Labelled Non-Zero Diffusion Particle Flow SMC-PHD Filtering for Multi-Speaker Tracking

Yang Liu, *Senior Member, IEEE,* Yong Xu, *Senior Member, IEEE,* Peipei Wu, Wenwu Wang, *Senior Member, IEEE*

***Abstract*—Particle flow (PF) is a method originally proposed for single target tracking, and used recently to address the weight degeneracy problem of the sequential Monte Carlo probability hypothesis density (SMC-PHD) filter for audio-visual (AV) multi-speaker tracking, where the particle flow is calculated by using only the measurements near the particle, assuming that the target is detected, as in a recent method based on non-zero particle flow (NPF), i.e. the AV-NPF-SMC-PHD filter. This, however, can be problematic when occlusion happens and the occluded speaker may not be detected. To address this issue, we propose a new method where the labels of the particles are estimated using the likelihood function, and the particle flow is calculated in terms of the selected particles with the same labels. As a result, the particles associated with detected speakers and undetected speakers are distinguished based on the particle labels. With this novel method, named as AV-LPF-SMC-PHD, the speaker states can be estimated as the weighted mean of the labelled particles, which is computationally more efficient than using a clustering method as in the AV-NPF-SMC-PHD filter. The proposed algorithm is compared systematically with several baseline tracking methods using the AV16.3, AVDIAR and CLEAR datasets, and is shown to offer improved tracking accuracy with a lower computational cost.**

***Index Terms*—Audio-Visual Tracking, SMC-PHD Filter, Particle Flow**

I. Introduction

Multi-speaker tracking in an enclosed space is an important task in several subject areas such as spatial audio [1], surveillance [2], sport video analysis [3], target discrimination [4], and speech recognition [5]. However, the number of active speakers is often unknown and time-varying. Apart from that, the multi-speaker tracker has to deal with many complex sources of uncertainty, such as measurement origin uncertainty, false alarm, noise, clutters, missing data, and births and deaths of targets.

To address these problems, multiple heterogeneous sensors can be exploited jointly for their complementarity. For example, if the speaker is visually occluded, they can be tracked using the audio information, and if the speaker stops talking, they can be tracked using the visual information. Other modalities, such as radar, sonar, electro-optical, infrared and unattended, can also be used for multi-speaker tracking. In this work, we consider the use of the audio-visual (AV) sensor, which has been widely adopted due to its low cost and easy installation [6].

With the measurements from multiple heterogeneous sensors, the Bayesian approach is a popular choice which provides an intuitive way for the estimation of speaker states [7]. Early methods include the Kalman filter (KF) [8], extended Kalman filter (EKF) [9] and particle filter [10], which can be used to track a fixed and known number of speakers, while more recent methods including random finite sets (RFS) [11], Gaussian mixture (GM) PHD filter [12], sequential Monte Carlo (SMC) PHD filter [13], cardinalized PHD filter [14], generalized labelled multi-Bernoulli (GLMB) RFS [15] and variational Bayesian methods [16], [17], [18] are employed to track an unknown and time-variant number of speakers. The SMC-PHD filter is widely used to estimate the target states under the non-linear model using a set of random particles. Compared to the trendy methods in deep learning, such as Yolo with DeepSort [19], the SMC-PHD filter has the following advantages. First, it offers an elegant mathematical framework for interpreting the relationship between the states of speakers and their measurements including mis-detections (e.g. during occlusions). Second, it does not involve model training, while deep learning models usually require the optimisation of their parameters by training on additional, often manually annotated data. Third, it is a flexible method and could be used together with deep learning models, whose detection results can be used as measurements for the SMC-PHD filters. Therefore, our focus here is on the SMC-PHD filter.

Although the SMC-PHD filter has a moderate computational cost, 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, after they are updated for a number of iterations. This problem also happens in other Monte Carlo based tracking methods, such as AV-GLMB [21], AV importance particle filter [22] and AV3T [23]. To address the weight degeneracy problem, several ideas, such as the auxiliary SMC [24], unscented SMC [25], and more recently, particle flow filters [26], [27], [28], [29], [30], [31] and end-to-end neural network [32], [33], have been developed. Although deep learning based methods [32], [33] have drawn increasing attention, their performance depends on the quality and size of the training data. In the particle flow methods, the particles are migrated from the prior density to the posterior density in terms of a homotopy function, with which either zero diffusion particle flow (ZPF) or non-zero diffusion particle flow (NPF) can be derived with different assumptions.

Y. Liu is with Meta, Seattle, USA. He was with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, U.K. E-mail: yangliuav@gmail.com.

Y. Xu is with Tencent AI Lab, Seattle, USA. E-mail: lucayongxu@global.tencent.com.

P. Wu and W. Wang are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, U.K. E-mail: [p.wu, w.wang]@surrey.ac.uk.

The ZPF and NPF have been used with the AV-SMC-PHD filter in the AV-ZPF-SMC-PHD filter [34] and the AV-NPF-SMC-PHD filter [35], respectively, to improve the estimation of the posterior density, and reduce the chances required for particle resampling. Apart from that, ZPF has also been used previously to improve the tracking performance of particle filter and PHD filter with a reduced number of particles in the particle flow particle filter (PFPF) [36] and the Gaussian mixture PHD filter [37]. However, the particle flow is designed for addressing the weight degeneracy issue for single-target tracking, without using any label information about the particles. For example, in both AV-ZPF-SMC-PHD [34] and AV-NPF-SMC-PHD [35], the particle flows are calculated only with the measurements near the target.

To address the multi-measurement problem of multi-target tracking, Gaussian particle flow implementation of PHD (GPF-PHD) filter [37], zero diffusion particle flow SMC delta-GLMB (ZPF-delta-GLMB) filter [38] and AV-GLMB [21] are proposed. In the GPF-PHD filter, particle flow is calculated based on multiple hypothesis tracking [39], where each particle is independently updated by each measurement. Therefore, the number of particles is linearly increased with the growth in the number of measurements, which increases computational cost, especially for a large number of clutters. The ZPF-delta-GLMB filter is proposed to calculate the label variables based on the maximum labelling probability density. The ZPF has been used to address the weight degeneracy issue of the delta-GLMB filter [38]. However, in ZPF-delta-GLMB, it is assumed that the speaker is always detectable. As a result, the particles associated with undetected targets may be incorrectly moved towards other speakers.

In this paper, we propose a labelled non-zero diffusion particle flow (LPF) SMC-PHD filter for audio-visual multi-speaker tracking, where the particle labels are estimated based on the likelihood function, and further used to calculate the particle flow. The labels of the particles include visual labels, audio labels and speaker labels, which are estimated independently. The visual and audio labels are calculated from the visual and audio measurements, respectively, while the speaker labels are estimated based the speaker states in the previous time frames. The particles with the same visual, audio and speaker labels can be considered as belonging to the same group. As a result, the mean and covariance of the particles can be estimated in terms of the label information, i.e. without having to cluster the particles using a k-means algorithm as in conventional SMC-PHD filters [13]. In addition, the scenario where the speakers are undetected is also considered and the states of the speakers are updated according to these labels.

Compared to our earlier work, i.e. AV-NPF-SMC-PHD [35], and another related method, i.e. ZPF-delta-GLMB filter [38], the particle flow in the proposed method is calculated with the particles selected in terms of their labels, and the particles associated with the undetected, occluded and detected speakers can be distinguished based on the label information. Compared to AV-GLMB [21] which is based on labelled RFS [40], [41], our method differs in the following three aspects. First, AV-GLMB uses the label information in the labelled RFS, while our method uses the label information in the particle flow, but the PHD filter we used is still an unlabelled RFS. Compared to GLMB, the PHD filter has a lower computational complexity [42]. Second, the labels used in AV-GLMB are unique and different for different RFS. However, the labels used in the proposed method are three independent labels (i.e. audio, visual and speaker labels). The particles belonging to the same speakers would have the same labels. Therefore, the meaning of "label" and the motivation for its use in AV-GLMB are significantly different from those in our method. Third, GLMB still has the weight degeneracy problem, as discussed in [38]. Our proposed LPF could be used with GLMB [21] to address the weight degeneracy problem of AV-GLMB.

Preliminary results were presented briefly in a conference paper [43]. Here we provide a comprehensive treatment of the proposed method, including further experimental results. There are three main improvements. First, we use the particle labels to distinguish the particles associated with audio measurements, visual measurements, and speakers, respectively. The particles with the same labels are used to estimate the particle covariance matrix and are updated by the associated audio and visual measurements in particle flow. Second, the labelled NPF is proposed to update the particles with a novel audio-visual likelihood function. The particle flow can be calculated when the speaker is silent or visually occluded. Third, with the label information, the clustering step, which is often used to estimate the speaker states in the PHD filter, can be exempted, and simply replaced by the mean of the labelled particles.

This paper is organised 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 using comprehensive experiments. Finally, Section V concludes the paper.

II. Problem Statement And Background

This section describes our problem formulation and the AV-NPF-SMC-PHD filter (in Algorithm 1). For clarity, the notations used in this paper are summarised in Table VIII in Appendix. We assume that the speaker dynamics and observations are described as:

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

$$\{\mathring{\boldsymbol{z}}_k^o\}_{o=1}^{\mathring{N}_k} = \mathring{\mathbf{F}}_{\boldsymbol{z}}\left(\{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k}, \mathring{\boldsymbol{\Psi}}_k\right) + \mathring{\boldsymbol{\epsilon}}_k. \tag{2}$$

$$\{\breve{\boldsymbol{z}}_k^u\}_{u=1}^{\breve{N}_k} = \breve{\mathbf{F}}_{\boldsymbol{z}}\left(\{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k}, \breve{\boldsymbol{\Psi}}_k\right) + \breve{\boldsymbol{\epsilon}}_k, \tag{3}$$

where $\tilde{\boldsymbol{m}}_k^j \in \mathbb{R}^M$ is the state vector of the jth speaker at time k, $\tilde{}$ is used to distinguish the speaker state from the particle state used later, and $\tilde{N}_k$ is the number of speakers at time k. Let $\{\mathring{\boldsymbol{z}}_k^o\}_{o=1}^{\mathring{N}_k}$ and $\{\breve{\boldsymbol{z}}_k^u\}_{u=1}^{\breve{N}_k}$ denote the set of $\mathring{N}_k$ audio and $\breve{N}_k$ visual measurements at time k, respectively where o and u are used to represent the index of the audio and visual measurements, respectively. Different from [43], we use bounding boxes to represent the speakers. The state $\tilde{\boldsymbol{m}}_k^j = [x_k^j, y_k^j, \dot{x}_k^j, \dot{y}_k^j, w_k^j, l_k^j]^T$ consists of positions (x_k^j, y_k^j), velocities $(\dot{x}_k^j, \dot{y}_k^j)$, and size (w_k^j, l_k^j) of the target j at time

k, while the measurement is a noisy version of the position. We define the measurement noise and clutter terms as $\mathring{\boldsymbol{\Psi}}_k$ and $\mathring{\boldsymbol{\epsilon}}_k$ for audio measurements, and $\breve{\boldsymbol{\Psi}}_k$ and $\breve{\boldsymbol{\epsilon}}_k$ for visual measurements, respectively. The state transition model is denoted as $\mathbf{F}_{\tilde{\boldsymbol{m}}}$, where the system noise is $\boldsymbol{\Upsilon}_k$. The nonlinear measurement models for audio and visual information are denoted as $\mathring{\mathbf{F}}_{\boldsymbol{z}}$ and $\breve{\mathbf{F}}_{\boldsymbol{z}}$, respectively.

In [35], an AV-NPF-SMC-PHD filter is presented for audio-visual multi-speaker tracking. The audio information and visual information are applied in the prediction and update steps. The particle set is defined as $\{\boldsymbol{m}^i_{k-1}, \omega^i_{k-1}\}^{N_k}_{i=1}$, where N_k is the number of particles at time k, and $\boldsymbol{m}^i_{k-1}$ and ω^i_{k-1} are the state and weight of the ith particle at time $k-1$. The particle state is obtained by the proposal distribution $q_k(\boldsymbol{m}^i_{k|k-1}|\boldsymbol{m}^i_{k-1}, \boldsymbol{Z}_k)$, where $\boldsymbol{Z}_k$ is the set of observations,

$$\boldsymbol{m}^i_{k|k-1} \sim q_k(\cdot|\boldsymbol{m}^i_{k-1}, \boldsymbol{Z}_k). \tag{4}$$

Their weights are predicted as,

$$\omega^i_{k|k-1} = \frac{\phi\left(\boldsymbol{m}^i_{k|k-1}|\boldsymbol{m}^i_{k-1}\right)\omega^i_{k-1}}{q_k\left(\boldsymbol{m}^i_{k|k-1}|\boldsymbol{m}^i_{k-1}, \boldsymbol{Z}_k\right)}, i = 1, ..., N_k, \tag{5}$$

where $\phi(\boldsymbol{m}^i_{k|k-1}|\boldsymbol{m}^i_{k-1})$ is the analogue of the state transition probability with the previous state $\boldsymbol{m}^i_{k-1}$. If a new speaker appears, N_B particles are sampled from the new born importance function p_k,

$$\boldsymbol{m}^i_{k|k-1} \sim p_k(\cdot|\boldsymbol{Z}_k). \tag{6}$$

Their weights are

$$\omega^i_{k|k-1} = \frac{\gamma_k(\boldsymbol{m}^i_{k|k-1})}{N_B p_k(\boldsymbol{m}^i_{k|k-1}|\boldsymbol{Z}_k)}, i = N_k+1, ..., N_k+N_B, \tag{7}$$

where γ_k is the PHD of new targets.

In the update step, the audio-visual likelihood function h^i_k is obtained as:

$$h^i_k = \frac{\mathring{\boldsymbol{h}}^{i\,T}_k \mathring{\boldsymbol{\omega}}_k + \breve{\boldsymbol{h}}^{i\,T}_k \breve{\boldsymbol{\omega}}_k}{\|\mathring{\boldsymbol{\omega}}_k\|_1 + \|\breve{\boldsymbol{\omega}}_k\|_1}, \tag{8}$$

where $\mathring{\boldsymbol{\omega}}_k$ and $\breve{\boldsymbol{\omega}}_k$ are the weights for the audio and visual likelihood, respectively, and $\|\cdot\|_1$ denotes the L_1 norm. Then the particle state is updated by the NPF,

$$\boldsymbol{m}^i_k \Leftarrow \boldsymbol{m}^i_k + \triangle\boldsymbol{m}^i_k \lambda, \tag{9}$$

where

$$\triangle\boldsymbol{m}^i_k = \boldsymbol{f}^i_k(\boldsymbol{m}^i_k, \lambda)\triangle\lambda + \upsilon^i_k \boldsymbol{w}^i_k, \tag{10}$$

where $\boldsymbol{f}^i_k \in \mathbb{R}^M$ is the particle flow vector and $\boldsymbol{w}^i_k \in \mathbb{R}^M$ is the Wiener process with the diffusion coefficient υ^i_k. It moves the particle $\boldsymbol{m}^i_{k|k-1}$ with the distance $\triangle\boldsymbol{m}^i_{k|k-1}$ for the time period $\triangle\lambda$. Based on the Fokker-Planck equation [44], the non-zero particle flow $\boldsymbol{f}^i_k$ is calculated by the partial differential equation:

$$\boldsymbol{f}^i_k = -[\boldsymbol{\nabla}^2 \log \psi^i_k]^{-1}(\boldsymbol{\nabla} \log h^i_k), \tag{11}$$

where

$$\boldsymbol{\nabla}^2 \log \psi^i_k \approx -(\boldsymbol{P}^i_{k|k-1})^{-1} + \lambda \boldsymbol{\nabla}^2 \log h^i_k, \tag{12}$$

where $\boldsymbol{P}^i_{k|k-1}$ is the covariance matrix of $\boldsymbol{m}^i_{k|k-1}$. The derivation of Eq. (11) can be found in [45]. The first and second derivative of the likelihood function can be found in [35]. Then the weights of the particles are calculated as

$$\omega^i_k = \left[1 - p^i_{D,k} + \sum_{\boldsymbol{z}^r_k \in \boldsymbol{Z}_k} \frac{p^i_{D,k} h^{i,r}_k}{\kappa_k(\boldsymbol{z}^r_k) + G^r_k}\right] \omega^i_{k|k-1}, \tag{13}$$

where

$$G^r_k = \sum_{i=1}^{N_k} p^i_{D,k} h^{i,r}_k \omega^i_{k|k-1}, \tag{14}$$

in which $\kappa_k(\boldsymbol{z}^r_k)$ denotes the clutter intensity of the rth measurement $\boldsymbol{z}^r_k$ at time k, $p^i_{D,k}$ is the detection probability at time k, and $h^{i,r}_k$ is the likelihood of the ith particle for the rth measurement at time k. The measurement $\boldsymbol{z}^r_k$ is calculated by $\mathring{\boldsymbol{z}}^o_k$ and $\breve{\boldsymbol{z}}^u_k$ [35]. The number of speakers is estimated as the sum of the weights. The states and weights of the speakers $\{\tilde{\boldsymbol{m}}^j_k, \tilde{\omega}^j_k\}^{\tilde{N}_k}_{j=1}$ can be calculated using a clustering step e.g. the k-means clustering method [46] or multi-expected a posterior (MEAP) [47]. Finally, resampling is performed when the effective sample size (ESS) [48] is smaller than half

Algorithm 1 AV-NPF-SMC-PHD Filter

```
Input: ${\boldsymbol{m}^i_{k-1}, \omega^i_{k-1}}^{N_{k-1}}_{i=1}$, $N_B$, $\boldsymbol{Z}_k$, $k$ and DOA lines.
Output: ${\tilde{\boldsymbol{m}}^j_k, \tilde{\omega}^j_k}^{\tilde{N}_k}_{j=1}$, and ${\boldsymbol{m}^i_k, \omega^i_k}^{N_k}_{i=1}$.
Initialize: $\boldsymbol{\tau}_k$, $q_k$, $\phi_{k|k-1}$, $p_k$, $\gamma_k$, $\kappa_k$, $P_{D,k}$, $\mathbf{F}_{\tilde{\boldsymbol{m}}}$, $\mathbf{F}_{\boldsymbol{z}}$ and speaker histograms.
Run:
Step 1: Prediction step
Propagate surviving particles ${\boldsymbol{m}^i_{k|k-1}}^{N_{k-1}}_{i=1}$.
Step 2: Particle birth and relocation step by Eq. (4).
if DOA lines exist then
    Concentrate particles around the DOA line by Eq. (6).
    if new speaker then Sample $N_B$ born particles around the DOA line.
Calculate ${\omega^i_{k|k-1}}^{N_{k-1}}_{i=1}$ by Eq. (5) and Eq. (7).
${\boldsymbol{m}^i_{k|k-1}, \omega^i_{k|k-1}}^{N_k}_{i=1}$ = ${\boldsymbol{m}^i_{k|k-1}, \omega^i_{k|k-1}}^{N_{k-1}}_{i=1}$ $\cup$ ${\boldsymbol{m}^i_{k|k-1}, \omega^i_{k|k-1}}^{N_{k-1}+N_B}_{i=N_{k-1}+1}$.
Step 3: Update step
for $i \in [1, ..., N_k]$ do
    Calculate the visual likelihood $\breve{\boldsymbol{h}}^i_k$.
    Calculate the audio likelihood $\mathring{\boldsymbol{h}}^i_k$.
    Calculate the audio-visual likelihood $h^i_k$ by Eq. (8).
    Calculate $\boldsymbol{\nabla} \log h^i_k$ and $\boldsymbol{\nabla}^2 \log h^i_k$.
    for $\lambda \in [0, \triangle\lambda, 2\triangle\lambda, \cdots, N_\lambda\triangle\lambda]$ do
        Evaluate flow $\boldsymbol{f}^i_k$ by Eq. (11).
        Update $\triangle\boldsymbol{m}^i_{k|k-1}$ by Eq. (9) and Eq. (10).
    Re-calculate the particle weights.
Update ${\omega^i_{k|k-1}}^{N_k}_{i=1}$ to obtain ${\omega^i_k}^{N_k}_{i=1}$ by Eq. (13) and calculate $\tilde{N}_k = \sum^{N_k}_{i=1} \omega^i_k$.
Set ${\boldsymbol{m}^i_k}^{N_k}_{i=1}$ as ${\boldsymbol{m}^i_{k|k-1}}^{N_k}_{i=1}$.
Get ${\tilde{\boldsymbol{m}}^j_k, \tilde{\omega}^j_k}^{\tilde{N}_k}_{j=1}$ by the k-means or MEAP method
if ESS $< N_k/2$ then (Optional) Resample ${\boldsymbol{m}^i_k, \omega^i_k}^{N_k}_{i=1}$.
```

of the number of particles. More details about the AV-NPF-SMC-PHD filter can be found in [35].

Although the NPF has successfully mitigated the weight degeneracy problem of the AV-SMC-PHD filter, it was based on the assumption that the speakers are detectable, and each speaker generates only one measurement. This is because the particle flow method was proposed originally for single-target tracking, and adapted for multi-target tracking in the AV-NPF-SMC-PHD filter. Therefore, the particle flow is only calculated by the audio-visual measurements near the particles. The particles associated with the occluded speakers may be incorrectly migrated towards other speakers or clutters by NPF. As illustrated in Fig. 1, there are two speakers, where the speaker with a white shirt occludes another speaker with a yellow shirt. One hundred particles are generated, as shown in blue dots. When $\lambda = 0$ (line 19 of Algorithm 1), the particles are located near the two speakers. With the particle flow, most of the particles are migrated from the occluded speaker towards the detected speaker. When $\lambda = 1$, most of the particles are around the detected speaker and therefore given high weights, which, therefore, mitigates the weight degeneracy problem. However, only a small number of particles are located near the occluded speaker, which can be easily missed by the tracker.

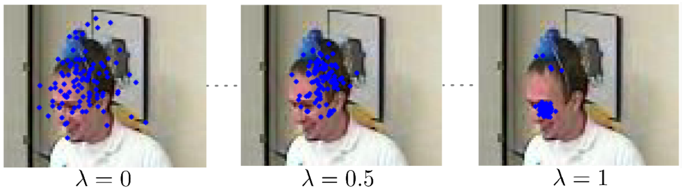

$\lambda = 0$ $\lambda = 0.5$ $\lambda = 1$

Fig. 1. Illustration of the non-zero diffusion particle flow on the frame 376 for Sequence 45 (camera 3) of the AV 16.3 dataset. The particles are shown as the blue dots.

Apart from that, a clustering step, e.g. the k-means or MEAP method, is applied for estimating the speaker state. However, the performance of the k-means is effected by the initialised random seed. Inappropriate choice of initial seeds may degrade the performance of the AV-NPF-SMC-PHD filter.

III. AUDIO-VISUAL LABELLED NON-ZERO DIFFUSION PARTICLE FLOW SMC-PHD FILTER

To address the above problems, we propose a new method based on the particle labels. The idea is to estimate the particle labels based on the likelihood function, and then calculate the particle flow in terms of the selected particles with the same labels. Therefore, the particles associated with detected speakers and undetected speakers are distinguished based on the particle labels. This novel method for calculating the particle flow in terms of particle labels, which we name as AV-LPF-SMC-PHD, offers significant improvement over AV-NPF-SMC-PHD in terms of tracking accuracy. Apart from that, with particle labels, the k-means clustering method (i.e. line 25 of Algorithm 1) can be replaced by the weighted mean of the labelled particles, which is computationally more efficient.

A. Particle label estimation

To accurately identify the particles associated with undetected or occluded speakers, we calculate their labels in the prediction step. We define the particle label as $\boldsymbol{l}_k^i = [a_k^i, v_k^i, t_k^i]^T$, where $a_k^i \in \{0, ...\mathring{N}_k^i\}$, $v_k^i \in \{0, ...\breve{N}_k^i\}$ and $t_k^i \in \{0, ...\tilde{N}_k^i\}$ are the index of audio measurement, visual measurement and speaker associated with the ith particle at frame k, respectively. $a_k^i = 0$ and $v_k^i = 0$ means that the speaker associated with the ith particle is not detected by audio sensor and visual sensor, respectively. Here $t_k^i = 0$ means the ith particle is associated with a new born speaker. t_k^i is calculated when the speaker states are estimated, which will be discussed in Section III-C. For example, $a_k^i = 2, v_k^i = 0, t_k^i = 1$ means that the candidate speaker associated with the ith particle is the first speaker and is only detected by the second audio measurement in time k.

In the prediction step, the labelled particle is represented as $\{\boldsymbol{m}_k^i, \omega_{k-1}^i, \boldsymbol{l}_{k-1}^i\}_{i=1}^{N_{k-1}}$, where a_k^i and v_k^i in $\boldsymbol{l}_{k-1}^i$ are calculated as,

$$a_k^i = H_{\mathring{r}_D^i}(1 - \mathring{P}_{D,k}^i) \arg\max_o (\mathring{r}_k^{i,o} \mathring{h}_{k|k-1}^{i,o}), \tag{15}$$

$$v_k^i = H_{\breve{r}_D^i}(1 - \breve{P}_{D,k}^i) \arg\max_u (\breve{r}_k^{i,o} \breve{h}_{k|k-1}^{i,u}), \tag{16}$$

where H is the Heaviside step function,

$$H_X(Y) = \begin{cases} 0, X \leq Y \\ 1, X > Y \end{cases} \tag{17}$$

where $\mathring{P}_{D,k}^i$ and $\breve{P}_{D,k}^i$ are audio and visual detection probabilities, respectively. $\mathring{r}_D^i$ and $\breve{r}_D^i$ are parameters with values ranging between 0 and 1, which are used to determine whether the particle is detected. For example, if $\mathring{r}_D^i \leq 1 - \mathring{P}_{D,k}^i$, $H_{\mathring{r}_D^i}(1 - \mathring{P}_{D,k}^i) = 0$ and the ith particle is undetected by audio measurement while if $\mathring{r}_D^i > 1 - \mathring{P}_{D,k}^i$, $H_{\mathring{r}_D^i}(1 - \mathring{P}_{D,k}^i) = 1$ and the ith particle is detected by audio measurement. Since both $\mathring{r}_D^i$ and $\breve{r}_D^i$ are valued in terms of uniform distributions, the probabilities for $H_{\mathring{r}_D^i}(1 - \mathring{P}_{D,k}^i) = 1$ and $H_{\breve{r}_D^i}(1 - \breve{P}_{D,k}^i) = 1$ are $1 - \mathring{P}_{D,k}^i$ and $1 - \breve{P}_{D,k}^i$, respectively. $\mathring{r}_k^{i,o}$ and $\breve{r}_k^{i,o}$ are also random values from 0 to 1 and they are used to select the audio and visual measurements associated with the ith particle. The term $\arg\max_o(\mathring{r}_k^{i,o} \mathring{h}_{k|k-1}^{i,o})$ means that the index of the audio measurement giving the highest likelihood is selected as the audio label. The visual label is obtained in a similar way. The audio and visual detection probabilities are defined as,

$$\mathring{P}_{D,k}^i = \sum_{\mathring{\boldsymbol{z}}_k^o \in \mathring{\boldsymbol{Z}}_k} \frac{\mathring{P}_{D,k-1}^i \mathring{h}_{k|k-1}^{i,o}}{\mathring{\kappa}_k(\mathring{\boldsymbol{z}}_k^o) + \sum_{i=1}^{N_k} \mathring{h}_{k|k-1}^{i,o}}, \tag{18}$$

$$\breve{P}_{D,k}^i = \sum_{\breve{\boldsymbol{z}}_k^o \in \breve{\boldsymbol{Z}}_k} \frac{\breve{P}_{D,k-1}^i \breve{h}_{k|k-1}^{i,u}}{\breve{\kappa}_k(\breve{\boldsymbol{z}}_k^u) + \sum_{i=1}^{N_k} \breve{h}_{k|k-1}^{i,u}}, \tag{19}$$

where $\mathring{h}_{k|k-1}^{i,o}$ and $\breve{h}_{k|k-1}^{i,u}$ are the oth audio and uth visual likelihood of the ith particle $\boldsymbol{m}_{k|k-1}^i$, respectively. The audio and visual likelihood densities can be calculated by different detectors, such as face detector or body detector. In this paper, the audio and visual likelihood density are given by

$\mathcal{N}(\mathring{\mathbf{F}}_{\boldsymbol{z}}(\boldsymbol{m}^i_{k|k-1})|\mathring{\boldsymbol{z}}^o_k, \mathring{\boldsymbol{\Psi}})$ and $\mathcal{N}(\breve{\mathbf{F}}_{\boldsymbol{z}}(\boldsymbol{m}^i_{k|k-1})|\breve{\boldsymbol{z}}^u_k, \breve{\boldsymbol{\Psi}})$, respectively, where $\mathring{\boldsymbol{z}}^o_k$ is calculated in terms of the direction of arrival (DOA) of the speakers [49] and $\breve{\boldsymbol{z}}^u_k$ is calculated by a face detector [50], since the face detector is not affected by the clothes on the body and the centre of the face is close to position of sound sources (i.e. around the mouth region). An advantage with the use of the PHD filter framework is that it can work with measurements obtained by a variety of detection methods, either conventional methods or state-of-the-art deep learning methods [51]. The index of audio and visual measurement with high likelihood density has a high probability to be set as a^i_k and v^i_k, as proved in Appendix.

The new born particles are created based on measurements and hence they are only used to represent the detected speakers. Therefore, for born particles, we have $\mathring{P}^i_{D,k} = 1$ and $\breve{P}^i_{D,k} = 1$, and a^i_k and v^i_k are calculated as,

$$a^i_k = \arg\max_o (r(1)\mathring{h}^{i,o}_{k|k-1}), \tag{20}$$

$$v^i_k = \arg\max_u (r(1)\breve{h}^{i,u}_{k|k-1}). \tag{21}$$

Fig. 2 represents the label space of $\boldsymbol{l}^i_k \in \mathbb{R}^{(\mathring{N}_k+1)\times(\breve{N}_k+1)\times(\tilde{N}_k+1)}$. Each point in this space represents a candidate speaker. For each layer t^i_k, there are four areas denoted by a, b, c and d. The particles in the area a, b and c are associated with the speakers detected by audio-visual measurement, visual measurement and audio measurement, respectively. The particles in the area d is associated with the undetected speaker.

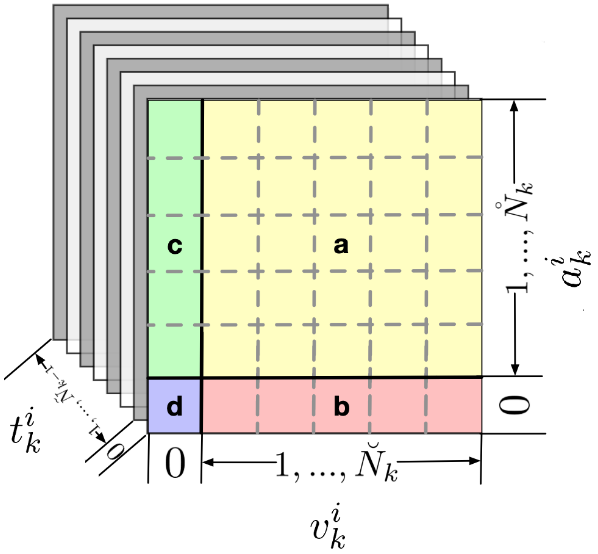


Fig. 2. Illustration of the label space. The yellow area (a) represents the speakers detected by the audio measurement and visual measurement. The red area (b) represents the speakers detected only by the visual measurement while the green area (c) represents the speakers detected only by the audio measurement. The blue area (d) represents the speakers undetected.

B. AV labelled non-zero diffusion particle flow

In the AV-NPF-SMC-PHD filter, it is assumed that all speakers are detected, however, this can be violated when occlusion happens. To address this problem, the audio-visual likelihood is defined based on the particle labels which are then used to calculate the particle flow, leading to the proposed LPF method. We use different strategies to update born particles, survival particles associated with detected speakers and survival particles associated with undetected speakers. This helps reduce the computational cost of LPF. Since the born particles ($t^i_{k-1} = 0$) are created based on measurements and the importance density at k, the posterior density is only calculated by the likelihood density and importance density. The particle flow will not be used to update the new born particles.

For survival particles associated with detected speaker, i.e. $t^i_{k-1} > 0$ and $a^i_k + v^i_k > 0$ as shown in Section III-A, where a^i_k and v^i_k are the index of the audio and visual measurements associated with the ith particle, respectively, the audio and visual likelihood densities of the ith particle, i.e. $\mathring{h}^{i,o}_{k|k-1}$ and $\breve{h}^{i,o}_{k|k-1}$ are normalised based on the particle labels, as follows,

$$\mathring{h}^i_k = \frac{\mathring{h}^{i,a^i_k}_{k|k-1}}{\mathring{\kappa}^i_k + \sum_{i'=1}^{N_k+N_B} \delta_{a^i_k}(a^{i'}_k)\mathring{h}^{i',a^i_k}_{k|k-1}\omega^{i'}_{k|k-1}}, \tag{22}$$

$$\breve{h}^i_k = \frac{\breve{h}^{i,v^i_k}_{k|k-1}}{\breve{\kappa}^i_k + \sum_{i'=1}^{N_k+N_B} \delta_{v^i_k}(v^{i'}_k)\breve{h}^{i',v^i_k}_{k|k-1}, \omega^{i'}_{k|k-1}} \tag{23}$$

where δ is the Dirac delta function.

$$\delta_X(Y) = \begin{cases} 1, X = Y \\ 0, X \neq Y \end{cases} \tag{24}$$

The function δ is used to select the i'th particle associated with the same measurements as the ith particle. The audio and visual clutter densities of the ith particle are denoted by $\mathring{\kappa}^i_k$ and $\breve{\kappa}^i_k$, respectively. The audio and visual measurement densities are defined as $\sum_{i'=1}^{N_k+N_B} \delta_{a^i_k}(a^{i'}_k)\mathring{h}^{i',a^i_k}_{k|k-1}\omega^{i'}_{k|k-1}$ and $\sum_{i'=1}^{N_k+N_B} \delta_{v^i_k}(v^{i'}_k)\breve{h}^{i',v^i_k}_{k|k-1}, \omega^{i'}_{k|k-1}$. Based on the novel audio and visual likelihood densities, the labelled particle flow $\boldsymbol{f}^i_k$ is calculated by the partial differential equation [45]:

$$\boldsymbol{f}^i_k = -[-(\boldsymbol{P}^i_{k|k-1})^{-1} + \lambda\boldsymbol{\nabla}^2 \log h^i_k]^{-1}(\boldsymbol{\nabla} \log h^i_k), \tag{25}$$

where

$$h^i_k = (1 - H_0(a^i_k) + H_0(a^i_k)\mathring{h}^i_k)(1 - H_0(v^i_k) + H_0(v^i_k)\breve{h}^i_k). \tag{26}$$

Applying Eq. (17) to $H_0(a^i_k)$ and $H_0(v^i_k)$, Eq. (26) can be simplified as,

$$h^i_k = \begin{cases} \mathring{h}^i_k\breve{h}^i_k & \text{, if } a^i_k > 0 \text{ and } v^i_k > 0. \\ \mathring{h}^i_k & \text{, if } a^i_k > 0 \text{ and } v^i_k = 0. \\ \breve{h}^i_k & \text{, if } a^i_k = 0 \text{ and } v^i_k > 0. \\ 0 & \text{, if } a^i_k = 0 \text{ and } v^i_k = 0. \end{cases} \tag{27}$$

Based on the definition of weighted covariance, the matrix $\boldsymbol{P}^i_{k|k-1}$ of the ith particle is calculated as

$$\boldsymbol{P}^i_{k|k-1} = \frac{\sum_{i1=1}^{N_k} s^{i,i1}_k[\omega^i_{k|k-1}\boldsymbol{e}(\boldsymbol{m}^i_{k|k-1})\boldsymbol{e}(\boldsymbol{m}^i_{k|k-1})^T]}{\sum_{i1=1}^{N_k} s^{i,i1}_k(t^{i1}_k)\omega^i_{k|k-1}}, \tag{28}$$

where

$$s_k^{i,i'} = \delta_{a_k^i}(a_k^{i'})\delta_{v_k^i}(v_k^{i'})\delta_{t_{k-1}^i}(t_{k-1}^{i'}). \tag{29}$$

$$\boldsymbol{e}(\boldsymbol{m}_{k|k-1}^i) = \boldsymbol{m}_{k|k-1}^i - \frac{\sum_{i'=1}^{N_k} s_k^{i,i'}\left(\omega_{k|k-1}^{i'}\boldsymbol{m}_{k|k-1}^{i'}\right)}{\sum_{i'=1}^{N_k} s_k^{i,i'}\omega_{k|k-1}^{i'}}. \tag{30}$$

where $\frac{\sum_{i'=1}^{N_k} s_k^{i,i'}\left(\omega_{k|k-1}^{i'}\boldsymbol{m}_{k|k-1}^{i'}\right)}{\sum_{i'=1}^{N_k} s_k^{i,i'}\omega_{k|k-1}^{i'}}$ is the centre state of the group of the particles which have the same labels as particles $\boldsymbol{m}_{k|k-1}^i$. Eq. (29) is used to select the particles whose visual, audio and speaker labels are the same, i.e. forming a same group. Eq. (30) represents the difference between the centre states of the group and the individual particle states. In the particle flow, the particle weight is modified as [36],

$$\omega_{k|k-1}^i \Leftarrow \frac{\phi\left(\boldsymbol{m}_k^i|\boldsymbol{m}_{k-1}^i\right)|\det(\boldsymbol{I}+\nabla\lambda\nabla f)|}{\phi\left(\boldsymbol{m}_{k|k-1}^i|\boldsymbol{m}_{k-1}^i\right)}\omega_{k|k-1}^i. \tag{31}$$

For the survival particles associated with undetected speakers i.e. $t_{k-1}^i > 0$, $a_k^i = 0$ and $v_k^i = 0$, there is no measurement associated with the particles. In this case, $P_{D,k}^i$ is 0 and the particle weight remains unchanged.

After the particle flow, the particle weights $\omega_{k|k-1}^i$ are updated to ω_k^i as follows,

$$\omega_k^i = \begin{cases} \omega_{k|k-1}^i, & \text{if } t_{k-1}^i > 0 \text{ and } a_k^i + v_k^i = 0 \\ h_k^i\omega_{k|k-1}^i, & \text{otherwise} \end{cases} \tag{32}$$

Based on Eq. (32), the particles associated with the undetected survival speakers retain the weights from the previous time frames, while the weights of the other particles are updated with our proposed likelihood function.

C. Estimating speaker states

In the baseline AV-NPF-SMC-PHD filter, the k-means algorithm was applied to estimate the speaker states following the update step. However, k-means often converges to local optimum and the result is sensitive to the choice of the initial seeds. Inappropriate initial seeds may degrade the estimation accuracy, especially for estimating the states of the occluded speakers. In addition, the clustering step results in increased computational cost. Here, in our proposed method, inspired by the weighted mean method of particle filters, we estimate the speaker states by weighting the particles in terms of their labels instead of applying the k-means clustering. With the label information, the weighted mean can be calculated directly without using the k-means algorithm. Compared to the k-means algorithm, grouping particles based on labels offers more accurate results in state estimation, since the visual and audio labels consider the information from the measurements and the speaker label considers the information from historical states. Apart from that, we have proposed a label update strategy for state estimation, where some particles are not used for calculating the weighted mean if the speaker to which these particles correspond is considered as moving out of the view of the camera.

As shown in Fig. 2, the particles with the same particle labels are used to estimate the same candidate speaker. However, since the measurement associated with survival particles is also used to create new born particles by Eq. (6), the state of the candidate speaker which has been estimated by survival particles may be repeatedly estimated by the birth particles. Therefore, we calculate the state of the candidate speaker associated with the ith particle with the set of particles, $\{\boldsymbol{m}_k^{i'}, \omega_k^{i'}\}_{i'\in\boldsymbol{\Lambda}(i)}$, where $\boldsymbol{\Lambda}(i)$ is a subset of $[1,\cdots,N_k+N_B]$.

For the ith survival particle, the particle set $\{\boldsymbol{m}_k^{i'}, \omega_k^{i'}\}_{i'\in\boldsymbol{\Lambda}(i)}$ is determined as follows,

$$H_{t_{k-1}^{i'}}(-1)(\delta_0(H_{i'}(N_k))s_k^{i,i'}+\delta_1(H_{i'}(N_k))\delta_{v_k^i}(v_k^{i'})\delta_{a_k^i}(a_k^{i'})) = 1, \tag{33}$$

where $H_{t_{k-1}^{i'}}(-1)$ means that the particle has not yet been used. If the i'th particle is a survival particle, we have $i' \leq N_k$, $\delta_0(H_{i'}(N_k)) = 1$ and $\delta_1(H_{i'}(N_k)) = 0$, while if the i'th particle is a new born particle, we have $i' > N_k$, $\delta_0(H_{i'}(N_k)) = 0$ and $\delta_1(H_{i'}(N_k)) = 1$. When the particle label set of the i'th particle and that of the ith particle are identical, $s_k^{i,i'}$ is equal to 1.

For birth particles ($i > N_k$), the particle set is determined via Eq. (34),

$$H_{t_{k-1}^{i'}}(-1)\delta_1(H_{i'}(N_k))\delta_{v_k^i}(v_k^{i'})\delta_{a_k^i}(a_k^{i'}) = 1. \tag{34}$$

Finally, $\{\boldsymbol{m}_k^{i'}, \omega_k^{i'}\}_{i'\in\boldsymbol{\Lambda}(i)}$ is the set of the survival particles and the born particles that have the same label as the i'th particle. To avoid the same i'th particle being repeatedly selected into the different sets $\boldsymbol{\Lambda}(i)$, $t_{k-1}^{i'}$ is set as -1 after Eq. (33) and Eq. (34), meaning that this particle has already been used. The state of the candidate speaker is estimated with the particle set as weighted states,

$$\tilde{\boldsymbol{m}}_k^j = \frac{\sum_{i'\in\boldsymbol{\Lambda}(i)}\omega_k^{i'}\boldsymbol{m}_k^{i'}}{\tilde{\omega}_k^j}, \tag{35}$$

where

$$\tilde{\omega}_k^j = \sum_{i'\in\boldsymbol{\Lambda}(i)}\omega_k^{i'}, \tag{36}$$

where $\tilde{\omega}_k^j$ is the weight of the candidate speaker. If the weight is lower than a threshold ξ ($0 < \xi < 1$), the state will be considered as corresponding to noise or clutters, otherwise, corresponding to the speaker. When the noise level of the measurements is high, ξ should be set as a low value. In our experiment, we set ξ as 0.5. The labels of the speakers are set as $a_k^j = a_k^i$ and $v_k^j = v_k^i$. Finally, the visual detection probability is updated as follows,

$$\overset{\circ}{P}{}_{D,k}^i = \min\left(\frac{w_k^i l_k^i}{w_{k-1}^i l_{k-1}^i}, 1\right), \tag{37}$$

where (w_k^i, l_k^i) are the weight and length of the bounding box of the ith particle at time k, respectively. The $w_k^i l_k^i$ and $w_{k-1}^i l_{k-1}^i$ are the areas of the ith particle at k and $k-1$, respectively. When the speaker associated with the ith particle is occluded, the area of the bounding box will be degraded and $\overset{\circ}{P}{}_{D,k}^i$ may become smaller than 1. However, when the speaker moves away from the camera, the bounding box size decreases, which may be incorrectly classified as the occlusion. To avoid this problem, the visual detection probability is updated in terms of the aspect ratio of the bounding box,

$$\mathring{P}_{D,k}^{i} = \frac{\min(\frac{l_k^i}{w_k^i}, \frac{l_{k-1}^i}{w_{k-1}^i})}{\max(\frac{l_k^i}{w_k^i}, \frac{l_{k-1}^i}{w_{k-1}^i})}. \tag{38}$$

When the aspect ratio of the bounding box changes sharply, the speaker associated with the ith particle is occluded. Since the speakers walk frequently towards or away from the cameras in the AV16.3 dataset, we use Eq. (38) in our experiments. Finally, t_k^i is set as the index of speaker j. The pseudo-code of the AV-LPF-SMC-PHD filter is presented in Algorithm 2.

Algorithm 2 AV-LPF-SMC-PHD Filter

```
1:  **Input:** ${\boldsymbol{m}_{k-1}^i, \omega_{k-1}^i, \boldsymbol{l}_{k-1}^i, \mathring{P}_{D,k-1}^i, \breve{P}_{D,k-1}^i}_{i=1}^{N_{k-1}}$, $N_B$, $k$, $\{\mathring{\boldsymbol{z}}_k^o\}_{o=1}^{\mathring{N}_k}$ and $\{\breve{\boldsymbol{z}}_k^u\}_{u=1}^{\breve{N}_k}$.
2:  **Output:** $\{\boldsymbol{m}_k^i, \omega_k^i, \boldsymbol{l}_k^i, \mathring{P}_{D,k}^i, \breve{P}_{D,k}^i\}_{i=1}^{N_k}$ and $\{\tilde{\boldsymbol{m}}_k^j, \tilde{\omega}_k^j, \tilde{\boldsymbol{l}}_k^j\}_{j=1}^{\widetilde{N}_k,}$
3:  **Initialize:** $\{\boldsymbol{\Upsilon}_k$, $\phi_{k|k-1}$, $p_k$, $\gamma_k$, $\breve{\boldsymbol{\Psi}}$ and $\mathring{\boldsymbol{\Psi}}.\}$
4:  **Run:**
5:  $N_k = N_{k-1}$
6:  Predict survival particles and create birth particles as in lines 5-11 of Algorithm 1.
7:  **for** $i \in \{N_k + 1, ..., N_k + N_B\}$ **do**
8:      Calculate $a_k^i$ and $v_k^i$ by Eq. (20) and Eq. (21), receptively.
9:      Set $t_{k-1}^i = 0$ and $\omega_k^i = h_k^i \omega_{k|k-1}^i$.
10: **for** $i \in \{1, ..., N_k + N_B\}$ **do**
11:     Calculate $a_k^i$ and $v_k^i$ by Eq. (15) and Eq. (16), receptively.
12:     **if** $t_{k-1}^i > 0$ and $a_k^i + v_k^i > 0$ **then**
13:         Calculate the audio-visual likelihood $h_k^i$ by Eq. (26).
14:         Calculate the covariance matrix by Eq. (28).
15:         **for** $\lambda \in [0, \triangle\lambda, 2\triangle\lambda, \cdots, N_\lambda \triangle\lambda]$ **do**
16:             Evaluate flow $\boldsymbol{f}_k^i$ by Eq. (25).
17:             Update $\triangle \boldsymbol{m}_{k|k-1}^i$ by Eq. (9) and Eq. (10).
18:         Re-calculate the particle weights by Eq. (31).
19:     Update particle weights by Eq. (32).
20: $j = 1$
21: **for** $i \in \{1, ..., N_k + N_B\}$ **do**
22:     **if** $t_{k-1}^i \neq -1$ **then**
23:         Select the particle set $\{\boldsymbol{m}_k^{i'}, \omega_k^{i'}\}_{i' \in \boldsymbol{\Lambda}(i)}$ by Eq. (33) and (34).
24:         Estimate the speaker weight by Eq. (36).
25:         **if** $\tilde{\omega}_k^j > \xi$ **then**
26:             Estimate the speaker state by Eq. (35).
27:             Set $\tilde{a}_k^j = a_k^i$, $\tilde{v}_k^j = v_k^i$ and $\tilde{t}_k^j = j$.
28:             Set $t_{k-1}^{i'} = -1$ for $i' \in \boldsymbol{\Lambda}(i)$.
29:             $j = j + 1$
30: **if** ESS $< N_k/2$ **then**
31:     {(Optional) Re-sample $\{\boldsymbol{m}_k^i, \omega_k^i\}_{i=1}^{N_k}$.}
```

IV. Experimental Evaluations

This section presents experimental evaluations of the proposed algorithms as compared with baseline algorithms. We start with a description of the experimental setup, datasets and performance metrics, before giving the analysis and comparison of the results.

A. Datasets and baselines

Several audio-visual datasets are publicly available, such as the AV16.3 [52], AVDIAR [53], AVTRACK-1 [54], AVASM [55], AMI [56], CLEAR [57], MVAD [58] and SPEVI [59]. We consider our requirements for choosing the datasets. 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 [52] consists of real-world data with both audio and video sequences. It provides the calibration information of the cameras to map the audio data 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 is recorded by three calibrated video cameras at 25 Hz, and each image frame has 288x360 pixels. The audio signals are recorded by two circular eight-element microphone arrays at a sample rate of 16 kHz. The audio and video streams are synchronised before running the algorithms. 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 [53] 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. The AVDIAR dataset provides training data to learn a mapping as in [53]. This dataset includes 23 sequences. Each image frame has 1920x1200 pixels. The audio and video were recorded at 48 kHz and 25 Hz, respectively, which were synchronised by an external trigger controlled by software. There are 12 different participants, and up to 4 people are recorded in each sequence.

AVTRACK-1 [54] and AVASM [55] are provided by the same institution as for AVDIAR. However, they are less challenging than AVDIAR. AMI [56] and MVAD [58], which are designed for speaker diarization, are not used in our tests since the speakers are mostly static or with small movements. In SPEVI [59], 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 [57] is chosen for our experiments since it has the largest number of speakers among these datasets.

Several baselines are considered for benchmarking our proposed algorithms, including the SMC-PHD filter [60], SAVMS-SMC-PHD filter [60], ZPF-SMC-PHD filter [20], NPF-SMC-PHD filter [35] and ZPF-delta-GLMB [38]. For convenience, the SMC-PHD, SAVMS-SMC-PHD, ZPF-SMC-PHD, NPF-SMC-PHD, ZPF-delta-GLMB and our proposed LPF-SMC-PHD filters are abbreviated as SMC, SAVMS, ZPF, NPF, GLMB and LPF, respectively. Their input measurements are same. Apart from that, the step in LPF for estimating the speaker state is compared to the k-means and MEAP clustering methods to show the improvement achieved by labelling. Finally, we include deep learning baselines, i.e. YoloV5-DeepSort and YoloV5-StrongSort [19], to demonstrate that our proposed method offers competitive performance, even though it does not involve model training.

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 [61] is defined as,

$$\text{OSPA}(\{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k}, \{\tilde{\mathfrak{m}}_k^{\tilde{j}}\}_{\tilde{j}=1}^{\tilde{\mathfrak{N}}_k}) = \sqrt[a]{\frac{\min\limits_{\pi \in \Pi_{\tilde{\mathfrak{N}}_k, \tilde{N}_k}} \sum_{j=1}^{\tilde{N}_k} \overline{d}^{(c)}(\tilde{\boldsymbol{m}}_k^j, \tilde{\mathfrak{m}}_k^{\pi(j)})^a + c^a(\tilde{\mathfrak{N}}_k - \tilde{N}_k)}{\tilde{\mathfrak{N}}_k}}, \quad (39)$$

where $\{\tilde{\mathfrak{m}}_k^1, ..., \tilde{\mathfrak{m}}_k^{\tilde{\mathfrak{N}}_k}\}$ are the ground truth speaker states, and $\{\tilde{\boldsymbol{m}}_k^1, ..., \tilde{\boldsymbol{m}}_k^{\tilde{N}_k}\}$ are 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. (39) is for $\tilde{N}_k \leq \tilde{\mathfrak{N}}_k$.

If $\tilde{\mathfrak{N}}_k < \tilde{N}_k$, then $\text{OSPA}(\{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k}), \{\tilde{\mathfrak{m}}_k^{\tilde{j}}\}_{\tilde{j}=1}^{\tilde{\mathfrak{N}}_k}) = \text{OSPA}(\{\tilde{\mathfrak{m}}_k^{\tilde{j}}\}_{\tilde{j}=1}^{\tilde{\mathfrak{N}}_k}, \{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k}))$. The function $\bar{d}^{(c)}(\cdot)$ is defined as $\min(c, \bar{d}(\cdot))$ where c is the cut-off value that considers 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 indicates a better tracking performance.

To assess the weight degeneracy problem, ESS is often used [36], [20], [31], defined as

$$\text{ESS} = \frac{(\sum_{i=1}^{N_k} \omega_k^i)^2}{\sum_{i=1}^{N_k} (\omega_k^i)^2}. \quad (40)$$

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

The performance of the methods for associating the particles with the speakers is normally evaluated by Silhouette Coefficient [62], R-Square [63] and Improved Hubert Γ Statistic [64]. However, when the speakers are occluded, the above metrics, which are based on distances between the particles, can become inaccurate since the particles associated with the speakers are overlapping with each other. In this work, we use Root Mean Squared Error (RMSE) to evaluate the performance in terms of the fitting of the posterior densities with a Gaussian distribution. This is because the likelihood and prior densities of the speakers are both Gaussian [65]. RMSE can represent the homogeneity score of the particle sets. Since different methods may give different estimates for the number of speakers, we use the mean RMSE of all the particles, which provides a more robust measure of the performance,

$$RMSE = \frac{\sum_{i'}^{N_k+N_B} \sqrt{\frac{\sum_{i \in \boldsymbol{\Lambda}(i')} \left\| \omega_{k|k-1}^i \boldsymbol{m}_{k|k-1}^i - \boldsymbol{e}(\boldsymbol{m}_{k|k-1}^i) \right\|_2}{\|\boldsymbol{\Lambda}(i')\|_1}}}{N_k + N_B}, \quad (41)$$

where $\|.\|_1$ and $\|.\|_2$ are the L_1 and L_2 norm, respectively.

C. Parameter settings

Compared to the baseline methods NPF and ZPF, LPF uses a smaller number of parameters and thresholds. In this section, the setting of particle flow, including the threshold Ξ is discussed, and other parameters are given as in the baseline methods.

The initial distributions of the particles are randomly sampled in the image frame. When the particles move out of the image frame, they will be removed from the particle set. The transition model is defined as

$$\mathbf{F}_{\tilde{\boldsymbol{m}}} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (42)$$

The DOAs of the speakers are used as the audio information in our experiments. It can be obtained by either a circular array (as in AV16.3) or a linear array. As the DOAs are determined by the relative delay between the pairs of the microphone signals [60], it shows only the approximate direction θ_k^o of the sound sources with respect to the microphones. The rectangular coordinate $[x_k^o, y_k^o]$ of $\mathring{\boldsymbol{m}}_k^o$ can be transformed to polar coordinate $[r_k^o, \theta_k^o]$, where r_k^o is the Euclidean distance from the state of the nearby speaker at the previous frame to the microphone position,

$$r_k^o = \left\| [\tilde{x}_{k-1}^{\hat{j}}, \tilde{y}_{k-1}^{\hat{j}}]^T - [x_{mic}, y_{mic}]^T \right\|_2, \quad (43)$$

where

$$\hat{j} = \arg\min_j \left\| \frac{\tilde{y}_{k-1}^j - y_{mic} - \tan\theta_k^o(\tilde{x}_{k-1}^j - x_{mic})}{\tilde{x}_{k-1}^j - x_{mic} - \tan\theta_k^o(\tilde{y}_{k-1}^j - y_{mic})} \right\|_1, \quad (44)$$

where $[\tilde{x}_{k-1}^j, \tilde{y}_{k-1}^j]^T$ and $[x_{mic}, y_{mic}]^T$ are the positions of the jth speaker at the previous frame and the position of the microphone array $\boldsymbol{m}_{mic}$, respectively. Face detection is used to provide the visual information in our experiment. The visual measurement model is defined as,

$$\check{\mathbf{F}}_{\boldsymbol{z}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (45)$$

For survival particles, the proposal distribution $q_k\left(\boldsymbol{m}^i_{k|k-1}|\boldsymbol{m}^i_{k-1},\boldsymbol{Z}_k\right)$ and the transition distribution $\phi(\boldsymbol{m}^i_{k|k-1}|\boldsymbol{m}^i_{k-1})$ are both simplified as $\mathcal{N}(\boldsymbol{m}^i_{k|k-1}|\mathbf{F}_{\tilde{\boldsymbol{m}}}\boldsymbol{m}^i_{k-1},\boldsymbol{\Upsilon}_k)$.

For birth particles, when the number of detected faces or DOA lines is greater than that of the estimated speakers N_{k-1}, LPF considers there are new speakers appearing. The birth density $\gamma_k(\boldsymbol{m}^i_{k|k-1})$ is given as

$$\gamma_k(\boldsymbol{m}^i_{k|k-1}) = \begin{cases} 1, & \text{if } x^i_{k|k-1} < 60 \text{ or } x^i_{k|k-1} > 300 \\ 0.8, & \text{otherwise} \end{cases} \tag{46}$$

The new speakers often appear at the edge of the image frame, and the birth density at the edge has a high value. The number of new birth particles N_B is set as 50 for each speaker. For particle flow, the pseudo time λ is increased incrementally from 0 to 1, with a step size $\Delta\lambda$, set as $\Delta\lambda = 0.01$ in our experiment.

The threshold ξ is used to detect the candidate speaker with a high weight. Different values are tested from 0 to 1 in the frames 200-320 and frames 870-910 for sequence 24 (camera 3) and frames 200-400 and frames 500-900 for sequence 45 (camera 3). In sequence 45, there are three speakers, while in sequence 24, there are two speakers. The occlusion mainly happens in frames 870-910 for sequence 24 and frames 500-900 for sequence 45 (camera 1), and the speakers are undetected in other frames. The results are shown in Table I. The value of ξ does not affect the running time of the proposed algorithm. When ξ has a low value such as 0.1, LPF over-estimates the number of the estimated speakers than the ground truth, since some clutters are estimated as the speaker. LPF would under-estimate the number of speakers than the ground truth when ξ has a high value such as 0.9, since candidate speakers with noisy measurements are not estimated as speakers. If the number of speakers is accurately estimated, our proposed method would give the lowest OSPA. At the frames 200-320 of sequence 24 and frames 200-400 of sequence 45, speakers are away from each other, OSPA gives the lowest value and the number of speakers can be accurately estimated at $\xi = 0.5$. When the occlusion frequently happens e.g. at frames 500-900 of sequence 45 and frames 870-910 of sequence 24, the weight of the occluded speaker becomes small and our proposed method offers the lowest OSPA at $\xi = 0.4$. Apart from that, Table I shows that the computational cost is not affected by the value of ξ. Therefore, it is reasonable to set ξ between 0.4 and 0.5 for AV16.3. Since the occlusion does not frequently happen in AV16.3, ξ is set as 0.5 in our experiment.

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 [34], [20], [43]. 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.

TABLE I
RUNNING TIME (S) AND OSPA OF LPF VERSUS ξ ON THE AV 16.3.

ξ	Sequence 24		Sequence 45		
	200-320	870-910	200-400	500-900	
0.1	17.4	5.8	44.2	8.8	time (s)
	24.8	25.8	36.1	38.1	OSPA
	3.6	2.9	4.9	4.2	N_k
0.3	17.4	5.8	44.2	8.8	time (s)
	19.5	20.4	32.9	34.2	OSPA
	2.9	2.6	4.3	3.7	N_k
0.4	17.4	5.8	44.2	8.8	time (s)
	17.2	**17.6**	31.4	**31.6**	OSPA
	2.1	**2.1**	3.4	**3.2**	N_k
0.5	17.4	5.8	44.2	8.8	time (s)
	16.8	19.6	**29.8**	33.7	OSPA
	2.0	1.8	**3.2**	2.5	N_k
0.7	17.4	5.8	44.2	8.8	time (s)
	18.6	21.8	30.8	35.8	OSPA
	1.8	1.7	3.0	2.4	N_k
0.9	17.4	5.8	44.2	8.8	time (s)
	23.8	27.7	36.7	38.9	OSPA
	1.5	0.8	2.4	2.0	N_k

D. Comparison with the baseline methods

In this subsection, we show the improvement achieved by our proposed ideas using frames from an example sequence, i.e. the frames 630-700 of Sequence 45 (camera 3), as they contain some challenging situations.

First, we compare our proposed particle label estimation method with the two particle-speaker association baselines, i.e. the particle-speaker association method in NPF using the measurements near the particles, and the particle-speaker association method in delta-GLMB based on all the measurements. Second, this section compares our proposed labelled particle flow, with two baseline methods, i.e. zero diffusion particle flow and non-zero diffusion particle flow. Third, this section compares our state estimation method based on the particle labels, with the k-means and and MEAP clustering methods for state estimation.

1) Labelling particles: In this section, we compare the particle label estimation method in LPF, with the particle-speaker association methods used in NPF and delta-GLMB. Fig. 3 shows an example for Sequence 45 (camera 3), where at frame 640, a speaker with a black shirt begins walking into the frame, and the speaker with a white shirt then occludes the speaker with a yellow shirt. The initial particles are spread randomly and in the same way for all the filters.

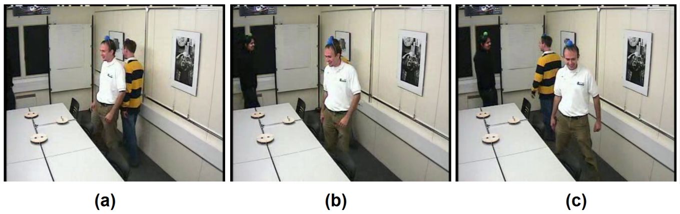

(a) (b) (c)

Fig. 3. Frames 640 (a), 660 (b) and 670 (c) for Sequence 45 (camera 3) in the AV16.3 dataset. There are three speakers. One of them is walking into the scene and the speaker with the yellow shirt is occluded in frame 660.

Fig. 4 shows the particle labels estimated for the frame 660 for Sequence 45 (camera 3) where the speaker in yellow shirt is occluded. Note, the face images were cropped manually from the video signal to visualise the distribution of the particles around the face area, which is very small in the whole image plane. The particles of the new speaker with the black shirt can be accurately labelled by the filters as it only has a nearby measurement, which is not shown in Fig. 4. For better visualisation, we only plot the particles with high weights. The green asterisks and red asterisks show the particles associated with the speaker with a white shirt and the occluded speaker with yellow shirt, respectively. Since the speaker in yellow shirt is occluded by another speaker, only one face can be detected in frame 660. In NPF, the particle flow is calculated with the audio-visual measurements near the particle, and the particles are classified based on the distance from the particles to the measurements. Therefore, the particles associated with the occluded speaker are labelled as the particles associated with the front speaker, and as a result, there are only green asterisks in Fig. 4(a). In GLMB (Fig. 4(c)), the labels of the particles are given when they are created, and the undetected speaker is considered. Therefore, nearly half of the particles are shown as the red asterisks, and there is a clear boundary between the two groups of particles. In our proposed LPF (Fig. 4.b), due to the random values in Eq. (16) and Eq. (15), the particles in green and the particles in red are mixed. As a result, most of the particles can still track their associated speakers. Fig. 4 shows that the proposed LPF provides a better data association than GLMB. In this example, one speaker is occluded by another, and their face states should be similar. However, for GLMB, the centre of the particles corresponding to the front speaker is lower than the actual position of their face, while in LPF, the particle distributions of the two speakers are mixed, meaning that the estimated states of these two speakers are closer to each other, thus the LPF provides a better fit with the ground truth.

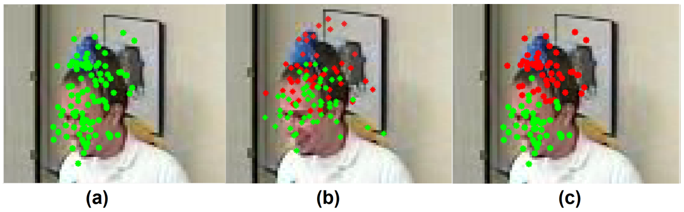


Fig. 4. The labelled particles of NPF (a), LPF (b) and GLMB (c) at frame 660 for Sequence 45 (camera 3). The particles of the front speaker and occluded speaker are shown as the green and red asterisks, respectively.

In Fig. 5, we show the RMSE of NPF, GLMB and LPF, respectively, for frames 630-700 of Sequence 45 (camera 3). The initial particle sets for the three filters are the same at frame 630. Apart from that, RMSE is calculated before the update step, such as in line 11 of Algorithm 2 and line 13 of the Algorithm 1. The only factor that affects the RMSE is the method for associating the particles of these filters. For better visualisation, we only plot $\log(\text{RMSE})$.

In the beginning, the RMSE of the compared filters is similar. Since the speakers have a long distance to other speakers, the label of particles can be accurately estimated by the nearby measurements in NPF. At frames 650 - 660, the occlusion happens and $\log(\text{RMSE})$ of NPF increases to -12.70, since NPF can not accurately label the particles for the occluded speaker. However, with our proposed LPF, $\log(\text{RMSE})$ is about -18.89, resulting in a 48% performance improvement over NPF thanks to the particle labels estimated by Eq. (15) and Eq. (16). GLMB gives an RMSE similar to LPF, as GLMB can estimate the particle labels accurately based on the Bernoulli filter.

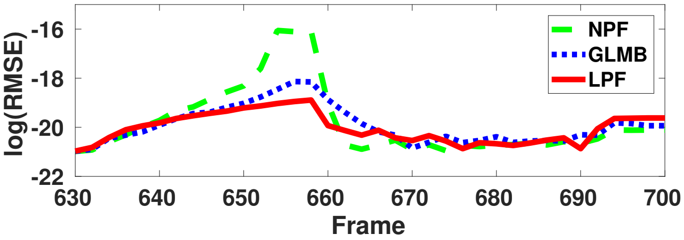


Fig. 5. $\log(\text{RMSE})$ of NPF, GLMB and LPF at the frames 630-700 for Sequence 45 (camera 3).

2) Particle flows: To evaluate the particle flow, ZPF, NPF, GLMB and LPF are compared. To allow for a fair comparison, the speakers are estimated by the k-means method. Fig. 6 shows how the particles are modified by these filters from $\lambda = 0$ to $\lambda = 1$. The three figures in each row are shown for $\lambda = 0, 0.5$ and 1, respectively. The rows show the tracking results of ZPF, NPF, GLMB and LPF, respectively. The particle flow of ZPF and GLMB is calculated based on the zero diffusion flow, while the particle flow of the NPF and GLMB is calculated based on the non-zero diffusion flow. The green asterisks show the front speaker and the red asterisks show the occluded speaker. Since ZPF and NPF are only calculated based on the measurements near the particles without the label information, there are only green asterisks for ZPF and NPF. Compared to the NPF and ZPF, LPF gives more accurate estimates for the speaker states, and more particles are located nearby the occluded speakers. Although GLMB uses the particle label information to update the particle weights, it assumes the speaker is always detected, and the particle flow is also modified towards the nearby measurements.

In Fig. 7, we show the variation of ESS of NPF, ZPF and LPF from $\lambda = 0$ to $\lambda = 1$ on the frame 660 for Sequence 45 (camera 3). At the beginning of the update step, one of the speakers is occluded, and the filters encounter with the weight degeneracy problem. Using ZPF, NPF, GLMB and LPF, the ESS is increased to 71.9, 71.7, 77.1 and 79.3, respectively. In NPF and ZPF, the particles associated with the occluded speaker are modified towards the front speaker with a white shirt. The number of particles associated with the front speaker is increased and the average weight of these particles is decreased. Therefore, ESS of the NPF and ZPF at $\lambda = 1$ is only increased by 10.6% and 10.3% as compared with the ESS at $\lambda = 0$. Although GLMB uses particle labels to update the particle weights, the particle flow of GLMB updates the particle states with the measurements. The particles of GLMB are modified towards the detected speakers. The improvement of ESS achieved by LPF is the highest among the tested filters. With the label information, the LPF provides an ESS that is

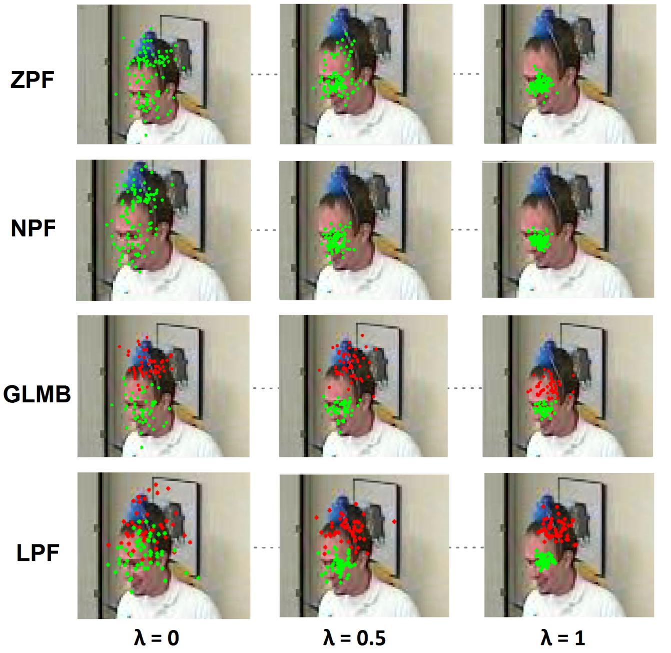


Fig. 6. The motion trails of the particles by ZPF, NPF, GLMB and LPF. The columns show the results for $\lambda = 0, 0.5, 1$ respectively in the frame 660 for Sequence 45.

11% higher than its baseline NPF.

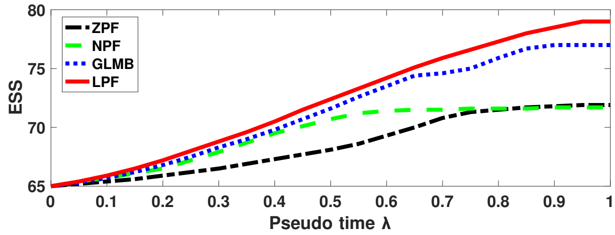


Fig. 7. The ESS and OSPA of ZPF, NPF, GLMB and SMC in the frame 660 for Sequence 45 (camera 3) changes with respect to λ.

Fig. 8 shows the average OSPA for the frames 630-700 of Sequence 45. It can be observed that LPF gives the smallest average OSPA. Due to the presence of occlusion from frames 645 to 660, the OSPAs for NPF and ZPF have increased. OSPA of the LPF remains low in most of the frames. At frame 660, LPF gives an average OSPA at about 13.5, resulting in a 29%, 28% and 6% performance improvement over ZPF, NPF and GLMB, respectively, thanks to the more accurate estimate of the number of speakers offered by the label information, as shown in Fig. 9.

3) Clustering: Here we compare our approach for estimating the speaker states using the labels estimated, with k-means and MEAP clustering, at frame 660 for Sequence 45. We add clutter with a clutter rate ranging from 0 to 40, including both audio clutter and visual clutter. The clutter is randomly distributed over the measurement space. Fig. 10 shows the frame 660 of Sequence 45 with audio clutters and visual clutters. The visual measurements and clutters are shown in the red and blue boxes, respectively, while the audio measurements and clutters are shown in the yellow and green lines, respectively. The number of visual and audio clutters is same for each experiment, and the clutters and measurements are the same for these methods.

Table II shows the OSPA and computational time with respect to different number of clutters. We observe that our proposed method is computationally most efficient. This is because its computational complexity is only $\mathcal{O}(N_K)$ while the k-means clustering and MEAP is a time-consuming iterative procedure. Our proposed method offers the lowest OSPA. When there is no clutter, our proposed method can reduce the OSPA by approximately 5% and 3% as compared with k-means and MEAP, respectively. However, the advantage of our proposed method in terms of OSPA tends to decrease with the growing number of clutters. The reason is that some particles may be associated with clutter, which increases the risk of generating false particle flows.

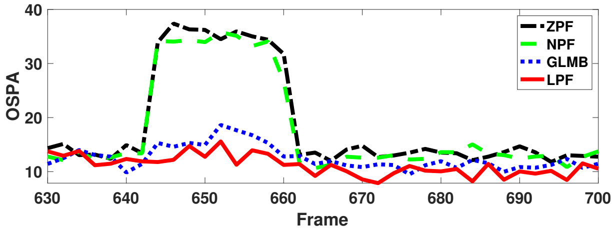


Fig. 8. The OSPA of ZPF, NPF, GLMB and LPF for Sequence 45 (camera 3) for frames 630-700.

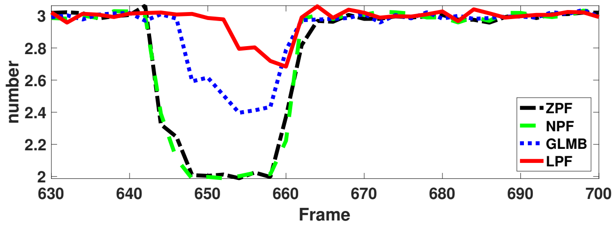


Fig. 9. The number of speakers estimated by ZPF, NPF, GLMB and LPF for Sequence 45 (camera 3) for frames 630-700.

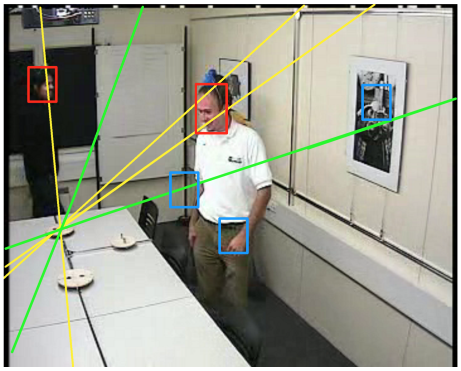


Fig. 10. The clutter at frame 660 of Sequence 45. The face detection, DOA lines, visual clutters, and the audio clutters are shown in the red boxes, yellow lines, blue boxes and green lines, respectively.

TABLE II
RUNNING TIME (S) AND OSPA OF K-MEANS, MEAP AND OUR ESTIMATION METHOD CHANGES WITH NUMBER OF CLUTTERS AT FRAME 660 OF SEQUENCE 45 OF AV 16.3.

Num. of clutters	0	10	20	30	40	
K-means	0.4	0.73	0.8	1.08	1.65	time (s)
	13.5	15.6	20.4	24.8	31.3	OSPA
MEAP	0.38	0.63	0.75	0.89	1.32	time (s)
	13.2	14.5	19.8	24.6	31.3	OSPA
Our method	**0.31**	**0.32**	**0.32**	**0.32**	**0.33**	time (s)
	12.8	**14.1**	**18.6**	**22.5**	**30.4**	OSPA

E. Comparison with other audio-visual algorithms

In this subsection, the proposed algorithm LPF is compared with several baselines, including ZPF [34], NPF [35], GLMB [38], GPF [37] and SMS algorithms [60]. GPF, GLMB and ZPF are implemented by the zero diffusion particle flow. LPF and NPF are implemented by the non-zero diffusion particle flow. Although the particle flow is not used in the SMS, the weight degeneracy of SMC filter is address by the mean-shift method. Table III reports the average OSPA. It can be observed that using LPF about 24% reduction in tracking error has been achieved as compared with NPF. The advantage of LPF is clear on Sequence 45, where occlusion happens frequently.

TABLE III
THE OSPA OF LPF, NPF, ZPF, GPF, GLMB AND SMS RUNNING ON THE AV 16.3.

Seq (Cam)	LPF	NPF	ZPF	GPF	GLMB	SMS
24 (1)	**10.13**	12.32	12.99	13.00	10.66	14.50
24 (2)	**10.34**	13.20	13.82	15.13	11.98	15.35
24 (3)	**9.38**	13.23	14.01	15.22	11.62	15.72
25 (1)	**12.04**	15.96	16.80	18.28	13.43	17.17
25 (2)	**12.35**	15.29	15.88	15.58	12.94	15.39
25 (3)	**12.13**	16.29	17.56	18.62	14.45	17.62
30 (1)	**12.64**	15.76	17.15	18.89	13.55	19.27
30 (2)	**10.24**	13.41	14.22	16.12	11.68	16.16
30 (3)	**12.44**	15.93	17.63	19.03	16.38	19.67
45 (1)	**13.12**	17.65	19.33	23.12	18.14	23.40
45 (2)	**14.24**	18.60	20.85	22.71	20.35	23.16
45 (3)	**14.12**	19.50	21.35	23.76	20.36	23.80
Avg. OSPA	**11.93**	15.60	16.80	18.28	14.63	18.43

To show the difference among the results of the tested algorithms in Table III, we have run the ANOVA based F-test [66] and present the results in Table IV. The significance value is set as 5%, and the degree of freedoms for all the significance tests is $(1, 22)$. The corresponding critical value F_{crit} is 4.30 in terms of the F-distribution table [66]. Note that 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 observed that the improvements of LPF over ZPF, NPF and GLMB are statistically significant.

TABLE IV
SIGNIFICANCE TEST FOR THE DIFFERENCE BETWEEN LPF, AND NPF, ZPF, GPF, GLMB AND SMS, RESPECTIVELY.

Method	NPF	ZPF	GPF	GLMB	SMS	
LPF	21.4	28.12	33.15	6.21	36.08	F
	0.0001	2.5e-05	8.60e-06	0.021	4.80e-06	p-value

Table V shows Frame rate per Sequence per Speaker (FPSS) and computational complexities of the compared algorithms. Since the measurements of different algorithms are the same, the computational complexities of the detector is not included in this table. As shown in Table V, LPF and NPF have a similar computational cost. Although LPF calculates the particle labels, LPF saves the cost at the clustering step. Since GLMB considers not only the particle labels but also the label history, the particles of GLMB may have multiple labels. However, the particles of LPF have only one label. Therefore, the number of particle flows in GLMB is greater than that in LPF, which leads to a higher computational cost. Although LPF and GPF use the same initial number of particles, the number of particles is drastically varying since a few particles are added in the update step of the GPF [37]. Therefore, LPF runs faster than the GPF. The FPSS of LPF is about 25 % higher than that of NPF. As the complexity of LPF does not depend on the number of measurements, LPF is computationally more efficient than GLMB.

TABLE V
FRAME RATE PER SEQUENCE PER SPEAKER (FPSS) COMPARISON FOR LPF, NPF, ZPF, GPF, GLMB AND SMS.

Seq	LPF	ZPF	NPF	GPF	GLMB	SMS
24	6.72	4.02	5.52	1.21	2.21	6.53
25	6.63	3.83	5.21	1.03	1.95	6.35
30	6.75	3.96	5.40	1.11	2.01	6.48
45	4.51	2.64	3.53	0.74	1.34	4.32
FPSS	**6.15**	**3.61**	**4.92**	**1.02**	**1.88**	**5.92**
Com	$N_k N_\lambda$	$N_k N_\lambda$	$N_k N_\lambda$	$N_k U_k N_\lambda$	$N_k U_k N_\lambda$	$U_k N_k$

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 [53], and the frames 100-170 (four speakers) and frames 180-250 (five speakers) for Sequence UKA from the CLEAR dataset [57]. Their average errors are summarised in Table VI. Our proposed LPF offers the lowest OSPA among all the filters. 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.

F. Comparison with deep learning methods

In this subsection, we compare the proposed method with deep learning based methods, i.e. YoloV5-DeepSort[1] and YoloV5-StrongSort[2], using Sequence 45 Camera 1 of AV16.3, under occlusions (frames 640 to 665) and non-occlusions (the remaining frames of this sequence), respectively. Considering the fact that these two deep learning based methods use visual information only, we adjust the proposed method accordingly by dropping the audio measurements to ensure a fair comparison. The audio labels were set as 0 in this case. All the

[1] https://github.com/HowieMa/DeepSORT_YOLOv5_Pytorch

[2] https://github.com/mikel-brostrom/Yolov5_StrongSORT_OSNet

TABLE VI
EXPERIMENTAL RESULTS FOR LPF, NPF, ZPF, GPF, GLMB AND SMS, IN TERMS OF THE OSPA ERROR FOR SEQUENCE 09 AND 32 OF THE AVDIAR DATASET AND FRAMES 100-170 AND FRAMES 180-250 FOR SEQUENCE UKA 20060726 OF THE CLEAR DATASET.

Filters	sequence 09	sequence 32	frames 100-170	frames 180-250
LPF	11.75	12.32	**24.53**	**26.65**
ZPF	13.72	14.37	28.62	31.57
SMS	13.95	14.90	29.35	36.68
GPF	13.82	14.78	30.25	37.84
GLMB	**11.74**	**12.25**	24.67	26.98
NPF	13.80	14.42	28.60	31.55

TABLE VII
EXPERIMENTAL RESULTS FOR LPF, YOLOV5-DEEPSORT AND YOLOV5-STRONGSORT, IN TERMS OF THE AVERAGE OSPA ERROR OF TWO COMPARISON GROUPS FOR SEQUENCE 45 CAMERA 1 OF THE AV16.3 DATASET (FRAMES 630-700).

Method	Non-Occlusion	Occlusion
LPF	14.45	**22.81**
YoloV5-DeepSort	14.27	27.54
YoloV5-StrongSort	**13.11**	24.22

compared methods are run under the same condition as in Section IV-D1 with the same measurements given by YoloV5 using the pre-trained model 'crowdhuman_yolov5m.pt'. Unlike YoloV5-DeepSort which uses the default settings for the Re-ID model, we select 'osnet_x0_25_market1501.pt' for YoloV5-StrongSort. The experimental results in terms of the average OSPA over these frames are given in Table VII. For the non-occlusion cases, all the three methods provide similar performance, with the deep learning methods providing slightly better results. However, for the occlusion cases, our proposed method offers better performance with a lower average OSPA error at 22.81, as compared with YoloV5-DeepSort and YoloV5-StrongSort. This is because our proposed method has exploited visual and speaker labels which characterise the historical information of the particles and speakers during tracking. Even if the speaker is occluded, its states could still be estimated with the label information. Although the two deep learning based methods use historical information within a Kalman filter framework, they did not consider cluster density and detection density, which results in failure during occlusion.

Our AV-LPF-SMC-PHD filter does not involve model training, which offers advantages in cases where no training data is available. In addition, it is a flexible method, and could be used together with a deep learning method, such as YoloV5-DeepSort and YoloV5-StrongSort, where the detection results from deep learning methods can be used as measurements in our AV-LPF-SMC-PHD filter. This can help leverage the excellent detection performance from deep learning and promising performance of the proposed method in tracking over occlusions.

V. CONCLUSION

We have presented a novel AV-LPF-SMC-PHD filter for audio-visual multi-speaker tracking using particle labels. Specifically, the audio and visual labels of the particles are independently estimated based on the likelihood density and detection probability. Based on our proposed label space, the states of the undetected speakers can be estimated by the audio and visual labels. The particles associated with the detected speaker are selected and updated by the labelled particle flow with our proposed likelihood, which considers four different situations in the label space. Finally, the weighted mean of the selected particles is used for calculating the states of the speakers, which replaces the clustering step widely used in conventional SMC-PHD filters. The proposed algorithm has been tested on the AV16.3, AVDIAR and CLEAR datasets, and compared with other particle flow methods, PHD filters, GLMB filter and deep learning methods. The experimental results show that the proposed filter offers a higher tracking accuracy than several baseline methods with a lower computational cost.

APPENDIX A
LIKELIHOOD COMPARISON AND PROBABILITY ANALYSIS

For the convenience of expression, some complex symbols are redefined in this appendix. Audio likelihood $\mathring{h}^{i,o}_{k|k-1}$ and visual likelihood $\breve{h}^{i,u}_{k|k-1}$ are both denoted as $\{h^z\}_{z=1}^N$, where z is the index of the measurement and N is the number of measurements. The maximun value of $\{h^z\}_{z=1}^N$ is h^ζ and $\arg\max_{z\in\{1,\ldots,N\}}(h^z)=\zeta$. We define r^z as a random value from 0 to 1 and z_1 and z_2 are two different indices of the measurement.

The probability of $r^{z_1}h^{z_1} \geq r^{z_2}h^{z_2}$ is defined as $P(r^{z_1}h^{z_1} \geq r^{z_2}h^{z_2})$ where $z_1 \in \{1,...,N\}$ and $z_2 \in \{1,...,N\}$. If $h^{z_1} \geq h^{z_2}$,

$$\begin{aligned}
&P(r^{z_1}h^{z_1} \geq r^{z_2}h^{z_2})\\
=&P(r^{z_1}\frac{h^{z_1}}{h^{z_2}} \geq 1) + P(r^{z_1}\frac{h^{z_1}}{h^{z_2}} \geq r^{z_2}|r^{z_1} < \frac{h^{z_2}}{h^{z_1}})P(r^{z_1} < \frac{h^{z_2}}{h^{z_1}})\\
=&\frac{h^{z_1}-h^{z_2}}{h^{z_1}} + \frac{1}{2}\frac{h^{z_2}}{h^{z_1}} = 1 - \frac{h^{z_2}}{2h^{z_1}} \geq \frac{1}{2}.
\end{aligned} \tag{47}$$

If $h^{z_1} < h^{z_2}$

$$\begin{aligned}
&P(r^{z_1}h^{z_1} \geq r^{z_2}h^{z_2}) = 1 - P(r^{z_1}h^{z_1} < r^{z_2}h^{z_2})\\
&= 1 - 1 + \frac{h^{z_1}}{2h^{z_2}} < \frac{1}{2}.
\end{aligned} \tag{48}$$

The probability of $\arg\max(r^z h^z) = z_1$ is

$$P(\arg\max(r^z h^z) = z_1) = \prod_{1\leqslant z\leqslant N} P(r^{z_1}h^{z_1} \geq r^z h^z). \tag{49}$$

If $z_1 \neq \zeta$, the probability of $\arg\max(r^z h^z) = \zeta$ is

$$P(\arg\max(r^z h^z) = \zeta) = \prod_{1\leqslant z\leqslant N} P(r^\zeta h^\zeta \geq r^z h^z). \tag{50}$$

To compare $P(\arg\max(r^z h^z) = \zeta)$ and $P(\arg\max(r^z h^z) = z_1)$, we need to compare $P(r^\zeta h^\zeta \geq r^z h^z)$ and $P(r^{z_1}h^{z_1} \geq r^z h^z)$. If $h^\zeta \geq h^z > h^{z_1}$, we can get

$$P(r^\zeta h^\zeta \geq r^z h^z) \geq \frac{1}{2}. \tag{51}$$

$$P(r^{z_1}h^{z_1} \geq r^z h^z) < \frac{1}{2}. \tag{52}$$

and

$$P(r^{\zeta}h^{\zeta} \geq r^z h^z) \geq \frac{1}{2} > P(r^{z_1}h^{z_1} \geq r^z h^z). \tag{53}$$

If $h^{\zeta} \geq h^{z_1} \geq h^z$,

$$\begin{aligned} &P(r^{\zeta}h^{\zeta} \geq r^z h^z) - P(r^{z_1}h^{z_1} \geq r^z h^z) \\ =&1 - \frac{h^z}{h^{\zeta}} - 1 + \frac{h^z}{h^{z_1}} \geq 0. \end{aligned} \tag{54}$$

Therefore, we can find that $P(r^{\zeta}h^{\zeta} \geq r^z h^z)$ always has a larger value than $P(r^{z_1}h^{z_1} \geq r^z h^z)$, where $z_1 \neq \zeta$. Based on Eq. (49) and Eq. (50), we get $P(\arg\max(r^z h^z) = \zeta) > P(\arg\max(r^z h^z) = z_1)$, where $z_1 \neq \zeta$ and $1 \leqslant z_1 \leqslant N$, which means that $\arg\max(r^i h^i)$ has a high probability to be equal to the index of the high likelihood ζ.

Acknowledgement

The authors wish to thank the reviewers and the associate editor for their constructive comments to further improve the paper. W. Wang was partially supported by the SIGNetS project. For the purpose of open access, the authors have applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising.

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] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1704–1716, 2013.

[4] R. Thiolliere, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling," in *Proc. INTERSPEECH*, 2015, pp. 3179–3183.

[5] P. Escudero, C. D. Bonn, R. N. Aslin, and K. E. Mulak, "Indexical and linguistic processing in infancy: Discrimination of speaker, accent and vowel differences," in *Proc. Int. Congress of Phonetic Sciences.*, May 2015, pp. 1–5.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

TABLE VIII
DESCRIPTION OF SYMBOLS.

Symbols	Description
j	Index of speaker
k	Index of time
$\tilde{\boldsymbol{m}}_k^j$	State vector
(x_k^j, y_k^j)	Speaker position
$(\dot{x}_k^j, \dot{y}_k^j)$	Speaker velocities
(w_k^j, l_k^j)	Size of a bounding box
$\tilde{N}$	Number of speakers
$\{\mathring{\boldsymbol{z}}_k^o\}_{o=1}^{\mathring{N}_k}$	Audio measurement set
$\mathring{N}_k$	Number of audio measurements
o	Index of audio measurements
$\mathring{\boldsymbol{\Psi}}_k$	Audio measurement noise
$\mathring{\boldsymbol{\epsilon}}_k$	Audio clutter
$\mathring{\mathbf{F}}_{\boldsymbol{z}}$	Audio measurement model
$\mathring{\boldsymbol{\omega}}_k$	Weight of the audio likelihood
$\{\breve{\boldsymbol{z}}_k^u\}_{u=1}^{\breve{N}_k}$	Visual measurement set
$\breve{N}_k$	Number of visual measurements
u	Index of visual measurements
$\breve{\boldsymbol{\Psi}}_k$	Visual measurement noise
$\breve{\boldsymbol{\epsilon}}_k$	Visual clutter
$\breve{\mathbf{F}}_{\boldsymbol{z}}$	Visual measurement model
$\breve{\boldsymbol{\omega}}_k$	Weight of the visual likelihood
$\mathbf{F}_{\tilde{\boldsymbol{m}}}$	State transition model
$\boldsymbol{\Upsilon}_k$	Transition system noise
i	Index of particle
N_k	Number of particles at k
$\boldsymbol{m}_k^i$	ith particle state at k
ω_k^i	ith particle weight at k
$\boldsymbol{Z}_k$	audio and visual measurement set
q_k	Proposal distribution
ϕ	Analogue of the state transition probability with the state
N_B	Number of the birth particles
p_k	New born importance function
γ_k	PHD of new targets
$\lVert\cdot\rVert_1$	L_1 norm
$\boldsymbol{w}_k^i$	Wiener process
$\triangle\lambda$	Particle pseudo time step
$\triangle\boldsymbol{m}_{k\vert k-1}^i$	Particle moving distance in $\triangle\lambda$
$\boldsymbol{P}_k^i$	Particle covariance matrix
$\kappa_k(\boldsymbol{z}_k^r)$	Clutter intensity of the rth measurement $\boldsymbol{z}_k^r$ at time k
$p_{D,k}^i$	Detection probability at time k
$h_k^{i,r}$	Likelihood of the ith particle for the rth measurement
$\boldsymbol{l}_k^i$	Set of labels
a_k^i	Audio measurement associated with the ith particle
v_k^i	Visual measurement associated with the ith particle
t_k^i	The speaker associated with the ith particle
$\mathring{h}_k^{i,o}$	The oth audio likelihood of the ith particle $\boldsymbol{m}_k^i$
$\breve{h}_k^{i,u}$	The uth visual likelihood of the ith particle $\boldsymbol{m}_k^i$
H	Heaviside step function
r_D^i	Parameters valued from 0 to 1 for particle detection
δ	Dirac delta function
$\boldsymbol{f}_k^i$	Labelled particle flow of the ith particle at time k
ϕ	Transition distribution
$\gamma_k(\boldsymbol{m}_{k\vert k-1}^i)$	Birth density
ξ	Weight threshold for clutters

[13] 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.

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

[15] D. Y. Kim, B.-T. Vo, and S. Nordholm, “Multiple speaker tracking with the GLMB filter,” in *Proc. IEEE Int. Conf. Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2017, pp. 38–43.

[16] Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, “Variational bayesian inference for audio-visual tracking of multiple speakers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1761–1776, 2019.

[17] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, “Online localization and tracking of multiple moving speakers in reverberant environments,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 88–103, 2019.

[18] Y. Ban, X. Alameda-Pineda, C. Evers, and R. Horaud, “Tracking multiple audio sources with the von mises distribution and variational em,” *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 798–802, 2019.

[19] M. Broström, “Real-time multi-camera multi-object tracker using YOLOv5 and StrongSORT with OSNet,” https://github.com/mikel-brostrom/Yolov5_StrongSORT_OSNet, 2022.

[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] S. Lin and X. Qian, “Audio-visual multi-speaker tracking based on the glmb framework.” in *INTERSPEECH*, 2020, pp. 3082–3086.

[22] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore, “Audio-visual speaker tracking with importance particle filters,” in *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, vol. 3. IEEE, 2003, pp. III–25.

[23] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, “Multi-speaker tracking from an audio-visual sensing device,” *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2576–2588, 2019.

[24] 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.

[25] 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.

[26] F. Daum and J. Huang, “Particle flow for nonlinear filters,” *Proc. SPIE Conf. Signal Processing, Sensor Fusion, and Target Recognition*, vol. 769704, pp. 5920–5923, Apr. 2011.

[27] ——, “Particle flow for nonlinear filters with log-homotopy,” *Proc. SPIE Conf. Signal Processing Sensor Fusion, Target Recognition*, vol. 6969, pp. 696 918–1 – 696 918–12, 2008.

[28] ——, “Renormalization group flow and other ideas inspired by physics for nonlinear filters, Bayesian decisions, and transport,” *Proc. SPIE Defense and Security*, pp. 90 910I–1 – 90 910I–14, 2014.

[29] ——, “Small curvature particle flow for nonlinear filters,” *Proc. SPIE Conf. Signal Processing, Sensor Fusion, and Target Recognition*, pp. 8393 – 8393–11, 2012.

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

[31] 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.

[32] N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[33] X. Zhou, V. Koltun, and P. Krähenbühl, “Tracking objects as points,” in *European Conference on Computer Vision*. Springer, 2020, pp. 474–490.

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

[35] 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,” *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

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

[37] 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, 2016, pp. 98 420D – 98 420D–10.

[38] A.-A. Saucan, Y. Li, and M. J. Coates, “Particle flow SMC delta-GLMB filter,” *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[39] S. S. Blackman, “Multiple hypothesis tracking for multiple target tracking,” *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 1, pp. 5–18, 2004.

[40] B.-T. Vo and B.-N. Vo, “Labeled random finite sets and multi-object conjugate priors,” *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3460–3475, 2013.

[41] B.-N. Vo, B.-T. Vo, and D. Phung, “Labeled random finite sets and the bayes multi-target tracking filter,” *IEEE Transactions on Signal Processing*, vol. 62, no. 24, pp. 6554–6567, 2014.

[42] Y. Sung and P. Tokekar, “GM-PHD filter for searching and tracking an unknown number of targets with a mobile sensor with limited fov,” *IEEE Transactions on Automation Science and Engineering*, 2021.

[43] Y. Liu, Q. Hu, Y. Zou, and W. Wang, “Labelled non-zero particle flow for SMC-PHD filtering,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5197–5201.

[44] L. P. Kadanoff, *Statistical Physics: Statics, Dynamics and Renormalization*. World Scientific Publishing Co Inc, 2000.

[45] F. Daum and J. Huang, “Particle flow with non-zero diffusion for nonlinear filters,” *Proc. SPIE Conf. Signal Processing, Sensor Fusion, and Target Recognition*, vol. 04, pp. 87 450P–87 450P–13, 2013.

[46] D. Arthur and S. Vassilvitskii, “K-means++: The advantages of careful seeding,” *Proc. the Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

[47] 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.

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

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

[50] P. Hu and D. Ramanan, “Finding tiny faces,” in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2017, pp. 951–959.

[51] Y. Liu, V. Kılıç, J. Guan, and W. Wang, “Audio–visual particle flow smc-phd filtering for multi-speaker tracking,” *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 934–948, 2019.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

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

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

[60] 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.

[61] 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.

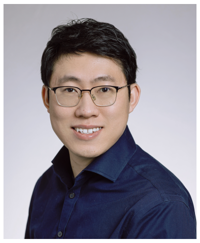
[62] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[63] A. C. Cameron and F. A. Windmeijer, "An R-squared measure of goodness of fit for some common nonlinear regression models," *Journal of Econometrics*, vol. 77, no. 2, pp. 329–342, 1997.

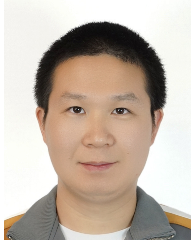
[64] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Academic Press, New York, 1999.

[65] P. Bromiley, "Products and convolutions of Gaussian probability density functions," *Tina-Vision Memo*, vol. 3, no. 4, p. 1, 2003.

[66] P. G. Hoel *et al.*, *Elementary Statistics*. John Wiley & Sons, London & New York, 1960.



Yang Liu (M'16-SM'22) currently holds the position of Senior Research Scientist at Meta, bringing his expertise from previous roles as a researcher at both Microsoft and Zoom. He obtained his doctorate from the renowned Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey, where he was supervised by esteemed Professors Wenwu Wang and Adrian Hilton. As an IEEE Senior Member, he has contributed significantly to the academic community by serving as a Session Chair at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2022 and 2023. He has also been a reviewer for over ten top conferences and journals, such as ICASSP, Interspeech, Neural Information Processing Systems (NeurIPS), Winter Applications of Computer Vision (WACV), and IEEE Transactions. His professional affiliations are extensive, including his membership in the IEEE Signal Processing Society, IEEE Young Professionals, CBAIA Research Fellow, Audio Engineering Society (AES) Professional Member, International Speech Communication Association (ISCA) Senior Member, and Association for Computing Machinery (ACM) Senior Member. His primary research interests lie in audio signal processing for edge devices, audio-visual multimodal learning, and audio generation technologies. Within these domains, his work is particularly focused on echo and noise cancellation, speech enhancement, and acoustic scene classification, all of which contribute to the advancement of audio processing and its applications.



Yong Xu (M'16-SM'23) is currently a Principal Researcher at Tencent America, Bellevue, WA, USA. He once worked at the University of Surrey, UK as a Research Fellow from 2016 to 2018. He got his Ph.D. degree from the University of Science and Technology of China (USTC) in 2015 and studied under a joint Ph.D. program at Georgia Institute of Technology, USA. He achieved the 2018 IEEE SPS Best Paper Award for his work on deep learning-based speech enhancement. He is a member of IEEE Signal Processing Society - Speech and Language Technical Committee (SLTC) (2023-2025). His work has got 5400+ citations on Google Scholar.

Peipei Wu has been working towards his PhD at the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey, UK since 2020. His research focuses on distributed audio-visual multi-target tracking. He embarked on his academic journey in 2018, earning a Bachelor's degree in Cyber Security. This was followed by a Master's degree in Computer Vision, Robotics, and Machine Learning from the University of Surrey in 2020. From 2018 to 2021, he undertook further studies at Nanchang Aviation University, earning a second Master's degree in Software Engineering in 2021.

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, Cardiff University, Tao Group Ltd. (now Antix Labs Ltd.), and Creative Labs, 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 is also an AI Fellow within the Surrey Institute for People Centred Artificial Intelligence. 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 300 publications in these areas. He is a (co-)author or (co-)recipient of over 15 awards including the 2022 IEEE SPS Young Author Best Paper Award, ICAUS 2021 Best Paper Award, DCASE 2020 Best Paper Award, DCASE 2019 and 2020 Reproducible System Award, LVA/ICA 2018 Best Student Paper Award, FSDM 2016 Best Oral Presentation, ICASSP 2019 and LVA/ICA 2010 Best Student Paper Award Nominees. He is an Associate Editor (2020-2025) for IEEE/ACM Transactions on Audio Speech and Language Processing. He was a Senior Area Editor (2019-2023) and an Associate Editor (2014-2018) for IEEE Transactions on Signal Processing. He is elected Chair (2023-2024) of IEEE SPS Machine Learning for Signal Processing Technical Committee, elected Vice Chair (2022-2024) of EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, an elected Member (2021-2023) of the IEEE Signal Processing Theory and Methods Technical Committee, and an elected Member (2019-) of the International Steering Committee of Latent Variable Analysis and Signal Separation. He was a Publication Co-Chair for ICASSP 2019, a Satellite Workshop Co-Chair for INTERSPEECH 2022 and ICASSP 2024, and a Technical/Program Committee Member of over 100 international conferences.