the tracking area. If the particles move out of the tracking area, we will reject and resample them. Resampling is performed when ESS is smaller than $N/2$. The order parameter a in OSPA is set to 2. These parameters are chosen empirically based on our earlier studies [7], [20]. All experiments are run on a computer with Intel i7-3770 CPU with a clock frequency of 3.40 GHz and 8G RAM. Each experiment is repeated 50 times, and the average results are presented.

1) Pseudo time: The pseudo time λ in the particle flow is increased incrementally from 0 to 1, with a step size $\Delta\lambda$, which is either fixed as in [75] or varied as in [43], [32]. Six situations are considered. Here, we have tested three fixed step sizes, i.e. $\Delta\lambda = 0.1$, 0.01 and 0.001, and three varied: $\Delta\lambda = 0.0385 \times 1.2^{\lambda \times N_\lambda}$, $2.4 \times 10^{-9} \times 1.2^{\lambda \times N_\lambda}$, and $1.3 \times 10^{-80} \times 1.2^{\lambda \times N_\lambda}$. In both cases, the number of steps N_λ is chosen as 10, 100, 1000, respectively. Sequence 01 (camera 1) from the AV16.3 dataset is used since there is only one speaker and it is easy to see the impact of different pseudo times.

TABLE II: Running time (s) and tracking accuracy in OSPA of ZPF versus λ steps.

step type	fixed			varied		
N_λ	10	100	1000	10	100	1000
time (s)	9.03	11.64	23.99	9.05	11.68	23.98
average OSPA	8.82	7.32	7.12	8.54	7.32	7.12

Table II shows the running time and OSPA of ZPF versus the step type and N_λ. It can be seen that a smaller step size leads to a smaller OSPA but with a longer running time. For the case of $N_\lambda = 100$ with a fixed time step size, a good balance between OSPA at about 7.3 and running time at about 11.6s is achieved, therefore, this is used later in our experiments.

2) Number of particles N_k: A large N_k can alleviate the weight degeneracy problem [76], but induce extra computational cost. Here N_k is set from 10 to 1000. Sequence 01 (camera 1) of the AV16.3 dataset is used. During the iterations of the algorithm, if N_k is greater than a preset value, the particles with low weights are removed from the particle set. If N_k is smaller than the preset value, the particles with high weights are duplicated and added into the particle set. The results are shown in Table III. It can be seen that with the increase in the number of particles, OSPA is reduced while the computational cost is increased. When N_k is larger than 50, the OSPA becomes stabilized at approximately 7 and the further improvement is small. For example, compared to the case $N_k = 50$, using $N_k = 500$, an OPSA of only 3.1% lower was achieved, at a cost of nearly ten times computational load. If N_k is smaller than 50, e.g. N_k = 10, OSPA is 37.0471, implying a higher tracking errors. Therefore, $N_k = 50$ is used later in our experiments.

TABLE III: Running time (s) and OSPA of ZPF versus the number of particles.

N_k	10	50	100	500	1000
time (s)	2.51	11.63	22.25	111.95	245.40
average OSPA	37.05	7.32	7.22	7.09	6.84

3) Thresholds $\xi_{\boldsymbol{f}}$, $\xi_{\boldsymbol{m}}$, and $\xi_{\boldsymbol{d}}$: The parameters $\xi_{\boldsymbol{f}}$, $\xi_{\boldsymbol{m}}$, and $\xi_{\boldsymbol{d}}$ are used, respectively, to guide the selection of particles into $\boldsymbol{\Lambda}(\tilde{\boldsymbol{m}}_{k|k-1}^j)$, to obtain the states for the candidate speakers, and to relocate the particles in the relocation steps. We have tested different values for these parameters ranging from 1 to 288 (for the image height at 288). Since these thresholds were used for different purposes, we have chosen different sequences and frames for the tests, i.e. all the frames in Sequence 45 (camera 1) for $\xi_{\boldsymbol{f}}$, frames 500-1000 of Sequence 45 (camera 1) for $\xi_{\boldsymbol{m}}$, and frames 280-500 of Sequence 24 (camera 1) for $\xi_{\boldsymbol{d}}$, all from the AV16.3 dataset.

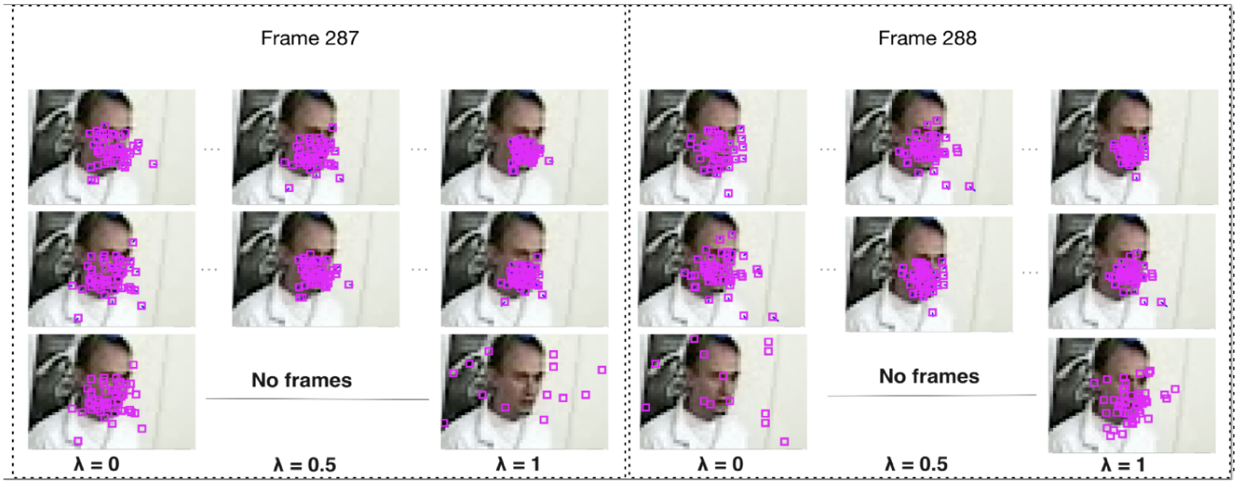


Fig. 2: The motion trails of the particles for the second speaker by NPF, ZPF and SMC. The columns show the results for $\lambda = 0, 0.5, 1$ respectively in the frame 287 and 288 for Sequence 24.

The results are shown in Tables IV. When $\xi_{\boldsymbol{f}}$ is increased to 25, the OSPA is decreased but the running time is increased since more particles are considered in the particle flow. However, with a further increase in $\xi_{\boldsymbol{f}}$, the performance in terms of OSPA starts to degrade, because the particles of the different speakers may be selected into the same set $\boldsymbol{\Lambda}$. When $\xi_{\boldsymbol{m}}$ is increased, OSPA is first decreased from 38.6 to 34.8 and then remains stable. When $\xi_{\boldsymbol{d}}$ is increased, the OSPA is decreased from 22.3 to 20.8 and then remains almost unchanged at 20.8. Therefore, for the AV16.3 dataset, we set $\xi_{\boldsymbol{f}}$, $\xi_{\boldsymbol{m}}$, and $\xi_{\boldsymbol{d}}$ empirically as 25 in our experiments. Note that the running time of our proposed algorithm is not dependent on $\xi_{\boldsymbol{m}}$ and $\xi_{\boldsymbol{d}}$, since only the DOA lines near the particles are considered.

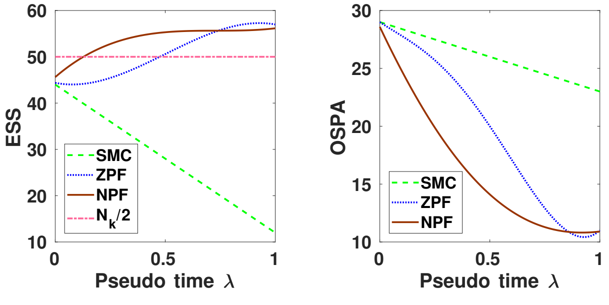
TABLE IV: Running time (s) and OSPA of ZPF versus $\xi_{\boldsymbol{f}}$, $\xi_{\boldsymbol{m}}$ and $\xi_{\boldsymbol{d}}$.

	1	10	25	50	100	288	
$\xi_{\boldsymbol{f}}$	96	286	348	394	493	673	time (s)
	26.5	23.8	19.3	24.8	28.5	36.5	OSPA
$\xi_{\boldsymbol{m}}$	146	146	146	146	146	146	time (s)
	38.6	34.8	31.6	31.6	31.6	31.6	OSPA
$\xi_{\boldsymbol{d}}$	52.1	52.1	52.1	52.1	52.1	52.1	time (s)
	22.3	21.5	20.8	20.8	20.8	20.8	OSPA

D. *Comparison with the baseline methods*

In this subsection, we show the improvement achieved by the particle flow, the novel relocation method and the candidate speaker state. First, we compare between particle flows and SMC. Second, we compare particle flows with candidate speaker state and color histograms, respectively. Each speaker is calculated by 50 particles. Third, we compare between our proposed relocation method with the relocation method in SMC [5]. When a new speaker is detected, 50 particles are created and added into the PHD filter. The parameters for the particle flow are set empirically, i.e. $\triangle\lambda = 0.01$ and $N_\lambda = 100$.

1) Particle flow for weight degeneracy problem: To evaluate the particle flow, ZPF, NPF and SMC are compared. Firstly, we only update the particle states by SMC in the frames 270-286 for Sequence 24 on camera 1. In the frames 0-270, there is no speaker or only one speaker, and the weight degeneracy issue does not usually occur. Therefore, these frames are not used for demonstrating the weight degeneracy effect. Note, however, that we have included the results for all the frames in Section IV-E, with a varying number of speakers from 0 to 3. Until frame 286, there are two speakers and most of the particles can track the speakers. In frame 287, the ESS of SMC is smaller than $N_k/2$ and SMC is encountered with the weight degeneracy problem. Then this particle set is separately updated by ZPF, NPF and the baseline SMC. In these filters, the candidate speaker state is used. In SMC, we assume $h_k^{i,j} \propto \mathcal{N}(\boldsymbol{m}_{k|k-1}^i - \tilde{\boldsymbol{m}}_{k|k-1}^j, \boldsymbol{R})$. $h_k^{i,r}$ and $\boldsymbol{z}_k^r$ in Eq. (5) are replaced by $h_k^{i,j}$ and $\tilde{\boldsymbol{m}}_{k|k-1}^j$. ESS is calculated at each pseudo time step.



(a) (b)

Fig. 3: The ESS and OSPA of ZPF, NPF and SMC in the frame 287 of Sequence 24 (camera 1) changes with respect to λ.

Fig. 2 shows how the particles are modified by these filters from $\lambda = 0$ to $\lambda = 1$. As an example, the first, second and third rows, show the tracking results of NPF, ZPF, and SMC, respectively, for frames 287 and 288. The green lines show the motion trails of the particles. Note, the face images were actually cropped manually from the video signal to visualize the distribution of the particles around the face area, which is very small in the whole image plane. For NPF and ZPF, three figures are shown for $\lambda = 0, 0.5$ and 1, respectively. SMC at the third row is shown as the baseline method, where only the predicted and updated results are shown as there is no particle flow step in this filter. Compared to SMC, NPF and ZPF give more accurate estimates for the speaker states. Before resampling, SMC has only four particles located on the speaker's face and the speaker in frame 287 is not well tracked.

In Fig. 3(a), we show the variation of ESS of NPF, ZPF and SMC from $\lambda = 0$ to $\lambda = 1$. As the baseline SMC does not use the particle flow, its ESS only has values at the beginning and end of the update step, as shown by the green dashed line. At the beginning of the update step, the ESS of the three filters is about 45, which is lower than $N_k/2$, shown in the pink dash-dot line. Using NPF and ZPF, the ESS is increased to 57 and 58, respectively, and therefore the particles do not need to be resampled. When $\lambda < 0.5$, the improvement of ESS given by NPF is higher than that given by ZPF. In Fig. 3(b), the OSPAs of the three filters are shown at each λ. After the update step, the OSPAs of NPF and ZPF are both decreased to 12. This shows that ZPF and NPF provide more accurate estimate of the speaker state than SMC. When $\lambda = 0.95$, the OSPA of ZPF increases slightly, due to the measurement errors. Fig. 4 shows the average OSPA and the number of speakers for the frames 287-500. It can be observed that ZPF and NPF give a smaller average OSPA than SMC. Due to the presence of occlusion from frames 300 to 500, the OSPAs for all methods have increased. At frame 345, ZPF and NPF give an average OSPA at about 19.3 and 18.9, respectively, resulting in a 24% and 29% performance improvement over SMC thanks to the more accurate estimate of the number of speakers offered by the particle flow.

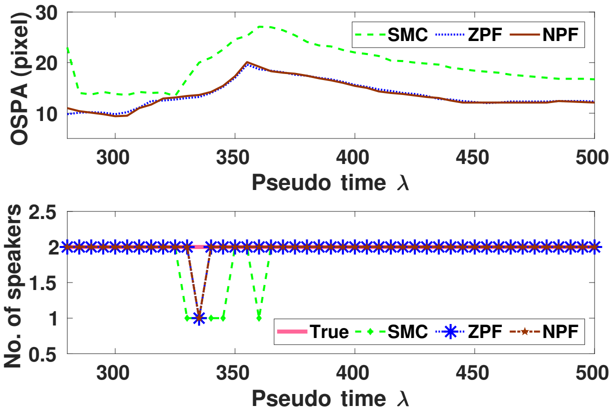


Fig. 4: The average OSPA of SMC, ZPF and NPF and the estimated number of speakers in the frames 287-500 of Sequence 24 (camera 1).

2) *Particle relocation methods:* Here the baseline SMC and SMC with the novel particle relocation methods are compared. To show the OSPAs of these two filtering algorithms, we ran experiments in frames 280-500 of Sequence 24 camera 1 which involves two speakers and visual occlusion. Fig. 5(a) shows the average OSPA. For convenience, the novel method in which the particles are relocated with the previous speaker states are abbreviated as Previous states. The average OSPA error is 20.8 for the novel method and 22.6 for the baseline method SMC. This means that the novel relocation method offers an 8% improvement over the baseline method.

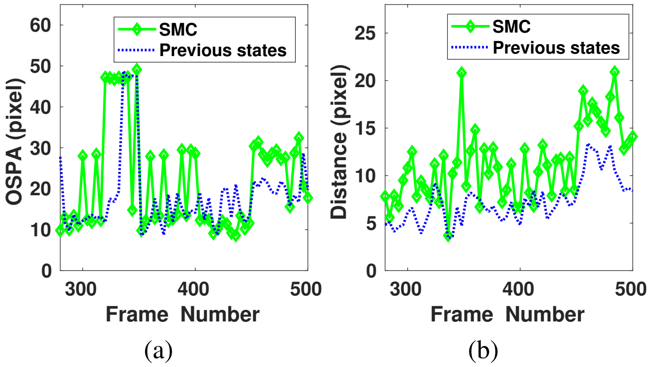


Fig. 5: Comparison the OSPA error (a) and the minimal distance d_m (b) between the novel relocation method with the speaker states at the previous time frame and the baseline relocation method of SMC.

Apart from that, Fig. 5(b) shows the distance measure in terms of Eq. (39). The average is 6.77 for the novel method and 10.78 for the baseline method. The novel relocation method offers 37% improvements. The running time of the novel relocation step (0.0529s) is higher than that of the baseline method (0.0218s), however, the running time of the overall algorithms are similar.

3) *Candidate speaker states:* To show the impact of using the candidate speaker states $\tilde{\boldsymbol{m}}^j_{k|k-1}$, ZPF is compared with the filter in our earlier work [7]. Although both methods use ZPF, DOA lines are used only in the update step of ZPF, rather than in that of the filter in [7]. Other steps and parameters of these filters are the same. The frames 500-1000 of Sequence 45 are used to test both methods since in these frames the speakers go out of the view of cameras which represents a challenging tracking scenario.

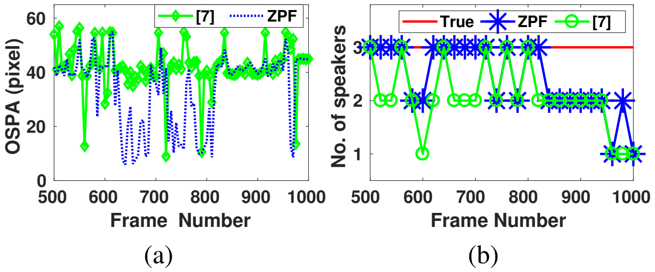


Fig. 6: Comparison of ZPF and the filter in [7] in terms of OSPA and the number of speakers.

Fig. 6(a) shows the average OSPA for this sequence, which is 38.6 for the filter in [7] and 31.6 for ZPF. This means that

TABLE V: The experimental results for NPF, ZPF and SMC with different levels of noise and different number of clutter in terms of the OSPA error

Method	$\boldsymbol{\sigma}_\varsigma$				N_c			
	0	10	20	30	0	5	10	50
NPF	19.41	23.54	26.71	32.94	19.41	19.75	21.81	24.37
ZPF	19.33	23.55	25.51	31.42	19.33	19.47	20.06	21.86
SMC	29.46	31.52	39.12	39.94	29.46	29.57	30.14	32.07

the proposed relocation method offers 19% improvements over the filter in [7]. In the frames 620-700, the average OSPA error is 21.6 for ZPF, and 38.2 for the filter in [7]. Fig. 6(b) shows that ZPF offers more accurate estimates for the number of speakers than the filter in [7], especially for the frames 620-700. The average running time of ZPF and the filter in [7] are 146.8s and 143.5s, respectively. This means the use of audio information leads to only 2.3% increase in running time.

4) Clutter and noise: We also evaluated the performance of our proposed filter in different levels of clutter and noise, as compared with SMC, using Sequence 45 of the AV16.3 dataset. Fig. 7 shows a frame of sequence 45 with clutter and noise. Gaussian noise $\varsigma_k \propto \mathcal{N}(0, \boldsymbol{\sigma}_\varsigma)$ and random clutter are added to the visual detection, where $\boldsymbol{\sigma}_\varsigma$ is the covariance matrix of noise. The clutter is shown in green stars. The visual detection without noise is shown in red points and its noise version in yellow diamonds. Table V shows the OSPA for different levels of noise with $\boldsymbol{\sigma}_\varsigma$ set from 0 to 30 pixels and the different number of clutter with N_c set from 0 to 50. We observe that particle flow gives a smaller OSPA, as compared with SMC, confirming that particle flow can improve the performance of SMC in different levels of noise. The positions of clutter are randomly set in the tracking area. The OSPA of three filters slightly increases with the level of clutter due to the RFS model used, however, the ZPF and NPF offer 32% and 24% improvement, respectively, even when $N_c = 50$.

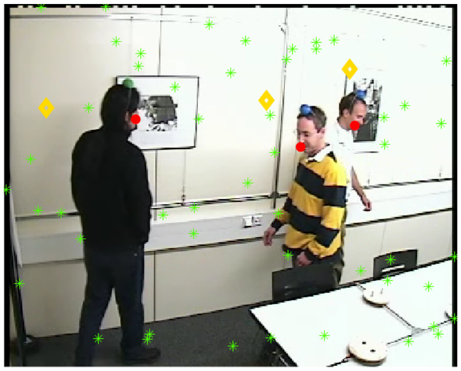


Fig. 7: The clutter and noise in Frame 325 of Sequence 45. The measurements, their noisy versions, and the clutter are shown in red points, yellow diamonds, and green stars, respectively.

5) Comparison with face detector: Here we compare the performance difference for using color histogram versus using measurements obtained by a face detector, e.g. the convolutional neural networks (CNNs) based face detector (Tiny) [10]. To distinguish the filters with face detectors from the filters with color histogram, ZPF, PPF, GPF, SMC and NPF with the face detector are renamed as AFZPF, AFPPF, AFGPF, AFSMC and AFNPF, respectively, where AF means both audio measurements (i.e. DoA) and visual measurements (obtained by face detector) are used in these algorithms. The frames 250-300 and 530-580 of the Sequence 45 camera 1 are used for evaluations since unreliable detection and occlusion happen in these frames. The OSPA and ESS of the different filters are shown in Table VI where ESS is not available for the Tiny face detector. AFZPF offers the lowest OSPA and highest ESS among all the filters except AFPPF, since the number of speakers is given in AFPPF. With the zero-diffusion particle flow, the OSPA of AFZPF is 51% and 12% lower than those of AFSMC and Tiny detector when speakers are occluded. We also tested these filters on frames 350-400 of the Sequence 45 camera 1 where occlusion does not happen. The face detector gives a lower OSPA than the filters with color histogram, while it gives a similar OSPA to AFZPF, i.e. ZPF with face detector.

TABLE VI: Experimental results for ZPF, NPF, AFZPF, AFPPF, AFGPF, AFSMC, AFNPF, and Tiny face detector in terms of the OSPA error and ESS for Sequence 45 camera 1.

Frame	ZPF	NPF	AFZPF	AFPPF	AFGPF	AFSMC	AFNPF	Tiny	
250-300	29.4	29.6	21.6	23.8	23.6	24.3	23.7	43.9	OSPA
	82	73	86	90	65	34	81	-	ESS
530-580	24.6	25.8	20.9	23.2	22.8	26.5	22.7	39.5	OSPA
	96	78	98	99	76	45	95	-	ESS
350-400	18.7	19.7	17.2	17.5	17.3	17.7	17.3	17.7	OSPA
	115	94	117	118	90	52	115	-	ESS

E. Comparison with other audio-visual algorithms

In this subsection, the proposed algorithms ZPF and NPF are compared with several baselines, including PPF [43], GPF [29], SMC [5], ASMC [5] and SMS algorithms [5]. Although Gebru et al. [47] also presented an audio-visual Bayesian framework, we did not consider this in our experiments, since it does not apply any particle flow or PHD filter. The same zero diffusion flow is used in PPF and GPF, as in our proposed ZPF.

In the PPF, the number of speakers is given when the particles are created due to the use of the particle filter framework. However, this information is unknown and time varying in our audio-visual tracking problem. Therefore, the following two situations are considered. For the AV16.3 dataset, the speakers utter together and continuously, therefore the number of speakers is the same as the number of estimated DOA lines. For the AVDIAR dataset, as the speakers are talking one by one, the number of speakers is given before running the algorithm. GPF is integrated with zero diffusion flow [29]. Based on the Gaussian mixture model, each particle has one dependent variance. Other parameters are the same as in ZPF. The NPF used is based on [45]. We set $\triangle\lambda = 0.01$ and $N_k = 50$, as already tested in the experiments shown in Section IV-C. To allow for fair comparison, the same measurements were used in all the filters compared.

Table VII reports the average OSPA over 50 random tests. All the frames of the 12 sequences have been used in these tests, where the number of speakers is varying with time. It can be observed that, using ZPF and NPF, about 31% reduction in tracking error has been achieved as compared with SMC. Compared with the ASMC filter, ZPF and NPF both give about 20% reduction in OSPA.

TABLE VII: The experimental results for ZPF, SMS, PPF, GPF, SMC, ASMC and NPF in terms of the OSPA error

Seq	ZPF	SMS	PPF	GPF	SMC	ASMC	NPF
24(1)	12.35	14.50	12.18	13.00	17.71	14.68	13.32
24(2)	13.24	15.35	13.12	15.13	19.83	16.15	13.20
24(3)	13.15	15.72	13.02	15.22	18.94	15.24	13.23
25(1)	15.94	17.17	14.90	18.28	19.13	16.21	15.96
25(2)	15.21	15.39	13.08	15.58	18.47	15.67	15.29
25(3)	16.22	17.62	14.98	18.62	21.61	17.95	16.29
30(1)	15.82	19.27	15.29	18.89	25.22	22.84	15.76
30(2)	13.43	16.16	13.86	16.12	19.37	16.17	13.41
30(3)	16.01	19.67	15.61	19.03	25.31	21.75	15.93
45(1)	17.60	23.40	24.50	23.12	29.46	26.07	17.65
45(2)	18.55	23.16	22.26	22.71	29.47	25.97	18.60
45(3)	19.54	23.80	24.34	23.76	28.43	26.41	19.50
Avg	**15.59**	**18.43**	**16.43**	**18.28**	**22.75**	**19.59**	**15.68**

To show how significant the difference is among the results of the tested algorithms in Table VII, the ANOVA based F-test [77] is applied and the significance test results are given in Table VIII. As the degree of freedoms for all the significance tests is $(1, 22)$ and the significance value is 5%, the corresponding critical value F_{crit} for $(1, 22)$ is 4.30 in terms of the F-distribution table [77] where the F-value is the ratio of the between-group variability to the within-group variability. The p-value is the probability of a more extreme result than the value achieved when the null hypothesis is true. According to the test, the results are considered as statistically significant if F-value $> F_{crit}$ and p-value is less than the significance value (0.05). It can be seen that the improvements of ZPF and NPF, over SMC, SMS and ASMC are statistically significant. However, the difference between ZPF and PPF is not significant. Nonetheless, in ZPF, the number of speakers is estimated, while in PPF, this is given as prior information. The difference between ZPF and NPF is also not significant.

TABLE VIII: Significance test for ZPF, SMS, PPF, GPF, ASMC and NPF.

Method	ZPF	SMS	PPF	GPF	ASMC	NPF	
ZPF	-	5.84	0.33	5.10	7.14	0.01	F
	-	0.024	0.057	0.034	0.014	0.921	p-value
NPF	0.01	5.64	0.027	4.89	6.98	-	F
	0.921	0.027	0.610	0.038	0.015	-	p-value
SMC	23.84	6.92	11.6	7.27	2.8	23.7	F
	7e-07	0.015	0.003	0.013	0.1082	7e-05	p-value

As shown in Table IX, ZPF has a lower computational cost than GPF. Although ZPF and GPF use the same initial number of particles, the number of particles is drastically varying and a few particles are added in the update step of the GPF. For further understanding, we calculated the total number of particles used in the update step of ZPF and GPF. In the update step of ZPF, the average number of particles is 108, while in GPF, it is 463. The computational complexities are shown in the last line of Table IX. The complexity (Com) of SMC, SMC and ASMC is the lowest at U_kN_k. The complexity of ZPF, PPF and NPF does not depend on the number of measurements.

TABLE IX: Computational cost comparison per Sequence (s/Sequence) for ZPF, SMS, PPF, GPF, ASMC and NPF.

Seq	ZPF	SMS	PPF	GPF	SMC	ASMC	NPF
24	234.5	146.2	211.3	435.3	80.6	102.5	174.6
25	236.0	147.2	210.6	435.5	83.6	105.2	175.7
30	235.1	146.8	211.7	436.8	83.7	105.4	174.9
45	347.8	208.5	315.9	655.3	124.3	172.9	264.6
Time	**263.4**	**162.2**	**237.4**	**490.7**	**93.1**	**121.5**	**197.5**
Com	N_kN_λ	U_kN_k	N_kN_λ	$U_kN_kN_\lambda$	U_kN_k	U_kN_k	N_kN_λ

To show the performance of the proposed method on other datasets rather than AV16.3, we selected sequence 32 (four speakers) and 09 (three speakers) from the AVDIAR dataset [47], and the frames 100-170 (four speakers) and frames 180-250 (five speakers) of sequence UKA from the CLEAR dataset [48]. Their average errors are summarised in Table X. Our proposed ZPF and NPF methods offer a similar OSPA which is the lowest OSPA among all the filters except PPF. Note that the number of speakers is given to PPF as a priori. However, as the speakers are talking one by one, the performance difference among the compared filters is not significant. The OSPA of all the methods is increased with the increase in the number of speakers.

TABLE X: Experimental results for ZPF, SMS, PPF, GPF, SMC and NPF in terms of the OSPA error for Sequence 09 and 32 of the AVDIAR dataset and frames 100-170 and frames 180-250 of sequence UKA 20060726 of the CLEAR dataset.

Filters	sequence 09	sequence 32	frames 100-170	frames 180-250
ZPF	13.72	14.37	28.62	31.57
SMS	13.95	14.90	29.35	36.68
PPF	11.68	12.14	24.21	26.65
GPF	13.82	14.78	30.25	37.84
SMC	14.96	16.86	31.58	38.61
NPF	13.80	14.42	28.60	31.55

V. Conclusion

We have presented a new method for mitigating the particle degeneracy issue in SMC-PHD filtering, and implemented both the zero-flow and non-zero flow algorithms for audio-visual multi-speaker tracking. We have demonstrated the advantages of the proposed algorithms as compared with several audio-visual tracking baselines, in terms of ESS and OSPA. The computational cost of AV-NPF-SMC-PHD is lower than that of AV-ZPF-SMC-PHD, while AV-ZPF-SMC-PHD is easier to implement. Apart from that, the speaker states and weights in the previous frames have been used for relocating particles with DOA lines. The proposed relocation method offers a lower OSPA than the baseline methods. The proposed methods could be further improved to allow better detection of speakers when silent speakers are visually present in the scene.

Acknowledgments

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 the EPSRC Grant EP/K014307/1 and the MOD University Defence Research Collaboration in Signal Processing. The authors thank Yunpeng Li and Qiangqiu Kong from the University of Surrey for their helpful comments. The authors also thank the associate editor and anonymous reviewers for the valuable suggestions they provided.

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] B. Menelas, L. Picinalli, B. F. Katz, and P. Bourdot, "Audio haptic feedbacks for an acquisition task in a multi-target context," *IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 51–54, 2010.

[4] E. D'Arca, N. M. Robertson, and J. R. Hopgood, "Robust indoor speaker recognition in a network of audio and video sensors," *Signal Processing*, vol. 129, pp. 137–149, 2016.

[5] 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.

[6] 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.

[7] 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.

[8] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, vol. 1, 2001, pp. 741–746.

[9] G. R. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Proc. IEEE Workshop on Applications of Computer Vision*, 1998, pp. 214–219.

[10] P. Hu and D. Ramanan, "Finding tiny faces," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[11] T.-J. Shan, M. Wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation of coherent signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 806–811, 1985.

[12] X. Qian, A. Xompero, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "3D mouth tracking from a compact microphone array co-located with a camera," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[13] G. Welch, G. Bishop *et al.*, "An introduction to the Kalman filter," 1995.

[14] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*, Oct. 2005, pp. 118–121.

[15] K. Okuma, A. Taleghani, N. d. Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," *Proc. IEEE. European Conference on Computer Vision (ECCV)*, pp. 28–39, 2004.

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

[17] 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.

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

[19] D. Y. Kim, B.-T. Vo, and S. Nordholm, "Multiple speaker tracking with the GLMB filter," in *Control, Automation and Information Sciences (ICCAIS), 2017 International Conference on*. IEEE, 2017, pp. 38–43.

[20] Y. Liu, W. Wang, and Y. Zhao, "Particle flow for sequential Monte Carlo implementation of probability hypothesis density," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2017, pp. 1371–1375.

[21] M. K. Pitt and N. Shephard, "Filtering via simulation: Auxiliary particle filters," *Journal of the American statistical association*, vol. 94, no. 446, pp. 590–599, 1999.

[22] R. Van Der Merwe, A. Doucet, N. De Freitas, and E. A. Wan, "The unscented particle filter," in *Advances in neural information processing systems*, 2001, pp. 584–590.

[23] N. Whiteley, S. Singh, and S. Godsill, "Auxiliary particle implementation of probability hypothesis density filter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 46, no. 3, pp. 1437–1454, 2010.

[24] M. R. Danaee, "Unscented auxiliary particle filter implementation of the cardinalized probability hypothesis density filter," *arXiv preprint arXiv:1506.02570*, 2015.

[25] C. Berzuini, N. G. Best, W. R. Gilks, and C. Larizza, "Dynamic conditional independence models and Markov chain Monte Carlo methods," *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1403–1412, 1997.

[26] W. R. Gilks and C. Berzuini, "Following a moving target Monte Carlo inference for dynamic Bayesian models," *Journal of the Royal Statistical Society*, vol. 63, no. 1, pp. 127–146, 2001.

[27] F. Daum and J. Huang, "Particle flow and Monge-Kantorovich transport," in *Proc. IEEE Int. Conf. Information Fusion (FUSION)*, 2012, pp. 135–142.

[28] V. Kilic, X. Zhong, M. Barnard, W. Wang, and J. Kittler, "Audio-visual tracking of a variable number of speakers with a random finite set approach," in *Proc. IEEE Int. Conf. Information Fusion (FUSION)*, 2014, pp. 1–7.

[29] L. Zhao, J. Wang, Y. Li, and M. J. Coates, "Gaussian particle flow implementation of PHD filter," in *Proc. SPIE Defense and Security*, vol. 9842, May 2016, pp. 98 420D – 98 420D–10.

[30] F. Daum and J. Huang, "Particle flow for nonlinear filters with log-homotopy," in *Proc. SPIE Conf. Signal Processing Sensor Fusion, Target Recognition*, vol. 6969, Apr. 2008, pp. 696 918–696 918–12.

[31] J. Heng, A. Doucet, and Y. Pokern, "Gibbs flow for approximate transport with applications to Bayesian computation," *arXiv preprint arXiv:1509.08787*, 2015.

[32] 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.

[33] 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.

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

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

[36] 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.

[37] ——, "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.

[38] T. Ding and M. J. Coates, “Implementation of the Daum-Huang exact-flow particle filter,” in *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, 2012, pp. 257–260.

[39] M. A. Khan, M. Ulmke, B. Demissie, F. Govaers, and W. Koch, “Combining log-homotopy flow with tensor decomposition based solution for Fokker-Planck equation,” in *Proc. IEEE Int. Conf. Information Fusion (FUSION)*, 2016, pp. 2229–2236.

[40] C. Chlebek, J. Steinbring, and U. D. Hanebeck, “Progressive Gaussian filter using importance sampling and particle flow,” in *Proc. IEEE Int. Conf. Information Fusion (FUSION)*, 2016, pp. 2043–2049.

[41] Y. Li and M. Coates, “Particle filtering with invertible particle flow,” *IEEE Trans. Signal Processing*, vol. 65, no. 15, pp. 4102–4116, 2016.

[42] L. Chen and R. K. Mehra, “A study of nonlinear filters with particle flow induced by log-homotopy,” in *Signal Processing, Sensor Fusion, and Target Recognition XIX*, vol. 7697. International Society for Optics and Photonics, 2010, p. 769706.

[43] Y. Li, L. Zhao, and M. Coates, “Particle flow for particle filtering,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, May 2016, pp. 3979–3983.

[44] A.-A. Saucan, Y. Li, and M. Coates, “Particle flow superpositional GLMB filter,” in *Proc. SPIE Conf. Signal Processing, Sensor Fusion, and Target Recognition*, vol. 10200. International Society for Optics and Photonics, 2017, p. 102000F.

[45] 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 (ICASSP)*. IEEE, 2018.

[46] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, “AV 16.3: an audio-visual corpus for speaker localization and tracking,” in *Proc. Int. Workshop on Machine Learning for Multimodal Interaction*. Springer, 2004, pp. 182–195.

[47] I. D. Gebru, S. Ba, X. Li, and R. Horaud, “Audio-visual speaker diarization based on spatiotemporal Bayesian fusion,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, pp. 7106–7113, 2017.

[48] M. H. Ooi, T. Solomon, Y. Podin, A. Mohan, W. Akin, M. A. Yusuf, S. del Sel, K. M. Kontol, B. F. Lai, D. Clear *et al.*, “Evaluation of different clinical sample types in diagnosis of human enterovirus 71-associated hand-foot-and-mouth disease,” *J. Clinical Microbiology*, vol. 45, no. 6, pp. 1858–1866, Apr. 2007.

[49] B. Ristic, D. Clark, and B.-N. Vo, “Improved SMC implementation of the PHD filter,” in *Proc. IEEE Int. Conf. Information Fusion (FUSION)*, Jul. 2010, pp. 1–8.

[50] R. Mahler and A. El-Fallah, “Cphd and phd filters for unknown backgrounds, part iii: tractable multitarget filtering in dynamic clutter,” in *Signal and Data Processing of Small Targets*, vol. 7698. International Society for Optics and Photonics, 2010, p. 76980F.

[51] R. P. Mahler, B.-T. Vo, and B.-N. Vo, “Cphd filtering with unknown clutter rate and detection profile,” *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3497–3513, 2011.

[52] T. Kailath, “The divergence and bhattacharyya distance measures in signal selection,” *IEEE transactions on communication technology*, vol. 15, no. 1, pp. 52–60, 1967.

[53] F. Y. Shih and C. C. Pu, “A skeletonization algorithm by maxima tracking on euclidean distance transform,” *Pattern Recognition*, vol. 28, no. 3, pp. 331–341, 1995.

[54] 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.

[55] 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.

[56] D. B. Rubin, “A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the sir algorithm,” *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 543–546, 1987.

[57] A. F. Smith and A. E. Gelfand, “Bayesian statistics without tears: a sampling–resampling perspective,” *The American Statistician*, vol. 46, no. 2, pp. 84–88, 1992.

[58] J. S. Liu and R. Chen, “Sequential Monte Carlo methods for dynamic systems,” *Journal of the American statistical association*, vol. 93, no. 443, pp. 1032–1044, 1998.

[59] G. Kitagawa, “Monte carlo filter and smoother for non-gaussian nonlinear state space models,” *Journal of computational and graphical statistics*, vol. 5, no. 1, pp. 1–25, 1996.

[60] M. A. Khan and M. Ulmke, “Improvements in the implementation of log-homotopy based particle flow filters,” in *Proc. IEEE Int. Conf. Information Fusion (FUSION)*, Jul. 2015, pp. 74–81.

[61] M. A. A. Khan, “Nonlinear filtering based on log-homotopy particle flow,” Ph.D. dissertation, Universitats und Landesbibliothek Bonn, 2018.

[62] F. E. De Melo, S. Maskell, M. Fasiolo, and F. Daum, “Stochastic particle flow for nonlinear high-dimensional filtering problems,” *arXiv preprint arXiv:1511.01448*, 2015.

[63] F. Daum, J. Huang, and A. Noushin, “Generalized gromov method for stochastic particle flow filters,” in *Signal Processing, Sensor/Information Fusion, and Target Recognition XXVI*, vol. 10200. International Society for Optics and Photonics, 2017, p. 102000I.

[64] G. Lathoud, “Spatio-temporal analysis of spontaneous speech with microphone arrays,” EPFL, Tech. Rep., 2007.

[65] 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.

[66] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, “A tutorial on particle filters for online nonlinear/non-Gaussian bayesian tracking,” *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 174–188, 2002.

[67] A. Doucet, S. Godsill, and C. Andrieu, “On sequential Monte Carlo sampling methods for Bayesian filtering,” *Statistics and computing*, vol. 10, no. 3, pp. 197–208, 2000.

[68] I. D. Gebru, S. Ba, G. Evangelidis, and R. Horaud, “Tracking the active speaker based on a joint audio-visual observation model,” in *Proc. IEEE Int. Conf. Computer Vision Workshops*, Dec. 2015, pp. 15–21.

[69] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin, “Co-localization of audio sources in images using binaural features and locally-linear regression,” *IEEE/ACM Trans. Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 4, pp. 718–731, 2015.

[70] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, “The AMI meeting corpus: A pre-announcement,” in *Proc. Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[71] V. P. Minotto, C. R. Jung, and B. Lee, “Multimodal multi-channel online speaker diarization using sensor fusion through SVM,” *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1694–1705, 2015.

[72] M. Taj, “Surveillance performance evaluation initiative (SPEVI) audio-visual people dataset,” *Internet: http://www. eecs. qmul. ac. uk/ andrea/spevi. html*, 2007.

[73] 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.

[74] J. H. DiBiase, “A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays,” Ph.D. dissertation, Brown University, 2000.

[75] F. Daum, J. Huang, and A. Noushin, “A friendly rebuttal to Mallick and Sindhu on particle flow for Bayes’ rule,” in *Proc. SPIE Defense and Security*. International Society for Optics and Photonics, 2016, pp. 98 420H–98 420H–11.

[76] F. Daum and J. Huang, “Curse of dimensionality and particle filters,” in *Proc. IEEE Aerospace Conference*, vol. 4, Mar. 2003, pp. 1979–1993.

[77] P. G. Hoel *et al.*, “Elementary statistics,” *Elementary statistics*, 1960.

Yang Liu received the B.Sc. and M.Sc. degrees from Harbin Engineering University, Harbin, China, in 2012 and 2015, respectively. He is currently a Ph.D. student in Electronics Engineering at University of Surrey, Guildford, U.K. He joined the Centre for Vision, Speech and Signal Processing, in University of Surrey, Guildford, U.K., in 2015. His current research interests include audio-visual signal processing, multimodal speaker tracking, Monte Carlo methods in high-dimensional spaces, particle flow and PHD filters.

Volkan Kılıç received the B.Sc. degree in electrical and electronics engineering from Anadolu University, Eskisehir, Turkey, in 2008. He received the M.Sc. degree in Electronics Engineering from Istanbul Technical University, Institute of Science and Technology, Istanbul, Turkey, and started the Ph.D. in the same university in 2010. He joined the Centre for Vision, Speech and Signal Processing in University of Surrey, Guildford, U.K in 2012 and completed his Ph.D. in 2016. He is currently Assistant Professor in Izmir Katip Celebi University, Izmir, Turkey. His current research interests include audio-visual signal processing, image processing, sensor fusion and smartphone sensing.

Jian Guan received his B.Sc. and M.Sc. degrees from the College of Computer Science and Technology, Jilin University, China, in 2005 and 2010 respectively. He received the Ph.D. degree in Computer Applied Technology from Harbin Institute of Technology in 2018. From 2011 to 2012, he was a Teaching Assistant at Zhuhai College of Jilin University. During the time from October 2014 to January 2017, he was a visiting Ph.D. student at the Centre for Vision, Speech and Signal Processing (CVSSP), the University of Surrey, UK. He is currently a Senior Lecturer at Harbin Engineering University, and a Guest Professor at State Key Laboratory of Space-Ground Integrated Information Technology. His research interests include audio and speech signal processing, image processing, remote sensing, and machine learning.

Wenwu Wang (M'02-SM'11) was born in Anhui, China. He received the B.Sc. degree in 1997, the M.E. degree in 2000, and the Ph.D. degree in 2002, all from Harbin Engineering University, China.

He then worked in King's College London (2002-2003), Cardiff University (2004-2005), Tao Group Ltd. (now Antix Labs Ltd.) (2005-2006), and Creative Labs (2006-2007), before joining University of Surrey, UK, in May 2007, where he is currently a Professor in Signal Processing and Machine Learning, and a Co-Director of the Machine Audition Lab within the Centre for Vision Speech and Signal Processing. He has been a Guest Professor at Qingdao University of Science and Technology, China, since 2018.

His current research interests include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He has (co)-authored over 250 publications in these areas.

His work has been recognised internationally. He won (with his team) the Best Paper Award on LVA/ICA 2018, the Best Oral Presentation on FSDM 2016, the Top Paper Award in IEEE ICME 2015, Best Student Paper Award shortlists on IEEE ICASSP 2019 and LVA/ICA 2010. His papers are among the Most Downloaded Papers in IEEE/ACM Transactions on Audio Speech and Language Processing in 2018 and 2019, and Featured Articles in IEEE Transactions on Signal Processing 2013. As a team member, he achieved the 3rd place (among 558 submitted systems) in the 2018 Kaggle Challenge ”Free-sound general purpose audio tagging”, the 1st place (among 35 submitted systems) in the 2017 DCASE Challenge on ”Large-scale weakly supervised sound event detection for smart cars”, the TVB Europe Award for Best Achievement in Sound in 2016 and the finalist for GooglePlay Best VR Experience in 2017, and the Best Solution Award on the Dstl Challenge ”Under-sampled signal recognition” in 2012.

He has been a Senior Area Editor (2019-) and an Associate Editor (2014-2018) for IEEE Transactions on Signal Processing. He is a Publication Co-Chair for ICASSP 2019, Brighton, UK.