Mean-Shift and Sparse Sampling-Based SMC-PHD Filtering for Audio Informed Visual Speaker Tracking

Volkan Kılıç, *Student Member, IEEE*, Mark Barnard, Wenwu Wang, *Senior Member, IEEE*, Adrian Hilton, *Member, IEEE*, and Josef Kittler, *Life Member, IEEE*

***Abstract*—The probability hypothesis density (PHD) filter based on sequential Monte Carlo (SMC) approximation (also known as SMC-PHD filter) has proven to be a promising algorithm for multispeaker tracking. However, it has a heavy computational cost as surviving, spawned, and born particles need to be distributed in each frame to model the state of the speakers and to estimate jointly the variable number of speakers with their states. In particular, the computational cost is mostly caused by the born particles as they need to be propagated over the entire image in every frame to detect the new speaker presence in the view of the visual tracker. In this paper, we propose to use the audio data to improve the visual SMC-PHD (V-SMC-PHD) filter by using the direction of arrival angles of the audio sources to determine when to propagate the born particles and reallocate the surviving and spawned particles. The tracking accuracy of the audio-visual SMC-PHD (AV-SMC-PHD) algorithm is further improved by using a modified mean-shift algorithm to search and climb density gradients iteratively to find the peak of the probability distribution, and the extra computational complexity introduced by mean-shift is controlled with a sparse sampling technique. These improved algorithms, named as AVMS-SMC-PHD and sparse-AVMS-SMC-PHD, respectively, are compared systematically with AV-SMC-PHD and V-SMC-PHD based on the AV16.3, AMI, and CLEAR datasets.**

***Index Terms*—Audio-visual tracking, mean-shift, multi-speaker tracking, probability hypothesis density (PHD) filter, sequential Monte Carlo (SMC) implementation, sparse particles.**

I. Introduction

SPEAKER tracking in enclosed spaces has received much interest in the fields of computer vision and signal processing, driven by applications such as video conferencing [1], speaker discrimination [2], acoustic beamforming [3], audio-visual speech recognition [4], video indexing and retrieval [5], human-computer interaction [6], and surveillance [7]. However, speaker tracking in real life scenarios involves several challenges such as estimation of the variable number of speakers and their states, and dealing with various conditions such as occlusion, limited view of cameras, illumination change and room reverberations.

One approach to overcome these challenges is to use multi-modal information, as it provides additional observations about the state of each speaker compared to single-modal tracking. The multi-modal information used for tracking can be collected by sensors such as audio, video, thermal vision, laser range finders and RFID [8]. In speaker tracking, audio and video sensors are widely applied compared to others, for their easier installation, cheaper cost, and more data processing tools. Hence, our tracking system is also based on audio and visual data.

Video tracking is generally reliable and accurate when the targets are in the camera field of view [9], but is limited when the targets are occluded by other speakers, when they disappear from the camera field of view, or the appearance of the targets or illumination has changed [10]–[15]. On the other hand, audio tracking [16] is not restricted by these limitations. However, it is prone to non-negligible tracking errors as audio data is intermittent over time and may be corrupted by background noise and room reverberations. Nevertheless, the audio and visual modalities contain complementary information that can be used to improve the tracking performance in the case that either modality is unavailable or both are corrupted [2], [6], [17], [18], which is our focus here.

Several approaches have been proposed to use the multimodal information which can be categorized into two classes: namely, deterministic (data-driven) and stochastic (model-driven) [19]–[21]. Deterministic approaches are often considered as an optimization problem based on a cost function. A representative method in this category is the mean-shift [22]–[24] where the cost function is defined in terms of color similarity measured by the Bhattacharyya distance. The stochastic approaches use a state-space approach based on the Bayesian framework [25], [26]. Representative methods include the Kalman filter (KF) [27], extended KF (EKF), and particle filter (PF) [28]. In comparison to the KF and EKF approaches, the PF approach is more robust for non-linear and non-Gaussian models as it easily approaches the Bayesian optimal estimate with a sufficiently large number of particles [8]. It has been widely employed for speaker tracking [25], [29], [30].

Manuscript received November 13, 2015; revised April 23, 2016; accepted August 2, 2016. Date of publication August 10, 2016; date of current version November 15, 2016. This work was supported by the Engineering and Physical Sciences Research Council of the U.K. under Grant EP/K014307/1 and Grant EP/L000539/1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sen-Ching Samson Cheung.

V. Kılıç was with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. He is now with the Department of Electrical and Electronics Engineering, Izmir Katip Celebi University, Cigli-Izmir 35620, Turkey (e-mail: volkan.kilic@ikc.edu.tr).

M. Barnard, W. Wang, A. Hilton, and J. Kittler are with the Centre for Vision, Speech and Signal Processing, Department of Electrical and Electronic Engineering, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: mark.barnard@surrey.ac.uk; w.wang@surrey.ac.uk; a.hilton@surrey.ac.uk; j.kittler@surrey.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2599150

The generic PF applied to multi-speaker AV tracking is often under the assumption that the number of speakers is known and invariant. In practice, however, the speakers to be captured by the AV sensors may appear or disappear in a random manner. As a result, the number of speakers that can be observed from the AV measurements may vary with time. To address this issue, the theory of random finite sets (RFSs) has been introduced for tracking unknown and variable number of speakers which allows multi-speaker filtering by propagation of the multi-speaker posterior [31], [32]. The computational complexity of RFS, however, grows exponentially with the number of speakers. To overcome this problem, the PHD filtering approach [32] was proposed as the first order approximation of the RFS, whose complexity scales linearly with the number of speakers. It has been found to be promising for multi-speaker tracking [31], [32]. Different from the Bayesian (Kalman or PF) approach, the PHD filter does not require the *a priori* knowledge of the number of targets, which is actually estimated during the tracking process.

The SMC implementation [33] is introduced to obtain practical solutions of the PHD filter. The SMC-PHD filter uses particles to model the surviving, spawned and born state of the speaker. In the standard implementation of the SMC-PHD based visual tracking [33], the born particles are propagated in every frame to detect the speaker presence in the view, which is computationally expensive. To address this limitation, we propose to use the DOA information obtained from audio for the propagation of the particles. More specifically, the propagation of the born particles is decided based on the DOA information and the particles are re-located around the line drawn upon the DOA. A similar approach has been used in [34]–[36] under the PF framework for a fixed number of speakers. Here, the SMC-PHD filter is used, and to our knowledge, audio information has not been previously used with visual information in a SMC-PHD filter as we do here.

The estimation accuracy of the SMC-PHD filter, however, is compromised due to the use of the first-order approximation derived from RFS. In this paper, we propose a new method by employing the mean-shift to improve the particle distribution within the SMC-PHD filter. The mean-shift is run on the particle set to pull the centre of the particle distribution towards the target centre. This leads to improvement in estimation accuracy as observed in our experiments shown in Section VI-C. Although mean-shift has been previously used with particle filtering in [19], [21], [37], [38], [39], in various frameworks, none of these were explicitly designed for a variable number of targets since the structure of both methods was devised for single target tracking scenarios.

The mean-shift approach is computationally efficient, but it may converge to saddle points in the case of multi-modal distribution [40], and may fail to track small and fast moving targets and is unable to recover a track after partial or total occlusions [37], [38], [41]. These problems can be easily handled by the SMC-PHD filter due to its ability to recover from lost tracks [38], and the use of multiple particles which can help mean-shift to detect the target even if some of the particles fall in local maxima or saddle points. Another problem with the mean-shift is its limitation in adapting to the size or scale of the target. However, this problem can be solved with the SMC-PHD filter since the scale is one of the states of the target.

The mean-shift process is used to move the particle towards the target location leading to error reduction, but repeating this process for all the particles induces extra computational cost. To overcome this problem, we propose a technique based on “sparse sampling” leading to a new concept “sparse particle”. The traditional way of using sparsity in tracking is to represent the target appearance or features with sparsity [18], [42], [43]. Unlike the traditional way, sparse particles are obtained with sparse sampling strategy, which, to our knowledge, has not been done before.

This paper is an extended version of our previous study described in [44]. The main modification lies in the formulation and justification of the improved tracking scheme, the mean-shift and sparse sampling integration, and more experiments. The major contributions of this paper can be summarized as follows.

1) Audio is used for particle propagation of the SMC-PHD filter and to improve the tracking performance and robustness of the visual tracker for a variable number of speakers.
2) A new method is developed by using mean-shift to improve particle distribution in the particle propagation step of the SMC-PHD filter.
3) A novel sparse sampling algorithm is proposed to generate sparse particles for which the mean-shift iteration is operated in order to reduce the computational cost.

The rest of this paper is organized as follows: the next section introduces the PHD filter for visual multi-speaker tracking. Section III describes our proposed audio-visual SMC-PHD (AV-SMC-PHD) filtering algorithm. In Section IV and V, the mean-shift and sparse sampling are integrated in the proposed AV-SMC-PHD filtering algorithm for further improvements. Section VI shows experimental results performed on the $AV16.3$, AMI and CLEAR datasets and compares the performance of the algorithms. Closing remarks are given in Section VII.

II. Multi-speaker Tracking With the PHD Filter

This section describes our problem formulation for multi-speaker visual tracking based on the PHD filter.

Let us represent the state of a speaker by a vector $\mathbf{x} = [x_1 \;\; \dot{x_1} \;\; x_2 \;\; \dot{x_2} \;\; s]^T$ in a single speaker tracking system where x_1 and x_2 are, respectively, the horizontal and vertical positions of the rectangle centred around the face that we wish to track, $\dot{x_1}$ and $\dot{x_2}$ are, respectively, the horizontal and vertical velocity, and s is the scale of the rectangle centred around (x_1, x_2). For the evolution of the time dependent speaker state, the constant velocity model is employed [36], [45] given as

$$\mathbf{x}_k = \mathbf{F}\mathbf{x}_{k-1} + \mathbf{q}_k \tag{1}$$

where $\mathbf{q}_k$ is the zero-mean Gaussian noise with covariance $\mathbf{Q}$, $\mathbf{q}_k \sim \mathcal{N}(0, \mathbf{Q})$ for speaker at time frame $k = 1, ..., K$ and $\mathbf{F}$ is the linear motion model.

In our work, the multi-speaker states and measurements are characterized by using the RFS theory [32], given by

$$\mathcal{X}_k = \{\mathbf{x}_{1,k}, ..., \mathbf{x}_{\Xi_k,k}\} \tag{2}$$

$$\mathcal{Z}_k = \{\mathbf{z}_{1,k}, ..., \mathbf{z}_{M_k,k}\} \tag{3}$$

where $\Xi_k = |\mathcal{X}_k|$ is the number of speakers, with $|\cdot|$ representing the cardinality of the set. $\mathcal{Z}_k$ consists of M_k observations which may be corrupted by noise due to clutter. Uncertainty in a single speaker Bayesian tracking is introduced by modelling $\mathbf{x}_k$ and $\mathbf{z}_k$ as random vectors. In multi-speaker case, uncertainty is introduced by modelling $\mathcal{X}_k$ and $\mathcal{Z}_k$ as RFSs [46]

$$\mathcal{X}_k = \mathcal{S}_k(\mathcal{X}_{k-1}) \cup \mathcal{B}_k(\mathcal{X}_{k-1}) \cup \Gamma_k \tag{4}$$

$$\mathcal{Z}_k = \Theta_k(\mathcal{X}_k) \cup \mathcal{C}_k \tag{5}$$

where '∪' denotes union, $\mathcal{S}_k(\mathcal{X}_{k-1})$ denotes the RFS of surviving speakers, $\mathcal{B}_k(\mathcal{X}_{k-1})$ is the RFS of speakers spawned from the previous set of speakers $\mathcal{X}_{k-1}$ and Γ_k is the RFS of the new speakers that appear spontaneously at time k [33]. $\Theta_k(\mathcal{X}_k)$ denotes the RFS of the measurements generated by the speakers $\mathcal{X}_k$, and $\mathcal{C}_k$ is the RFS of clutter or false alarms. Besides, the dynamics in the state evolution $\mathcal{X}_k$ and the randomness in the observations are described by the multi-speaker transition density $f_{k|k-1}(\mathcal{X}_k|\mathcal{X}_{k-1})$ and likelihood $g_k(\mathcal{Z}_k|\mathcal{X}_k)$, respectively. Then, the RFS formulation can be employed in the optimal multi-speaker Bayesian filter by propagating the posterior density using Bayes recursion. Nevertheless, the RFS approach is computationally intractable since multiple integrals are involved in the recursion of multi-speaker posterior and the computational complexity increases exponentially with the number of speakers. To alleviate the computational complexity, the PHD filter is proposed which propagates the first-order moment of the posterior instead of the posterior itself [32] as described next.

A. PHD Filter

The PHD filter is defined as the intensity $v_{k|k}$ whose integral gives the expected number of speakers. The PHD filter consists of two iterative steps: prediction and update. The prediction step of the PHD is shown as [32]

$$v_{k|k-1}(\mathbf{x}_k) = \xi_k(\mathbf{x}_k) + \int \phi_{k|k-1}(\mathbf{x}_k|\mathbf{x}_{k-1})\, v_{k-1|k-1}(\mathbf{x}_{k-1})\, d\mathbf{x}_{k-1} \tag{6}$$

where $\xi_k(\mathbf{x}_k)$ is the intensity function of the new speaker birth RFS Γ_k, and $\phi_{k|k-1}(\mathbf{x}_k|\mathbf{x}_{k-1})$ is the analog of the single-speaker state transition probability [32]

$$\phi_{k|k-1}(\mathbf{x}_k|\mathbf{x}_{k-1}) = p_{S,k}(\mathbf{x}_{k-1})\, f_{k|k-1}(\mathbf{x}_k|\mathbf{x}_{k-1}) + \beta_{k|k-1}(\mathbf{x}_k|\mathbf{x}_{k-1}) \tag{7}$$

where $p_{S,k}(\mathbf{x}_{k-1})$ denotes the survival probability for the speakers still existing and $f_{k|k-1}(\mathbf{x}_k|\mathbf{x}_{k-1})$ denotes the single-speaker state transition density. The intensity function of RFS $\mathcal{B}_k(\mathcal{X}_{k-1})$ is denoted by $\beta_{k|k-1}(\mathbf{x}_k|\mathbf{x}_{k-1})$ for the speaker spawned at time k with previous state $\mathbf{x}_{k-1}$. The PHD update is given as [32]

$$v_{k|k}(\mathbf{x}_k) = [1 - p_{D,k}(\mathbf{x}_k)]\, v_{k|k-1}(\mathbf{x}_k) + \sum_{\mathbf{z}_k \in \mathcal{Z}_k} \frac{p_{D,k}(\mathbf{x}_k)\, g_k(\mathbf{z}_k|\mathbf{x}_k)\, v_{k|k-1}(\mathbf{x}_k)}{\kappa_k(\mathbf{z}_k) + \int p_{D,k}(\mathbf{x}_k)\, g_k(\mathbf{z}_k|\mathbf{x}_k)\, v_{k|k-1}(\mathbf{x}_k)} \tag{8}$$

where $p_{D,k}(\mathbf{x}_k)$ denotes detection probability and $g_k(\mathbf{z}_k|\mathbf{x}_k)$ denotes the single-speaker likelihood defining the probability that $\mathbf{z}_k$ is generated by a speaker state $\mathbf{x}_k$. The intensity of clutter RFS $\mathcal{C}_k$ is given as $\kappa_k(\mathbf{z}_k) = \Psi \mathrm{u}(\mathbf{z}_k)$, where Ψ is the average number of Poisson clutter points per scan and $\mathrm{u}(\mathbf{z}_k)$ is the probability distribution of each clutter point.

The PHD recursion involves multiple integrals in (6) and (8) that have no closed-form solutions in general. To obtain a numerical solution, two implementation methods can be used, i.e., the Gaussian mixture PHD (GM-PHD) [47] and sequential Monte Carlo PHD (SMC-PHD) [33]. Different from the GM-PHD filter where a linear and Gaussian model is assumed, the SMC-PHD filter has the ability to handle non-linear and non-Gaussian problems in multi-speaker tracking. For this reason, we prefer the SMC-PHD algorithm, which is summarized next.

B. SMC-PHD Filter

At time step $k-1$, the PHD $v_{k-1|k-1}(\mathbf{x}_{k-1})$ is approximated by $\{w_{k-1}^{(n)}, \mathbf{x}_{k-1}^{(n)}\}_{n=1}^{N_{k-1}}$ of N_{k-1} particles and their corresponding weights as

$$v_{k-1|k-1}(\mathbf{x}_{k-1}) \approx \sum_{n=1}^{N_{k-1}} w_{k-1}^{(n)} \delta\left(\mathbf{x}_{k-1} - \mathbf{x}_{k-1}^{(n)}\right) \tag{9}$$

where $\delta(.)$ is a Dirac delta function. Prediction of the PHD $v_{k|k-1}(\mathbf{x}_k)$ is obtained with particles $\tilde{\mathbf{x}}_k$ and their weights $\tilde{w}_{k|k-1}$, $\{\tilde{w}_{k|k-1}^{(n)}, \tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{N_{k-1}+J_k}$. Here, N_{k-1} particles of $\tilde{\mathbf{x}}_k$ are first drawn from importance sampling $q_k(\tilde{\mathbf{x}}_k^{(n)}|\mathbf{x}_{k-1}^{(n)}, \mathcal{Z}_k)$ to propagate the particles from time step $k-1$, then J_k particles of $\tilde{\mathbf{x}}_k$ from the new born importance function $p_k(\tilde{\mathbf{x}}_k^{(n)}|\mathcal{Z}_k)$ are drawn to model the state of new speakers appearing in the scene. The PHD prediction is given as

$$v_{k|k-1}(\mathbf{x}_k) \approx \sum_{n=1}^{N_{k-1}+J_k} \tilde{w}_{k|k-1}^{(n)} \delta\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_k^{(n)}\right) \tag{10}$$

where J_k new particles are generated in the birth process. By replacing (9) into (6) and employing importance sampling, the predicted weights $\tilde{w}_{k|k-1}^{(n)}$ are obtained as [33]

$$\tilde{w}_{k|k-1}^{(n)} = \begin{cases} \dfrac{\phi_{k|k-1}\left(\tilde{\mathbf{x}}_k^{(n)}, \mathbf{x}_{k-1}^{(n)}\right) w_{k-1}^{(n)}}{q_k\left(\tilde{\mathbf{x}}_k^{(n)}|\mathbf{x}_{k-1}^{(n)}, \mathcal{Z}_k\right)}, & n = 1, ..., N_{k-1} \\[2ex] \dfrac{\xi_k\left(\tilde{\mathbf{x}}_k^{(n)}\right)}{J_k\, p_k\left(\tilde{\mathbf{x}}_k^{(n)}|\mathcal{Z}_k\right)}, & n = N_{k-1}+1, ..., N_{k-1}+J_k. \end{cases} \tag{11}$$

The update step of the PHD recursion is approximated by updating the weight of the predicted particles when the likelihood $g_k\left(\mathbf{z}_k|\tilde{\mathbf{x}}_k^{(n)}\right)$ is obtained. By substituting $v_{k|k-1}(\mathbf{x}_k)$ into (8), the predicted weights are updated as

$$\tilde{w}_k^{(n)} = \left[\left[1-p_D\left(\tilde{\mathbf{x}}_k^{(n)}\right)\right]+\sum_{\mathbf{z}_k\in\mathcal{Z}_k}\frac{p_D\left(\tilde{\mathbf{x}}_k^{(n)}\right)g_k\left(\mathbf{z}_k|\tilde{\mathbf{x}}_k^{(n)}\right)}{\kappa_k(\mathbf{z}_k)+C_k(\mathbf{z}_k)}\right]\tilde{w}_{k|k-1}^{(n)} \tag{12}$$

where

$$C_k(\mathbf{z}_k)=\sum_{j=1}^{N_{k-1}+J_k}p_D\left(\tilde{\mathbf{x}}_k^{(j)}\right)g_k\left(\mathbf{z}_k|\tilde{\mathbf{x}}_k^{(j)}\right)\tilde{w}_{k|k-1}^{(j)}. \tag{13}$$

Here, J_k new particles are sampled for the born speakers at each iteration and added to the old ones $N_k = N_{k-1}+J_k$ which increases the number of particles over time and makes the PHD filter intractable. Besides, to concentrate the particles on the zones around the speakers, the low weight particles need to be removed and particles with high weights should be duplicated. To this end, a resampling step is performed after the update step. N_k particles are resampled from $\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{N_{k-1}+J_k}$ where $\hat{\Xi}_{k|k}$ is the total mass and $\hat{\Xi}_{k|k}=\sum_{n=1}^{N_{k-1}+J_k}\tilde{w}_k^{(n)}$. N_k is estimated by $N_k=\eta\hat{\Xi}_{k|k}$ where η is the constant number of particles per speaker. So, the complexity of the SMC-PHD filter grows *linearly* with the number of speakers. After the resampling step, new weights of the set $\{w_k^{(n)}, \mathbf{x}_k^{(n)}\}_{n=1}^{N_k}$ are normalized to preserve the total mass.

The SMC-PHD filter propagates the surviving, spawned and born particles to model the new and existing speakers. Conventionally, these particles are used every frame which increases the computational complexity. To address this problem, we introduce audio information, i.e. the DOA data, into the visual SMC-PHF filter, as discussed next.

III. Audio-Visual Tracker With SMC-PHD Filter

The DOA data is introduced to the SMC-PHD filter based on [34] and [36] where the efficiency of the particles is improved under a particle filter framework by re-allocating all the particles around the DOA line which is drawn from the center of the microphone array to a point in the image frame estimated by the projection of DOA to $2D$ image plane. However, different from [34] and [36] in which the DOA is used in the same way for all the particles, here the contribution of the DOA information is varied depending on the type of the particles. Similar to [34] and [36], we also use the sam-spare-mean (SSM) method [48] for the DOA estimation which is further enhanced by a third-order Auto-Regressive (AR) model. We should note that there are other audio features and algorithms for extracting these features that could be used in our proposed system, however, exploring other audio detection methods is beyond the scope of this work.

To address the aforementioned complexity issue, we propose to generate the born particles only when the detection of a new speaker occurs via audio. In other words, we assume that the DOA information is available to control particle distribution. As a result, the born particles can be uniformly distributed around the DOA line as illustrated in Fig. 1(b), rather than over the whole image as in Fig. 1(a). In Fig. 1(a), the born particles are distributed to detect the speaker on the restricted region of the frame as this region covers both sides of the scene that new speaker may enter. The DOA is also used for the surviving and spawned particles to concentrate them around the DOA line. The missing DOA data is completed by interpolation in the case of a short silence. However, the DOA data will be lost when the speaker stops talking for a long time. Then, our proposed algorithm continues tracking without the DOA information. With re-allocation of the particles around the DOA line, speaker detection and tracking is likely improved since the DOA indicates the approximate direction of the sound emanating from the speaker.

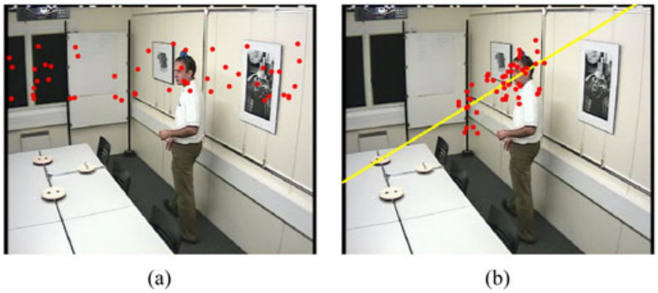

Fig. 1. Distribution of 50 particles for (a) the visual case and (b) the audio-visual case.

The surviving and spawned particles are defined as $\tilde{\mathbf{x}}_{s,k}$ for time k since the DOA information is used for surviving and spawned particles in the same way. In addition, the born particles are denoted as $\tilde{\mathbf{x}}_{b,k}$. The surviving particles from the previous iteration and the particles spawned from the surviving particles are distributed by a dynamic model given in (1). Details on the generation of the surviving, spawned and born particles can be found from [33] and [46].

If the DOA is available in current frame, the DOA line is drawn [34] and the perpendicular Euclidean distances $\mathbf{d}_k = [\,d_k^{(1)} \quad \ldots \quad d_k^{(N_{k-1})}\,]$ of the particles to the DOA line are computed. If there are multiple DOA lines, the one closest to the particles is chosen, as long as the distance to the DOA line is smaller than a pre-determined threshold to prevent the particles from converging to the DOA line which belongs to other speakers or is created due to noise or clutter. Then, the movement distances of the particles $\hat{\mathbf{d}}_k$ are calculated as [36]

$$\hat{\mathbf{d}}_k = \frac{\mathbf{d}_k}{\|\mathbf{d}_k\|_1}\odot\mathbf{d}_k \tag{14}$$

where $\|.\|_1$ is the ℓ_1 norm and $\odot$ is the element-wise product. $\hat{\mathbf{d}}_k$ is used to relocate the $\tilde{\mathbf{x}}_{s,k}$ particles around the DOA line

$$\tilde{\mathbf{x}}_{s,k} = \tilde{\mathbf{x}}_{s,k}\oplus\mathbf{h}_k\hat{\mathbf{d}}_k \tag{15}$$

where $\oplus$ is the element-wise addition and $\mathbf{h}_k = [\cos(\theta_k) \;\; 0 \;\; \sin(\theta_k) \;\; 0 \;\; 0]^T$ which is used to update only the position (x_1, x_2) of the particle state vector $[\,x_1 \;\; \dot{x}_1 \;\; x_2 \;\; \dot{x}_2 \;\; s\,]^T$ in order to provide the perpendicular movement to the DOA line.

Then, new speaker case is checked using the DOA information. If the number of DOA lines is larger than the number of estimated speakers in $k-1$, it may imply the presence of a new speaker in the scene. To detect the new speaker, J_k born particles $\tilde{\mathbf{x}}_{b,k}$ are generated and distributed uniformly around the new DOA line. The prediction step is employed to calculate the weights of particles $\tilde{w}_{k|k-1}$ after all the particles are combined under $\tilde{\mathbf{x}}_k$. Then, the update step is performed to calculate $\tilde{w}_k$ after the estimation of color likelihood. Assuming the noise on the color likelihood function to be Gaussian, the likelihood function of each measured color histogram can be written as [49]

$$g^{(m)}(\mathbf{z}_k|\mathbf{x}_k) \propto \mathcal{N}(\mathbf{z}_k; 0, \sigma_c^2) = \frac{1}{\sigma_c\sqrt{2\pi}} \exp\left\{-\frac{\{D^{(m)}(\mathbf{x}_k)\}^2}{2\sigma_c^2}\right\} \qquad (16)$$

where σ_c^2 is the variance of noise for the color likelihood, and $D^{(m)}(\mathbf{x}_k)$ are the color similarities calculated as the Bhattacharyya distance between the reference models and the candidate speaker, i.e.,

$$D^{(m)}(\mathbf{x}_k) = \sqrt{1 - \sum_{u=1}^{U} \sqrt{q_u(\mathbf{x}_k) r_u^{(m)}}} \qquad (17)$$

where $q_u(\mathbf{x}_k)$ is the color histogram for the state $\mathbf{x}_k$ extracted from the rectangle area centred around the location (x_k, y_k) on the frame by which the speaker candidate is defined, and $\{r_u^{(1)}, r_u^{(2)}, ..., r_u^{(M)}\}$, with u being the index of the histogram bins, are the color models of the speakers.

The number of estimated speakers is computed using the total mass which is the sum of the weights of the particles. After the resampling step is performed, the positions of the speakers are estimated using the K-means clustering algorithm. Lastly, the identity of the speakers is detected by measuring the similarity between the color histogram of the estimated speakers and that of the reference speakers. The pseudo code of the proposed AV-SMC-PHD filtering algorithm [44] is depicted in Algorithm 1.

Algorithm 1: Proposed AV-SMC-PHD filtering algorithm.

```
Initialize: $\eta$, $\sigma^2$, $U$, $T$, $\mathbf{F}$, $\lambda$, $r$, $u$, $p_M$, $p_D$, $\sigma_c^2$, $\sigma_s^2$, $p_S$,
$k$, $K$, $N_0$, $\mathbf{x}_0$
while $k < K$ do
    For $n = 1, ..., N_{k-1}$ sample $\tilde{\mathbf{x}}_k \sim q_k\left(\cdot|\mathbf{x}_{k-1}^{(n)}, \mathcal{Z}_k\right)$,
    where $\tilde{\mathbf{x}}_k \in \mathbb{R}^{5\times N_{k-1}}$
    Propagate surviving and spawned particles:
    $\tilde{\mathbf{x}}_{s,k} = \mathbf{F}\tilde{\mathbf{x}}_k + \mathbf{q}_k$
    Get the corresponding DOA angle $\theta_k$
    if DOA exists then
        // For surviving and spawned particles
        Calculate distances $\mathbf{d}_k = \left[d_k^{(1)} \quad ... \quad d_k^{(N_{k-1})}\right]^T$
        Calculate movement distances $\hat{\mathbf{d}}_k$ using Equation (14)
        Concentrate $\tilde{\mathbf{x}}_{s,k}$ around the DOA line :
        $\tilde{\mathbf{x}}_{s,k} = \tilde{\mathbf{x}}_{s,k} \oplus \mathbf{h}_k\hat{\mathbf{d}}_k$
        if new speaker then
            // For born particles
            For $n = N_{k-1}+1, ..., N_{k-1}+J_k$ sample
            $\tilde{\mathbf{x}}_{b,k} \sim p_k(\cdot|\mathcal{Z}_k)$ uniformly around the DOA line
        end
    end
    Combine all the particles:
    $\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_{s,k} \cup \tilde{\mathbf{x}}_{b,k}$
    Prediction: For $n = 1, ..., N_{k-1}+J_k$ calculate
    $\tilde{w}_{k|k-1}^{(n)}$ using Equation (11)
    Estimate colour likelihood using Equation (16)
    Update: For $n = 1, ..., N_{k-1}+J_k$ calculate $\tilde{w}_k^{(n)}$
    using Equation (12)
    Calculate the total mass $\hat{\Xi}_{k|k} = \sum_{n=1}^{N_{k-1}+J_k} \tilde{w}_k^{(n)}$
    Resampling: Resample $\left\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \tilde{\mathbf{x}}_k^{(n)}\right\}_{n=1}^{N_{k-1}+J_k}$ to
    get $\left\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \mathbf{x}_k^{(n)}\right\}_{n=1}^{N_k}$ where $N_k = \eta\hat{\Xi}_{k|k}$
    Multiply the weights by $\hat{\Xi}_{k|k}$ to get $\left\{\tilde{w}_k^{(n)}, \mathbf{x}_k^{(n)}\right\}_{n=1}^{N_k}$
    Cluster the particles and get the positions of the speakers
    $k = k + 1$
end
```

IV. Mean-Shift-Based AV-SMC-PHD Filtering

As mentioned earlier, the tracking performance of the PHD filter is compromised due to the first-order approximation of the RFS. To address this limitation, we propose a new and improved version of the AV-SMC-PHD algorithm based on the well-known mean-shift technique. The idea is to shift the particles to a local maximum of the distribution function so that they are closer to the speaker position, compared to their original positions. This is achieved by searching and climbing density gradients iteratively to find the peak of the probability distribution. This algorithm, which we name as AVMS-SMC-PHD, offers significant improvement over AV-SMC-PHD in terms of tracking accuracy. Despite its popularity, to our knowledge, mean-shift has never been used in the way as proposed here.

The mean-shift approach aims to find the target in the next image frame that is most similar to the initialised target (reference model) in the current frame by iteratively searching the next frame with a non-parametric kernel [22]. Such similarity is measured as the Bhattacharyya distance between the histogram of the target model and that of the candidate target in the next frame [23]. The mean-shift approach is originally designed for single target tracking [23]. For multi-target tracking, however, this approach needs to be adapted, which we propose to modify as follows.

A. Multiple-Speaker Mean-Shift

During tracking, the target is detected based on the comparison of the similarity between the pdf of the candidate target and the pdf of the reference model, measured by the Bhattacharyya distance

$$d(\mathbf{y}) = \sqrt{1 - \rho\left[\mathbf{q}(\mathbf{y}), \mathbf{r}\right]} \qquad (18)$$

where $\mathbf{r} = \{r_u\}_{u=1,\ldots,U}$ $(\sum_{u=1}^{U} r_u = 1)$ is the U-bin color histogram of the reference image of the target, $\mathbf{q}(\mathbf{y}) = \{q_u(\mathbf{y})\}_{u=1,\ldots,U}$ $(\sum_{u=1}^{U} q_u = 1)$ is the color histogram of the image region centered at the point $\mathbf{y}$, and $\rho[\mathbf{q}(\mathbf{y}), \mathbf{r}]$ is the Bhattacharyya coefficient, given by

$$\rho\left(\mathbf{y}\right) \equiv \rho\left[\mathbf{q}\left(\mathbf{y}\right), \mathbf{r}\right] = \sum_{u=1}^{U} \sqrt{q_u(\mathbf{y}) r_u} \tag{19}$$

where ρ takes values between 0 and 1, with a greater value representing a higher similarity in their pdfs.

Using Taylor expansion, the Bhattacharyya coefficient in (19) can be approximated as follows:

$$\rho\left(\mathbf{y}\right) \approx \frac{1}{2}\sum_{u=1}^{U} \sqrt{q_u\left(\mathbf{y}_0\right) r_u} + \frac{1}{2}\sum_{u=1}^{U} q_u\left(\mathbf{y}\right)\sqrt{\frac{r_u}{q_u\left(\mathbf{y}_0\right)}} \tag{20}$$

where $\mathbf{y}_0$ is the location of the target in the previous frame.

Using a kernel-based histogram representation for $q_u\left(\mathbf{y}\right)$ [23], equation (20) can be further written as

$$\rho(\mathbf{y}) \approx \frac{1}{2}\sum_{u=1}^{U} \sqrt{q_u\left(\mathbf{y}_0\right) r_u} + \frac{C_h}{2}\sum_{i=1}^{n_h} \mathcal{W}_i \, \mathscr{k}\left(\left\|\frac{\mathbf{y}-\mathbf{x}_i}{h}\right\|^2\right) \tag{21}$$

where C_h is a normalization constant, $\mathscr{k}$ is the kernel function (giving higher weights to the pixels at the center of the target region), and $\mathcal{W}_i$ are the weights given by

$$\mathcal{W}_i = \sum_{u=1}^{U} \sqrt{\frac{r_u}{q_u\left(\mathbf{y}_0\right)}} \delta[b(\mathbf{x}_i) - u] \tag{22}$$

where $b(\mathbf{x}_i)$ is a function which assigns one of the histogram bins to a given color at location $\mathbf{x}_i$. By employing the mean-shift procedure [22], we can find the mode of the density in the neighbourhood of $\mathbf{x}_i$. In this procedure, the kernel is applied recursively from the current location $\mathbf{y}_0$ to the next, i.e. $\mathbf{y}_1$, which is related to $\mathbf{y}_0$ as follows:

$$\mathbf{y}_1 = \frac{\sum_{i=1}^{n_h} \mathbf{x}_i \mathcal{W}_i g(\|\frac{\mathbf{y}_0 - \mathbf{x}_i}{h}\|)}{\sum_{i=1}^{n_h} \mathcal{W}_i g(\|\frac{\mathbf{y}_0 - \mathbf{x}_i}{h}\|)} \tag{23}$$

where $g(x) = -\mathscr{k}'(x)$, and $\mathscr{k}'(x)$ is the derivative of $\mathscr{k}(x)$ assuming that $\mathscr{k}'(x)$ exists for all $x \in [0, \infty)$, except for a finite set of points.

Several kernels could be used such as normal, uniform and Epanechnikov. As recommended in [23], we choose the Epanechnikov kernel here which is defined as

$$\mathscr{k}\left(\mathbf{x}\right) = \begin{cases} \left(1 - \mathbf{x}\right), & \|\mathbf{x}\| \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

In this case, its derivative $g(x)$ a constant

$$g(x) = -\mathscr{k}'\left(\mathbf{x}\right) = \begin{cases} 1, & \|\mathbf{x}\| \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Hence, (23) reduces to

$$\mathbf{y}_1 = \frac{\sum_{i=1}^{n_h} \mathbf{x}_i \mathcal{W}_i g(\|\frac{\mathbf{y}_0 - \mathbf{x}_i}{h}\|)}{\sum_{i=1}^{n_h} \mathcal{W}_i g(\|\frac{\mathbf{y}_0 - \mathbf{x}_i}{h}\|)} = \frac{\sum_{i=1}^{n_h} \mathbf{x}_i \mathcal{W}_i}{\sum_{i=1}^{n_h} \mathcal{W}_i}. \tag{26}$$

B. Particle Distribution With Mean-Shift

The above mean-shift algorithm can then be used to distribute the particles in AV-SMC-PHD filter, as follows. First, the iteration is initialized before it is run over the nth particle $\tilde{\mathbf{x}}_k^{(n)}$ at the time frame k. In this step, the horizontal and vertical positions within $\tilde{\mathbf{x}}_k^{(n)}$ are assigned to $\mathbf{y}_0 = [\tilde{\mathbf{x}}_k^{(n)}(1), \tilde{\mathbf{x}}_k^{(n)}(3)]$ since $\tilde{\mathbf{x}} = [\, x_1 \; \dot{x}_1 \; x_2 \; \dot{x}_2 \; s \,]^T$. The candidate speaker model, $q_u\left(\mathbf{y}_0\right)$, evaluated at the centre $\mathbf{y}_0$, is compared with all the models from the reference template. The closest reference model $\mathbf{r} = \{r_u\}_{u=1,\ldots,U}$ is then selected and used to move the particle towards the speaker in the following steps of the mean-shift iteration. The Bhattacharyya coefficient $\rho[\mathbf{q}\left(\mathbf{y}_0\right), \mathbf{r}]$ is calculated before running the mean-shift iteration for the particle $\tilde{\mathbf{x}}_k^{(n)}$. The mean-shift iteration is performed in a loop controlled by two parameters, namely, iteration flag $ContIter$ and iteration number $NumIter$.

The iteration of the mean-shift process continues if the predefined condidtions for the two parameters are satisfied. Otherwise, the loop will be broken and the same process will be repeated for the next particle with $NumIter$ set to 0 and $ContIter$ set $true$. In the first step of the loop, the weights are derived according to (22). Then, the next location $\mathbf{y}_1$ is calculated via (26). In practice, it may not be the correct direction to move the particle towards $\mathbf{y}_1$. To avoid this issue, the Bhattacharyya coefficient of $\mathbf{y}_1$, $\rho[\mathbf{q}(\mathbf{y}_1), \mathbf{r}]$, is calculated and compared with $\rho[\mathbf{q}(\mathbf{y}_0), \mathbf{r}]$. If $\rho[\mathbf{q}(\mathbf{y}_1), \mathbf{r}] < \rho[\mathbf{q}(\mathbf{y}_0), \mathbf{r}]$, it means that $\mathbf{y}_1$ is not a good estimate and the loop will be broken. If $\rho[\mathbf{q}(\mathbf{y}_1), \mathbf{r}] > \rho[\mathbf{q}(\mathbf{y}_0), \mathbf{r}]$, the amount of shifting needs to be checked. If $\|\mathbf{y}_1 - \mathbf{y}_0\| > \zeta$, where ζ is a threshold value, $\mathbf{y}_1$ is set to $\mathbf{y}_0$ and then used for the next iteration provided that $NumIter < MaxIter$. If $\|\mathbf{y}_1 - \mathbf{y}_0\| < \zeta$, the iteration will be broken again. This process is repeated for all the particles.

We refer to the above process as $\mathbb{MS}$, which is performed after the audio contribution is considered, and all the born, spawned and surviving particles are combined as $\tilde{\mathbf{x}}_k$. This enables the $\mathbb{MS}$ process to be applied to all types of the particles even without the DOA information. The pseudo code of the $\mathbb{MS}$ is given in Algorithm 2.

The algorithm is also illustrated in Fig. 2. Here, suppose 10 particles $\tilde{\mathbf{x}}_k^{(n)}$, $n = 1, ..., 10$, are given, which have different Bhattacharyya coefficients $\rho\left[\mathbf{q}\left(\mathbf{y}_0\right), \mathbf{r}\right]$. With these coefficients, the mean-shift iteration is performed to move the particles to the local maxima of the measurement function. As a result, the particles $\hat{\mathbf{x}}_k$ that have higher values of the Bhattacharyya coefficients tend to be closer to the speaker position.

In the end, the shifted particles provide good local characterization of the likelihood which allows the multi-mode distribution to be maintained with the use of a fewer number of particles [50]. After the $\mathbb{MS}$, the step of weight prediction is performed, followed by the remaining steps as in the AV-SMC-PHD filter.

V. Sparse Sampling for AVMS-SMC-PHD Filtering

The use of mean shift in AV-SMC-PHD leads to a reduction in tracking error (to be demonstrated in Section VI). However, the computational cost is increased due to the application of the

Algorithm 2: $\mathbb{MS}$ function for the mean-shift iteration.

```
Given: $\tilde{\mathbf{x}}_k$, $N_{k-1}$, $J_k$, $MaxIter$, $U$, $\zeta$
for $n = 1, ...N_{k-1} + J_k$ do
    Assign position coordinates of the particle to
    $\mathbf{y}_0 = [\tilde{\mathbf{x}}_k^{(n)}(1), \tilde{\mathbf{x}}_k^{(n)}(3)]$
    Find the closest reference model $r_u$ for $\tilde{\mathbf{x}}_k^n$ by
    comparing the candidate speaker model $q_u(\mathbf{y}_0)$ and
    the reference models.
    Evaluate: $\rho\left[\mathbf{q}\left(\mathbf{y}_0\right), \mathbf{r}\right] = \sum_{u=1}^{U} \sqrt{q_u(\mathbf{y}_0) r_u}$
    Set iteration number: $NumIter = 0$;
    Set iteration flag: $ContIter = true$;
    while $ContIter == True$ or $NumIter < MaxIter$ do
        Derive the weights according to Equation (22)
        Find the next location $\mathbf{y}_1$ by Equation (26)
        Compute: $\rho\left[\mathbf{q}\left(\mathbf{y}_1\right), \mathbf{r}\right] = \sum_{u=1}^{U} \sqrt{q_u(\mathbf{y}_1) r_u}$
        // Continue mean-shift iteration as long as the
        Bhattacharyya coefficient goes up. Otherwise,
        stop iteration
        if $\rho\left[\mathbf{q}\left(\mathbf{y}_1\right), \mathbf{r}\right] > \rho\left[\mathbf{q}\left(\mathbf{y}_0\right), \mathbf{r}\right]$ then
            // If position change exceeds the threshold
            value $\zeta$, then continue mean-shift iteration.
            Otherwise, stop iteration
            if $\|\mathbf{y}_1 - \mathbf{y}_0\| > \zeta$ then
                $\mathbf{y}_0 = \mathbf{y}_1$;
                $ContIter = true$;
            else
                $ContIter = false$;
            end
        else
            $\mathbf{y}_1 = \mathbf{y}_0$
            $ContIter = false$;
        end
        $NumIter = NumIter + 1$;
    end
    $[\tilde{\mathbf{x}}_k^{(n)}(1), \tilde{\mathbf{x}}_k^{(n)}(3)] = \mathbf{y}_1$;
end
```

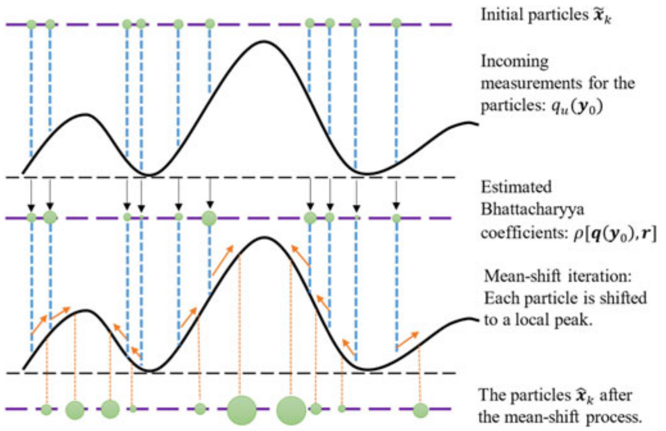
Initial particles $\tilde{\mathbf{x}}_k$
Incoming measurements for the particles: $q_u(\mathbf{y}_0)$
Estimated Bhattacharyya coefficients: $\rho[\mathbf{q}(\mathbf{y}_0), \mathbf{r}]$
Mean-shift iteration: Each particle is shifted to a local peak.
The particles $\tilde{\mathbf{x}}_k$ after the mean-shift process.

Fig. 2. Mean-shift process for the particles.

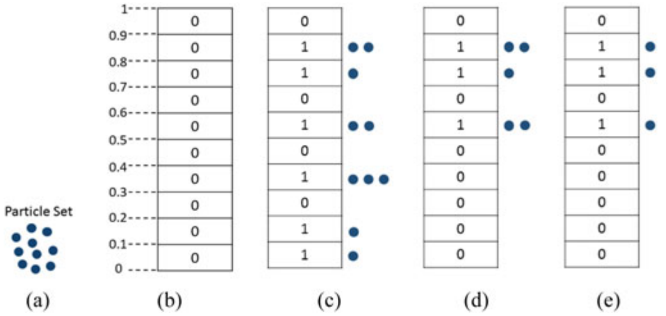

Fig. 3. (a) Set of 10 particles is given. (b) Initial $\mathbb{B}$ is illustrated. (c) Distribution of the particles on $\mathbb{B}$ according to Bhattacharyya coefficients. (d) Updated $\mathbb{B}$ after the estimation of the number of particles with KLD-Sampling is given. (e) Sparse particle selection is shown.

mean shift to all the particles [41]. To reduce the complexity, we introduce a sampling technique to select a subset of most relevant particles before applying the mean-shift, leading to a filtering algorithm of improved computational efficiency and similar accuracy, termed as sparse-AVMS-SMC-PHD.

More specifically, one dimensional bins, $\mathbb{B}$ with τ subintervals are created on the interval $[0, 1]$. Each subinterval, denoted by $\mathbb{B}_i$, has a range of $[\{1 - (i-1)/\tau\}, \{1 - i/\tau\}]$. The number of bins in $\mathbb{B}$, and hence the choice of τ, will affect the sparse sampling results. Experimental studies suggest that, as a practical choice [51], τ can be set to a constant number of particles per speaker, η.

Fig. 3 gives a demonstration of the proposed sampling algorithm. Suppose a set of 10 particles is given in Fig. 3(a), and $\mathbb{B}$ is created for this set, as illustrated in Fig. 3(b). We can consider η as 10, and as a result, τ can also be set to 10. Hence, each bin is allocated a subinterval with a length of $1/10 = 0.1$, and these subintervals are sorted in a descending order, as shown on the left side of Fig. 3(b). The top subinterval has a range from 0.9 to 1 while the bottom subinterval is ranged from 0 to 0.1.

At the beginning, each bin is set to $\mathbb{B}_{i=1,\ldots,\tau} = 0$ as it is empty. For each particle, the Bhattacharyya coefficient ρ is then calculated using (19), which has a value between $[0, 1]$ with 1 being the best similarity matching between the reference and candidate histograms, and 0 the worst. After ρ is obtained, its corresponding range $\mathbb{B}_i$ is found and updated from 0 to 1. Then, the number of particles, $\mathfrak{N}_i$, in $\mathbb{B}_i$ is increased by one. In the end, ρ is estimated for all the particles and the allocation of these particles in $\mathbb{B}$ is shown in Fig. 3(c), where the particles are sorted in terms of the values of ρ. Such representation resembles sparse categorization of the particles, hence the proposed method is referred to as "sparse sampling". Here, the total number of true bins b is estimated via $b = \sum_{i=1}^{\tau} \mathbb{B}_i$, which, in this case, gives a value of 6 since only 6 subintervals are updated from 0 to 1, as shown in Fig. 3(c).

The number of bins b can be used to estimate the number of particles, $\mathscr{N}$, as follows [52]:

$$\mathscr{N} = \frac{b-1}{\epsilon}\left\{1 - \frac{2}{9(b-1)} + \sqrt{\frac{2}{9(b-1)}}\, z_{1-\delta}\right\}^3 \tag{27}$$

where ϵ is the upper error bound given by the KL-divergence, b is the number of bins, and $z_{1-\delta}$ is the upper $1-\delta$ quantile of the standard normal distribution $\mathcal{N}(0, 1)$.

Algorithm 3: $\mathbb{SS}$ function for sparse sampling.

```
Given: $\tilde{\mathbf{x}}_k$, $N_{k-1}$, $J_k$, $\tau$, $\epsilon$, $z_{1-\delta}$
Create $\mathbb{B}$ with $\tau$ subintervals.
for $j = 1, ...N_{k-1} + J_k$ do
    Calculate $\rho_j$ using Equation (19)
    Find $i$ where $\rho_j \in [\{1 - (i-1)/\tau\}, \{1 - i/\tau\}]$
    Set $\mathbb{B}_i = 1$
    Increase the particle counter for $\mathbb{B}_i$, $\mathfrak{N}_i = \mathfrak{N}_i + 1$
end
Estimate the total bin number. $b = \sum_{i=1}^{\tau} \mathbb{B}_i$
Calculate $\mathscr{N}$ using Equation (27)
Choose the sparse particles $\bar{\mathbf{x}}_k$ by taking one particle
from the subintervals $\mathbb{B}_{i=1:t}$ where $t$ is the upper bound
for $\sum_{i=1}^{t} \mathfrak{N}_i = \mathscr{N}$
```

In this example, $\mathscr{N}$ is estimated as 5, which means that only 5 particles need to be chosen from the right side of Fig. 3(c). This selection starts from the top subintervals of $\mathbb{B}$ as the particles are already sorted according to ρ in a descending order. In Fig. 3(d), $\mathbb{B}$ is updated by removing all the particles except those in the first t top subintervals for which $\sum_{i=1}^{t} \mathfrak{N}_i = \mathscr{N}$. Therefore, here $t = 5$ since only the first 5 subintervals contain $\mathscr{N} = 5$ particles. These five particles can be employed as the "sparse particles". However, we take one step further in the selection of the sparse particles. It is observed that the mean-shift process to be performed afterwards tends to move the particles within the same subinterval to the same local maxima. As a result, it is unnecessary to have more than one particle from the same interval. Therefore, one particle from each of the intervals that have non-zero number of particles is chosen and added to the sparse particles $\bar{\mathbf{x}}_k$, as illustrated in Fig. 3(e). It can be seen from the above example that (27) plays a key role in the estimation of the number of particles, $\mathscr{N}$, for generating a smaller subset from the source particles.

The sparse sampling process is denoted as function $\mathbb{SS}$. The pseudo code of $\mathbb{SS}$ is presented in Algorithm 3.

The $\mathbb{SS}$ function is integrated into the AVMS-SMC-PHD filter as follows. First, all the particles are combined as $\tilde{\mathbf{x}}_k$ after incorporating the DOA contribution. Then the $\mathbb{SS}$ is applied to obtain the sparse particles $\bar{\mathbf{x}}_k$

$$\bar{\mathbf{x}}_k = \mathbb{SS}\left(\tilde{\mathbf{x}}_k\right) \tag{28}$$

before applying the $\mathbb{MS}$ operation. Since $\bar{\mathbf{x}}_k$ has less particles than $\tilde{\mathbf{x}}_k$, the number of particles used in the mean-shift iteration is reduced with the sparse sampling method, which leads to a significant reduction in the computational cost.

VI. Experimental Results

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. Setup and Performance Metric

Several publicly available audio-visual datasets could be used for the evaluation of the proposed algorithms, such as "$AV16.3$" [53], "CLEAR" [54], "AMI" [55], and "SPEVI" [56]. However, there are several requirements in our evaluations that narrow down the choice of the suitable datasets. First, the dataset should consist of real-world scenarios with both audio and video sequences.

Second, the calibration information of the cameras should be available for the projection of DOA from the physical space to the image plane. Third, the audio detection and localization algorithm employed here is compatible only with circular microphone arrays.

Finally, apart from these physical features, the dataset should contain challenging scenarios such as occlusion and rapid movements of the speakers, and audio-visual sequences with mostly talking speakers. Here, having mostly talking speakers enables the DOA information to be detected and used for generating the born particles. Among these datasets, the $AV16.3$ offers the best fit to the requirements. Therefore, sequences from $AV16.3$ are mostly used for quantitative evaluation of the baseline and proposed algorithms. To show the flexibility of the proposed algorithms, sequences from the AMI and CLEAR datasets are also used in our tests, as shown in Section VI-D.

The $AV16.3$ consists of sequences where the speakers are moving and speaking at the same time whilst being recorded by three calibrated video cameras and two circular eight-element microphone arrays. The audio and video were recorded at 16 kHz and 25 Hz, respectively, and synchronized before being used in our system. The size of each image frame is 288 × 360 pixels. The speakers wear a colored ball for annotation purpose, which however is never used in our tracking algorithms. All the algorithms are tested for two and three speaker cases with all three different camera angles of four sequences: Sequences 24, 25, 30 and 45, which are the most challenging sequences in term of movements of the speakers and the number of occlusions.

To measure the tracking performance, the Optimal Subpattern Assignment for Tracks (OSPA-T) metric [57] is employed which is widely used for the evaluation of multi-speaker tracking systems. The OSPA-T is an extension of the OSPA metric [58] for tracking management evaluation. The OSPA metric, which uses a penalty to transfer the cardinality error into the state error, is able to evaluate the performance in both source number estimation and speaker position estimation

$$e_{\text{OSPA}}(\hat{\mathcal{X}}_k, \mathcal{X}_k) = \min_{\pi \in \Pi_{\hat{\Xi}_k, \Xi_k}} \sqrt[a]{\frac{1}{\Xi_k}\left(\sum_{i=1}^{\hat{\Xi}_k} \bar{d}^{(c)}(\hat{\mathbf{x}}_{i,k}, \mathbf{x}_{\pi_i,k})^a + c^a(\Xi_k - \hat{\Xi}_k)\right)} \tag{29}$$

where it is assumed that $\hat{\mathcal{X}}_k = \{\hat{\mathbf{x}}_{1,k}, ..., \hat{\mathbf{x}}_{\hat{\Xi}_k,k}\}$ is an estimation of the ground truth state set $\mathcal{X}_k = \{\mathbf{x}_{1,k}, ..., \mathbf{x}_{\Xi_k,k}\}$ and $\Pi_{\hat{\Xi}_k,\Xi_k}$ is the set of maps $\pi : 1, ..., \hat{\Xi}_k \to 1, ..., \Xi_k$. Here the state cardinality estimation $\hat{\Xi}_k$ may not be the same as the ground truth

Ξ_k. The OSPA error given in Equation (29) is for $\hat{\Xi}_k \leq \Xi_k$. If $\Xi_k < \hat{\Xi}_k$, then $e_{\text{OSPA}}(\hat{\mathcal{X}}_k, \mathcal{X}_k) = e_{\text{OSPA}}(\mathcal{X}_k, \hat{\mathcal{X}}_k)$. The function $\bar{d}^{(c)}(\cdot)$ is defined as $\min(c, \bar{d}(\cdot))$ where c is the cut-off value which determines the relative weighting of the penalties assigned to cardinality and localization errors. In addition, a describes the metric order which determines the sensitivity to outliers.

In addition to the OSPA-T metric, we used the Wasserstein distance [59] to enable the comparison of the proposed algorithms with another baseline algorithm by Pham *et al.* [15], since the results reported in this baseline algorithm are based on the Wasserstein distance. These results are given in Section VI-C.

In our evaluations, we used twelve multi-speaker sequences and the average results are shown in Table II at the end of this section. As it is not feasible to plot the results of all these sequences, only the results of two sequences are illustrated by the plots. The first one is Sequence 24 camera #1 where two moving speakers are walking back and forth, crossing the field of view twice and occluding each other. The second is Sequence 45 camera #3 where three moving speakers occlude each other many times. In these two sequences, the speakers are speaking continuously and the number of speakers varies between 0 to 3. Here, the experiments are run on Intel core $i7$ 2.2 GHz processor with 8 GB memory under Windows 7 operating system. Each experiment is repeated 10 times and the average results are presented with plots and tables.

B. V-SMC-PHD Versus AV-SMC-PHD on AV16.3

First, we compare between the V-SMC-PHD and AV-SMC-PHD filters. The parameters for the SMC-PHD are set as: $p_D = 0.98$, $p_S = 0.99$, $\lambda = 0.26$ and $\sigma_c = 0.1$. The uniform density u is $(360 \times 280)^{-1}$ and the number of particles per speaker is $\eta = 50$. In this case, the cut-off parameter $c = 65$, the OSPA-T metric order parameter $a = 2$. These parameters are set empirically based on extensive experimental studies in [51], where the impact of these parameters on the tracking performance is also studied and is omitted here for space limitations.

To show the computational efficiency of these two filtering algorithms, we ran experiments on Sequence 24 camera #1 and Sequence 45 camera #3. The number of particles per speaker changes from 25 to 500. The experiments are repeated 10 times and Fig. 4 shows the average time costs.

It can be observed that the computational cost of V-SMC-PHD is higher than that of AV-SMC-PHD and they both increase with the number of particles. The time required for processing Sequence 45 is higher than for Sequence 24 since the maximum number of speakers to be tracked is three in Sequence 45 while it is two in Sequence 24. Using audio information introduces some computational cost, however, as shown in Fig. 4, this cost is negligible as compared with that for propagating the particles. In fact, AV-SMC-PHD is computationally more efficient than V-SMC-PHD as the born particles are propagated only when necessary.

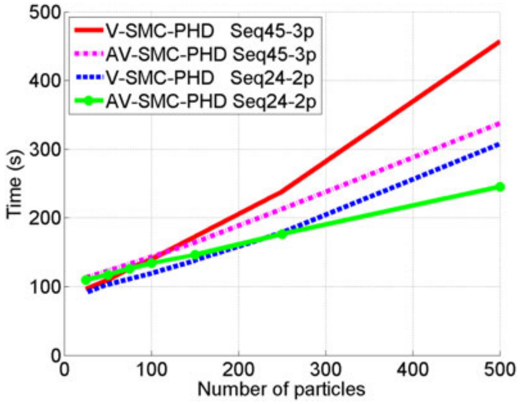

Fig. 4. Computational cost of the V-SMC-PHD and the proposed AV-SMC-PHD filters measured on Sequence 24 camera #1 and Sequence 45 camera #3.

The following experiments aim to investigate the estimation accuracy of the algorithms. Some frames from Sequence 24 camera #1 are shown in Fig. 5. The first row shows the results of V-SMC-PHD, while the second row for AV-SMC-PHD.

In the first two columns, both speakers are detected by our proposed AV-SMC-PHD filter while only one speaker is detected by the V-SMC-PHD filter. After occlusion, in the third and fourth columns, our proposed AV-SMC-PHD filter tracks the speakers more accurately. In the fifth column, the DOA information is available only for one speaker while in the last column, there is no DOA information. Nevertheless, our proposed AV-SMC-PHD filter is still able to track both speakers while the V-SMC-PHD filter fails to track one of them. This can be explained by the fact that the surviving particles are always dense until the DOA information is lost and more particles survive for the next frame in the AV-SMC-PHD filter. Even when the DOA information no longer exists after some points, the AV-SMC-PHD filter still has an advantage over the V-SMC-PHD filter on the number of surviving particles. Fig. 6 shows the estimation of the number of speakers. Here, we performed down-sampling to the plots for better visualization. The number of active speakers varies from 2 to 0, and as can be observed, our proposed AV-SMC-PHD filter gives better performance than the V-SMC-PHD filter.

The same experiments are performed on Sequence 45 camera #3 and some chosen frames are given in Fig. 7. Here, occlusion happens between the three speakers many times and the AV-SMC-PHD filter is able to detect and follow all the speakers even after the occlusions. The number of speakers estimated for Sequence 45 camera #3 is given in Fig. 8. Similarly, we can observe the improved performance of the AV-SMC-PHD filter over the V-SMC-PHD filter.

Fig. 9(a) and Fig. 9(b) show the OSPA-T errors for Sequence 24 camera #1 and Sequence 45 camera #3, respectively, averaged over 10 experiments. In Fig. 9(a), the average OSPA-T error is 27.12 for V-SMC-PHD and 17.71 for AV-SMC-PHD. This means that AV-SMC-PHD offers 34.69% improvements over V-SMC-PHD for Sequence 24 camera #1. The average OSPA-T errors for Sequence 45 camera #3 in Fig. 9(b)

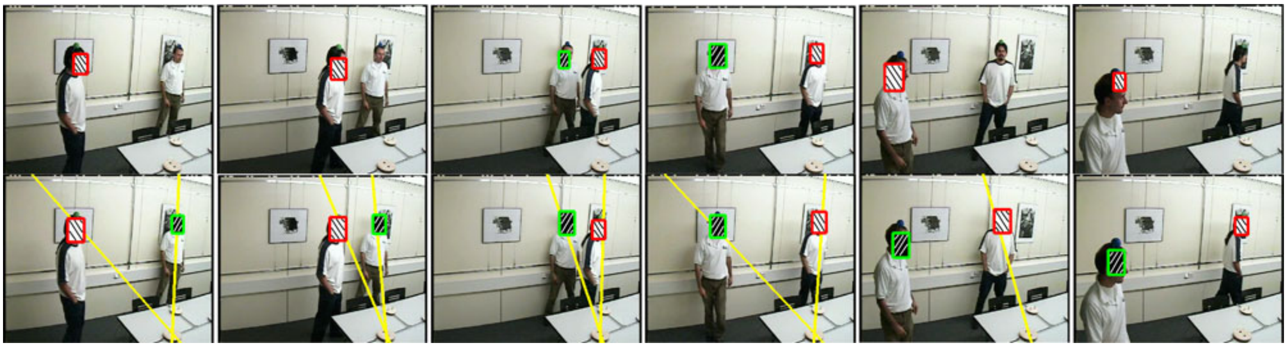

Fig. 5. Sequence 24 camera #1: two speakers with occlusions. The first row shows the results of the V-SMC-PHD filter and the second row for our proposed AV-SMC-PHD filter.

are 39.09 and 28.43 for V-SMC-PHD and AV-SMC-PHD, respectively. In this case, AV-SMC-PHD offers a 27.27% improvement over V-SMC-PHD.

C. AVMS-SMC-PHD and Sparse-AVMS-SMC-PHD on AV16.3

To make a fair comparison, we use the same parameters as those in the previous section for the evaluation of the AVMS-SMC-PHD filter. The mean-shift method has two specific parameters, i.e. the threshold for shifting distance ζ and the maximum number of iterations, which are set to 0.5 and 6, respectively. These parameters are set empirically based on extensive experimental studies in [51], similar to the experiments presented in the previous section.

The proposed AVMS-SMC-PHD algorithm was tested on Sequence 24 camera #1 and Sequence 45 camera #3. Because of the space constraints, plots for the AVMS-SMC-PHD could not be presented separately. Numerically, the AVMS-SMC-PHD algorithm gives an average error of 13.93 for Sequence 24 camera #1, resulting a 21.33% performance improvement over AV-SMC-PHD and a 48.64% improvement over V-SMC-PHD. For Sequence 45 camera #3, the average error by the AVMS-SMC-PHD algorithm is 22.43, showing an improvement over V-SMC-PHD and AV-SMC-PHD by 42.61% and 21.10%, respectively.

The proposed sparse-AVMS-SMC-PHD algorithm was tested with the same sequences and parameters as those used in the AVMS-SMC-PHD algorithm. In addition, the design parameters of Equation (27) are set to $\epsilon = 0.25$ and $z_{1-\delta} = 0.99$ based on empirical tests. A key parameter in sparse-AVMS-SMC-PHD is the dimension of the bins τ which may cause either performance failure, or an increase in the computation cost, depending on its size. To get a practical guidance for the selection of τ, pilot simulations were conducted on Sequence 24 camera #1 and Sequence 45 camera #3. We found that it seems to be reasonable to set τ as η. More details about these simulations can be found in [51].

As discussed earlier, the motivation for using sparse sampling is to reduce the computational cost of the AVMS-SMC-PHD filter. To this end, we measure the computational cost of the V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD and sparse-AVMS-SMC-PHD filters when applied to Sequence 24 camera #1, as illustrated in Fig. 10.

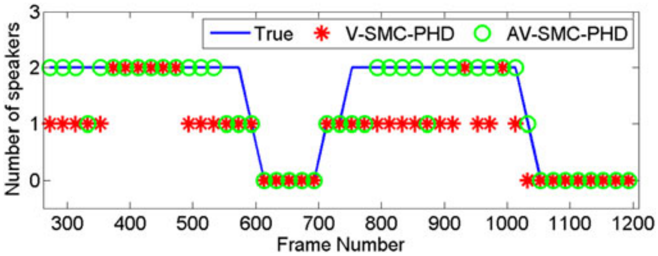

Fig. 6. Comparison of the V-SMC-PHD and the proposed AV-SMC-PHD filters in estimation of the number of speakers for Sequence 24 camera #1.

The integration of mean-shift to the AV-SMC-PHD causes a dramatic increase in computational cost. However, using sparse particles with the mean-shift iteration reduces the computational cost significantly by approximately 10 times, as the sparse sampling process generates a small subset from the source particles. To see the estimation accuracy, the proposed sparse-AVMS-SMC-PHD algorithm is further compared with the previous algorithms, the results on Sequence 24 camera #1 and Sequence 45 camera #3 are plotted all together in Fig. 11, which depicts the mean absolute error at each time step.

From this figure, it can be observed that sparse-AVMS-SMC-PHD and AVMS-SMC-PHD filters perform better than the AV-SMC-PHD and V-SMC-PHD filters, and the AVMS-SMC-PHD filter is slightly better than the sparse-AVMS-SMC-PHD filter. All the three algorithms perform better on Sequence 24 than on Sequence 45. This result is not surprising as the three-speaker sequence is more complex in terms of the movement of the speakers and the number of occlusions, which result in an increase in the estimation error.

A bar plot is also given in Fig. 12 to show the average results of the four algorithms over all the frames. According to these plots, the AVMS-SMC-PHD filter performs only 3.94% and 5.74% better than the sparse-AVMS-SMC-PHD filter for Sequence 24 camera #1 and Sequence 45 camera #3, respectively.

These trackers are also run over the remaining sequences and the results are given in Table I. The average error for V-SMC-PHD and AV-SMC-PHD is 32.01 and 22.75 respectively, which shows that with the contribution of audio, 28.93% reduction in tracking error has been achieved. This clearly demonstrates

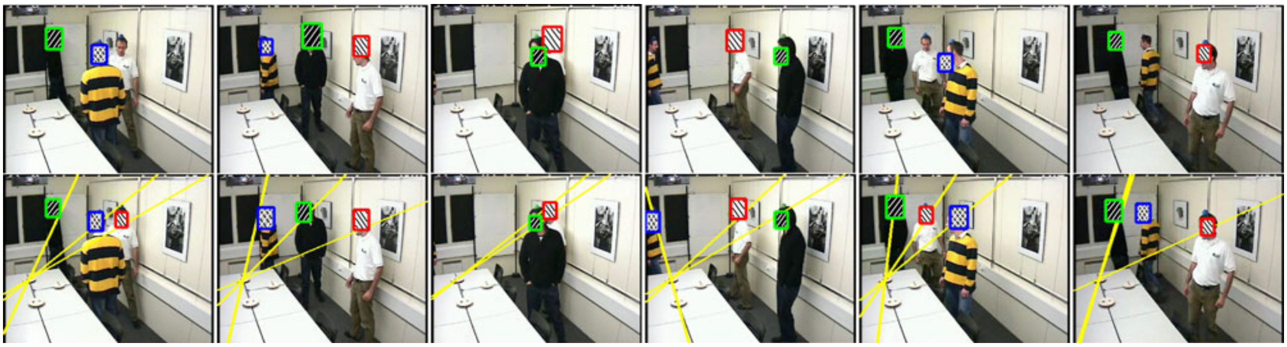

Fig. 7. Sequence 45 camera #3: three speakers with occlusions. The first and second row show the tracking results of the V-SMC-PHD and the proposed AV-SMC-PHD filter, respectively.

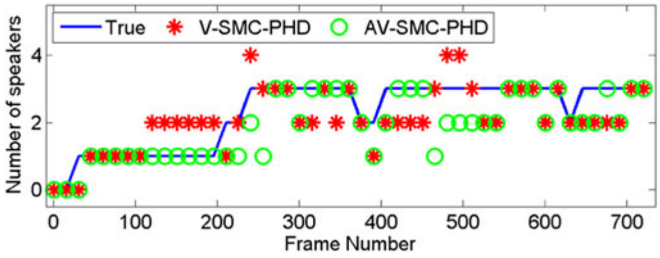

Fig. 8. Comparison between V-SMC-PHD and AV-SMC-PHD for estimating the number of speakers in Sequence 45 camera #3.

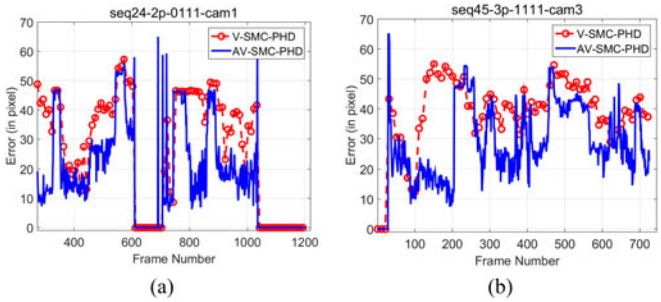

Fig. 9. Performance comparison of the V-SMC-PHD and the proposed AV-SMC-PHD filters in terms of the OSPA-T error. The data points on the V-SMC-PHD curve were down sampled for better visualisation.

that adding the audio information to the visual tracker leads to improvement in performance. In addition, Table I shows that the AVMS-SMC-PHD filter improves the estimation accuracy by 24.02% and 46.00% over the AV-SMC-PHD and V-SMC-PHD algorithms, respectively. Taking the average of all the experiments, sparse-AVMS-SMC-PHD outperforms AV-SMC-PHD and V-SMC-PHD by 18.96% and 42.41%, respectively. Its performance is slightly reduced by 6.65% as compared with AVMS-SMC-PHD. However, it is a reasonable sacrifice, given a ten-fold reduction in the computational cost as shown in Fig. 10. To further understand the cost reduction offered by the sparse-AVMS-SMC-PHD, we calculated the total number of particles used in each frame of the whole sequence for both AVMS-SMC-PHD and sparse-AVMS-SMC-PHD algorithms. In AVMS-SMC-PHD, the total number of particles for all the

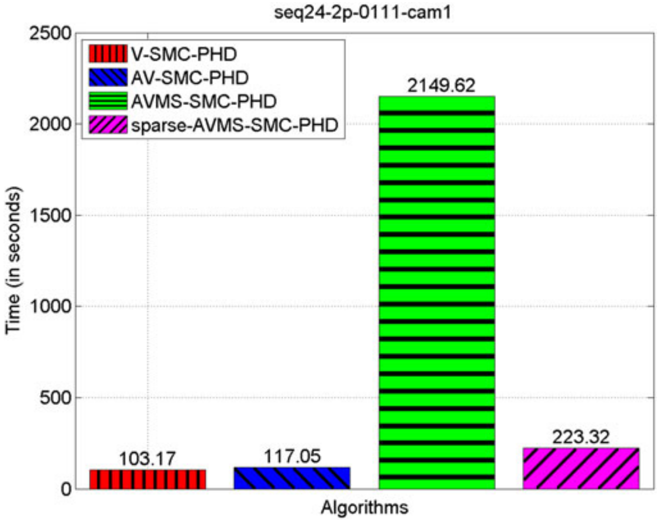

Fig. 10. Computational cost comparison between the V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD, and sparse-AVMS-SMC-PHD filters.

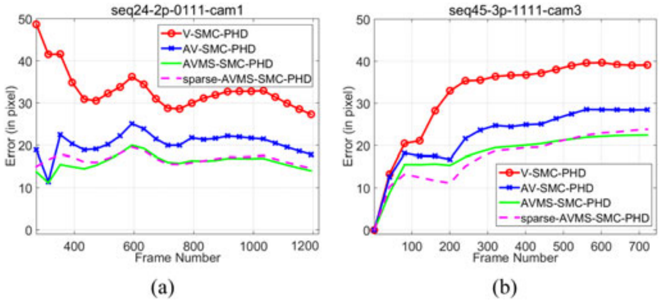

Fig. 11. Comparison of the V-SMC-PHD and the proposed algorithms AV-SMC-PHD, AVMS-SMC-PHD, and sparse-AVMS-SMC-PHD using mean absolute OSPA-T error. The data points on the V-SMC-PHD curve were down sampled for better visualisation.

speakers is 61853, while in sparse-AVMS-SMC-PHD it is 6301. With the proposed sparse sampling, the number of particles has been reduced to almost 10% which, in other words, leads to a ten-fold improvement in computational efficiency.

In addition, we used another baseline algorithm [18] for comparison. The results available in [18] cover only Sequences 24,

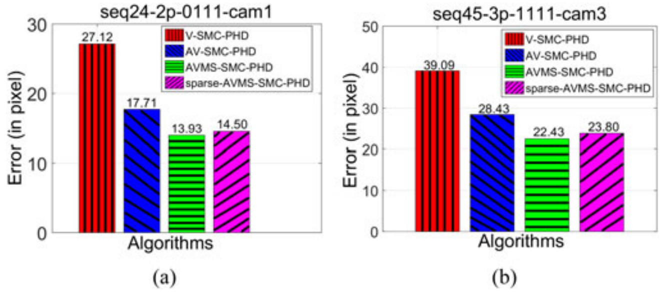

Fig. 12. Performance comparison of the V-SMC-PHD with the proposed algorithms AV-SMC-PHD, AVMS-SMC-PHD, and sparse-AVMS-SMC-PHD with bar plots.

TABLE I
EXPERIMENTAL RESULTS FOR [18], V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD, AND SPARSE-AVMS-SMC-PHD

		Tracking algorithm [18]	V SMC-PHD	AV SMC-PHD	AVMS SMC-PHD	sparse AVMS SMC-PHD
seq24	cam1	22.28	27.12	17.71	13.93	14.50
	cam2	17.60	25.91	19.83	14.97	15.35
	cam3	28.18	24.32	18.94	14.12	15.72
seq25	cam1	21.49	25.84	19.13	15.72	17.17
	cam2	19.17	25.66	18.47	13.93	15.39
	cam3	29.35	29.99	21.61	17.07	17.62
seq30	cam1	35.98	35.60	25.22	16.65	19.27
	cam2	28.40	24.97	19.37	14.86	16.16
	cam3	34.60	37.64	25.31	19.29	19.67
seq45	cam1	NA	48.68	29.46	22.95	23.40
	cam2	NA	39.24	29.47	21.47	23.16
	cam3	NA	39.09	28.43	22.43	23.80
Average		**26.34**	**32.01**	**22.75**	**17.28**	**18.43**

TABLE II
SIGNIFICANCE TEST

	V SMC-PHD	AV SMC-PHD	AVMS SMC-PHD	sparse-AVMS SMC-PHD	
V	NA	12.51	35.59	30.23	F
SMC-PHD	NA	1.9E-3	5.28E-6	1.59E-5	p-value
AV	12.51	NA	11.12	6.92	F
SMC-PHD	1.9E-3	NA	3E-3	1.53E-2	p-value

25 and 30. With the average of 26.34 achieved on these sequences, the algorithm in [18] outperforms the V-SMC-PHD filter. However, the proposed algorithms show better performance than [18].

In order to show how significant the difference is between the results of the tested algorithms in Table I, the ANOVA based F-test [60] is applied and the significance test results are given in Table II. The results of Sequence 45 are missing in [18], therefore the corresponding column could not be used in significance test. For all the significance tests, we found the same degree of freedom $(1, 22)$ and so, the corresponding F_{crit} value for $(1, 22)$ is 4.30 from the F-distribution table given in [60]. The p-value (or probability value) is the probability of a more extreme result than what we actually achieved when the null hypothesis is true. The F-value is defined as the ratio of the variance of the group means to the mean of the within group variances. The F-test was carried out at a 5% significance level. According to this test, the results are considered as statistically significant if $F > F_{\mathrm{crit}}$ and p-value is less than 0.05 (for a 5% significance level). From the test results, we can observe that the results of Table I are indeed statistically significant.

TABLE III
COMPARISON OF TRACKING RESULTS IN TERMS OF MEAN WASSERSTEIN DISTANCE (IN PIXEL)

seq24	Tracking algorithm [18]	Tracking algorithm [15]	V SMC-PHD	AV SMC-PHD	AVMS SMC-PHD	sparse AVMS SMC-PHD
cam1	9.02	7.20	16.96	7.94	6.67	7.45
cam2	6.4	4.80	19.17	7.59	5.24	5.73
Average	**7.71**	**6.00**	**18.06**	**7.76**	**5.96**	**6.59**

Our tracking results are also compared with those of Pham *et al.* [15] where the Wasserstein distance [59] was used for evaluating the tracking results of Sequence 24 cameras #1 and #2. Hence, the results of [18] and ours are also evaluated in terms of the Wasserstein distance and given in Table III. Among the six methods, the proposed AVMS-SMC-PHD outperforms the others.

D. Evaluations on the AMI and CLEAR Datasets

In order to show the performance of the proposed algorithms on other datasets than AV16.3, we selected sequences from another two multiple-subject datasets, namely, the AMI dataset [55] and the CLEAR dataset [54].

In our proposed algorithm, it is assumed that the born particles are generated and propagated only when a new speaker is detected in terms of the DOA information derived from audio. The main purpose of this assumption is to reduce the computational cost induced by propagating new born particles in each time frame. Different from the AV16.3 dataset, however, the speakers in both AMI and CLEAR datasets are talking one by one. For the visual tracker it is convenient to detect all the speakers as the born particles are propagated in each time frame, while in audio-visual tracking, it happens only if the speaker talks.

Therefore, to evaluate the tracking algorithms on these two datasets, we allow the proposed audio-visual tracker to run on the sequence from the beginning until the image frame where each of the speakers talks at least once, and the tracking errors were measured from this particular frame onwards. Another issue about these two datasets is that the calibration information of the cameras is not available, which prevents us from projecting the DOA information from 3D to 2D (i.e. image plane). To allow fair comparison, we have used noisy DOA information which was obtained by adding noise to the results from the annotation of the video frames.

As an example, we include the results for Sequence IS1001a, and Sequence UKA_20060726. Some frames of the tracking results are shown in Figs. 13 and 14 for V-SMC-PHD and AV-SMC-PHD. The average errors for V-SMC-PHD,

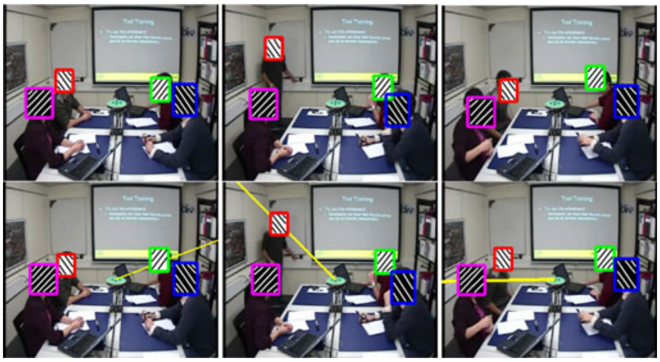

Fig. 13. Sequence IS1001a from AMI. The first and second row show the results of the V-SMC-PHD and the proposed AV-SMC-PHD filter, respectively.

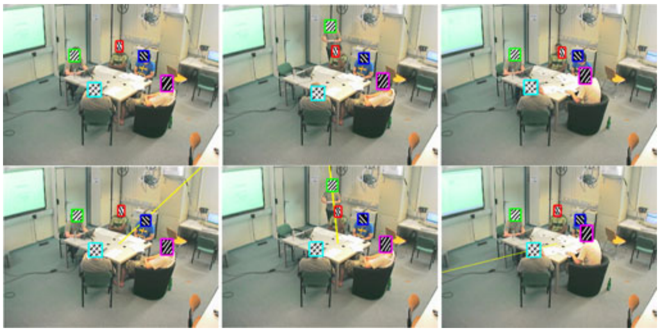

Fig. 14. Sequence UKA_20060726 from CLEAR. The first and second row show the results of the V-SMC-PHD and the proposed AV-SMC-PHD filter, respectively.

TABLE IV
AVERAGE TRACKING ERRORS (IN PIXEL) OF THE ALGORITHMS ON THE CHOSEN SEQUENCES FROM THE AMI AND CLEAR DATASETS

Methods– Sequences	V SMC-PHD	AV SMC-PHD	AVMS SMC-PHD	sparse AVMS SMC-PHD
IS1001a	25.32	21.51	18.91	20.37
UKA_20060726	28.33	25.94	23.14	24.82

AV-SMC-PHD, AVMS-SMC-PHD, and sparse-AVMS-SMC-PHD are summarised in Table IV.

As the speakers are talking one by one, performance difference between visual and audio-visual trackers is less significant. In this case, the audio-visual tracker acts similarly to a visual tracker for the silent speakers, while it is more effective for the talking speakers.

VII. CONCLUSION

In this study, we have presented several contributions for multi-speaker tracking. First, we have introduced a SMC-PHD approach for tracking a variable number of speakers in a smart room environment using audio-visual measurements. The efficient distribution of the born particles based on the DOA information reduces both the computational complexity and the estimation error. The mean-shift method is introduced to further improve the estimation accuracy of the AV-SMC-PHD filter by driving the particles to their neighbouring local peaks. We have also proposed the use of sparse sampling, to allow the mean-shift to run on a subset of the particles, thus significantly reducing the extra computational cost induced by the mean-shift with only a very small sacrifice in estimation accuracy. The proposed algorithms have been tested on the $AV16.3$ dataset for two and three-speaker scenarios, where the number of speakers varies over time. In addition, these algorithms have been tested with sequences from AMI and CLEAR datasets for four and five-speaker scenarios. Experimental results demonstrated that the proposed techniques can reliably estimate both the number of speakers and the positions of the speakers with significant improvement. The proposed tracking system could be further improved by formulating the sparse sampling process with a sparse coding framework, and extended to include other audio information or microphone array types.

ACKNOWLEDGMENT

The authors wish to thank the associate editor and the anonymous reviewers for their contributions to improving the quality of this paper. The authors confirm that all data underlying the findings are fully available without restriction. Details of the data and how to get access are available from the University of Surrey: http://dx.doi.org/10.15126/surreydata.00811616.

REFERENCES

[1] Q. Liu, Y. Rui, A. Gupta, and J. J. Cadiz, "Automating camera management for lecture room environments," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2001, pp. 442–449.

[2] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio–visual active speaker tracking in cluttered indoors environments," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 3, pp. 799–807, Jun. 2008.

[3] M. Wölfel and J. W. McDonough, "Combining multi-source far distance speech recognition strategies: Beamforming, blind channel and confusion network combination," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2005, pp. 3149–3152.

[4] G. Potamianos, C. Neti, and S. Deligne, "Joint audio-visual speech processing for recognition and enhancement," in *Proc. Int. Conf. Audio-Visual Speech Process.*, 2003, pp. 95–104.

[5] M. R. Naphade, T. Kristjansson, B. Frey, and T. S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems," in *Proc. Int. Conf. Image Process.*, 1998, pp. 536–540.

[6] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 5, pp. 882–894, Oct. 2010.

[7] A. Hampapur *et al.*, "Smart video surveillance: Exploring the concept of multiscale spatiotemporal tracking," *IEEE Signal Process. Mag.*, vol. 22, no. 2, pp. 38–51, Mar. 2005.

[8] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010.

[9] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 125–141, 2008.

[10] O. Lanz, "Approximate Bayesian multibody tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1436–1449, Sep. 2006.

[11] M. Isard and J. MacCormick, "Bramble: A Bayesian multiple-blob tracker," in *Proc. Int. Conf. Comput. Vis.*, 2001, vol. 2, pp. 34–41.

[12] C.-T. Chu, J.-N. Hwang, H.-I Pai, and K.-M. Lan, "Tracking human under occlusion based on adaptive multiple kernels with projected gradients," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1602–1615, Nov. 2013.

[13] X. Suau, J. Ruiz-Hidalgo, and J. R. Casas, "Real-time head and hand tracking based on 2.5D data," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 575–585, Jun. 2012.

[14] K.-W. Chen, C.-C. Lai, P.-J. Lee, C.-S. Chen, and Y.-P. Hung, "Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 625–638, Aug. 2011.

[15] N. T. Pham, W. Huang, and S. H. Ong, "Tracking multiple objects using probability hypothesis density filter and color measurements," in *Proc. Int. Conf. Multimedia Expo*, 2007, pp. 1511–1514.

[16] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1409–1415, May 2012.

[17] V. Cevher, A. C. Sankaranarayanan, J. H. McClellan, and R. Chellappa, "Target tracking using a joint acoustic video system," *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 715–727, Jun. 2007.

[18] M. Barnard *et al.*, "Robust multi-speaker tracking via dictionary learning and identity modelling," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 864–880, Apr. 2014.

[19] C. Shan, Y. Wei, T. Tan, and F. Ojardias, "Real time hand tracking by combining particle filtering and mean shift," in *Proc. Int. Conf. Autom. Face Gesture Recog.*, 2004, pp. 669–674.

[20] S. K. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1491–1506, Nov. 2004.

[21] J. Sullivan and J. Rittscher, "Guiding random particles by deterministic search," in *Proc. Int. Conf. Comput. Vis.*, 2001, vol. 1, pp. 323–330.

[22] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[23] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–577, May 2003.

[24] E. Polat and M. Ozden, "A nonparametric adaptive tracking algorithm based on multiple feature distributions," *IEEE Trans. Multimedia*, vol. 8, no. 6, pp. 1156–1163, Dec. 2006.

[25] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," in *Proc. Int. Conf. Image Process.*, Sep. 2003, vol. 3, pp. 25–28.

[26] P. Cui, L.-F. Sun, F. Wang, and S.-Q. Yang, "Contextual mixture tracking," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 333–341, Feb. 2009.

[27] Y. Bar-Shalom and T. Fortmann, *Tracking and Data Association*. San Francisco, CA, USA: Academic, 1998.

[28] M. Isard and A. Blake, "Condensation—Conditional density propagation for visual tracking," *Int. J. Comput. Vis.*, vol. 29, pp. 5–28, 1998.

[29] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1154–1164, 2002.

[30] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proc. Int. Conf. Multimodal Interfaces*, 2005, pp. 61–68.

[31] B. -N. Vo, S. Singh, and W. K. Ma, "Tracking multiple speakers using random sets," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 2004, vol. 2, pp. 357–360.

[32] R. P. S. Mahler, *Statistical Multisource-Multitarget Information Fusion*. Norwood, MA, USA: Artech House, 2007.

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

[34] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Audio constrained particle filter based visual tracking," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 3627–3631.

[35] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Adaptive particle filtering approach to audio-visual tracking," in *Proc. Eur. Signal Process. Conf.*, Sep. 2013, pp. 1–5.

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

[37] K. Deguchi, O. Kawanaka, and T. Okatani, "Object tracking by the mean-shift of regional color distribution combined with the particle-filter algorithms," in *Proc. Int. Conf. Pattern Recog.*, 2004, vol. 3, pp. 506–509.

[38] E. Maggio and A. Cavallaro, "Hybrid particle filter and mean shift tracker with adaptive transition model," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, vol. 2, pp. 221–224.

[39] J. Wang and Y. Yagi, "Adaptive mean-shift tracking with auxiliary particles," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 6, pp. 1578–1589, Dec. 2009.

[40] C. Chang and R. Ansari, "Kernel particle filter for visual tracking," *IEEE Signal Process. Lett.*, vol. 12, no. 3, pp. 242–245, Mar. 2005.

[41] S. Zhong and F. Hao, "Hand tracking by particle filtering with elite particles mean shift," in *Proc. Jpn.-China Joint Workshop Frontier Comput. Sci. Technol.*, Dec. 2008, pp. 163–167.

[42] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2259–2272, Nov. 2011.

[43] B. Ma *et al.*, "Visual tracking using strong classifier and structural local sparse descriptors," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1818–1828, Oct. 2015.

[44] V. Kilic, M. Barnard, W. Wang, A. Hilton, and J. Kittler, "Audio informed visual speaker tracking with SMC-PHD filter," in *Proc. Int. Conf. Multimedia Expo*, Jul. 2015, pp. 715–720.

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

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

[47] B.-N. Vo and W.-K. Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4091–4104, Nov. 2006.

[48] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, vol. 3, pp. 265–268.

[49] J. Czyz, B. Ristic, and B. Macq, "A color-based particle filter for joint detection and tracking of multiple objects," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, pp. 217–220.

[50] C. Shan, T. Tan, and Y. Wei, "Real-time hand tracking using a mean shift embedded particle filter," *Pattern Recog.*, vol. 40, no. 7, pp. 1958–1970, 2007.

[51] V. Kilic, "Audio-visual tracking of multiple moving speakers," Ph.D. dissertation, Dept. Electr. Electro. Eng., Univ. Surrey, Guildford, U.K., 2015.

[52] D. Fox, "Adapting the sample size in particle filters through KLD-Sampling," *Int. J. Robot. Res.*, vol. 22, pp. 985–1003, 2003.

[53] G. Lathoud, J. M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Proc. Mach. Learn. Med. Imag. Workshop*, 2005, pp. 182–195.

[54] R. Stiefelhagen *et al.*, "The CLEAR 2007 evaluation," in *Proc. Int. Eval. Workshops Multimodal Technol. Perception Humans*, 2008, pp. 3–34.

[55] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *Proc. 2nd Int. Conf. Mach. Learn. Multimodal Interaction*, 2006, pp. 28–39.

[56] M. Taj, "Surveillance performance evaluation initiative (spevi) audiovisual people dataset," 2007, accessed on: Aug. 24, 2014. [Online]. Available: http://www.eecs.qmul.ac.uk/˜andrea/spevi.html.

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

[58] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3447–3457, Aug. 2008.

[59] J. R. Hoffman and R. Mahler, "Multitarget miss distance via optimal assignment," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 34, no. 3, pp. 327–336, May 2004.

[60] A. G. Bluman, *Elementary Statistics*. New York, NY, USA: McGraw-Hill, 2013.

Volkan Kılıç (S'04) received the B.Sc. degree in electrical and electronics engineering from Anadolu University, Eskisehir, Turkey, in 2008, and the M.Sc. degree in electronics engineering from the Institute of Science and Technology, Istanbul Technical University, Istanbul, Turkey, in 2010, and the Ph.D. degree from the Department of Electrical and Electronic Engineering, University of Surrey, Guildford, U.K., in 2016.

He is currently an Assistant Professor with Izmir Katip Celebi University, Izmir, Turkey. He joined the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K., in 2012. His current research interests include audio-visual signal processing, multimodal speaker tracking, particle, and PHD filters.

Mark Barnard received the Ph.D. degree from EPFL, Lausanne, Switzerland, in 2005.

While working toward the Ph.D. degree, he worked with the IDIAP Research Institute, Martigny, Switzerland, as a Research Assistant. In 2006, he joined the Machine Vision Group, University of Oulu, Oulu, Finland, where he was Postdoctoral Researcher for three years. He is currently a Research Fellow with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford, U.K. His current research interests include audio-visual tracking, dictionary-based image representation, and audio head pose estimation.

Wenwu Wang (M'02–SM'11) received the B.Sc., M.E., and Ph.D. degrees from the Harbin Engineering University, Harbin, China, in 1997, 2000, and 2002, respectively.

Since May 2007, he has been with the Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, U.K., where he is currently a Reader and a Codirector of the Machine Audition Lab. He has authored or coauthored more than 150 publications. His current research interests include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, machine audition (listening), and statistical anomaly detection.

Dr. Wang is an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Adrian Hilton (M'98) received the B.Sc. (Hons.) and D.Phil. degrees from the University of Surrey, Guildford, U.K., in 1988 and 1992, respectively.

He is currently a Professor of computer vision and the Director of the Centre for Vision, Speech and Signal Processing, University of Surrey, Guilford, U.K. His contributions include technologies for the first hand-held 3D scanner, modeling of people from images, and 3D video for games, broadcast, and film. He currently leads research investigating the use of computer vision for applications in entertainment content production, visual interaction, and clinical analysis. His current research interests include robust computer vision to model and understand real-world scenes.

Prof. Hilton is a Chartered Engineer.

Josef Kittler (M'74–LM'12) received the B.A., Ph.D., and Sc.D. degrees from the University of Cambridge in 1971, 1974, and 1992, respectively.

He is a Professor of machine intelligence with the Centre for Vision, Speech and Signal Processing, Department of Electrical and Electronic Engineering, University of Surrey, Guildford, U.K. He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He authored the textbook *Pattern Recognition: A Statistical Approach* (Prentice-Hall, 1982) and more than 170 journal papers.

Prof. Kittler serves on the Editorial Board of several journals in pattern recognition and computer vision.