Audio Assisted Robust Visual Tracking With Adaptive Particle Filtering

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

***Abstract*—The problem of tracking multiple moving speakers in indoor environments has received much attention. Earlier techniques were based purely on a single modality, e.g., vision. Recently, the fusion of multi-modal information has been shown to be instrumental in improving tracking performance, as well as robustness in the case of challenging situations like occlusions (by the limited field of view of cameras or by other speakers). However, data fusion algorithms often suffer from noise corrupting the sensor measurements which cause non-negligible detection errors. Here, a novel approach to combining audio and visual data is proposed. We employ the direction of arrival angles of the audio sources to reshape the typical Gaussian noise distribution of particles in the propagation step and to weight the observation model in the measurement step. This approach is further improved by solving a typical problem associated with the PF, whose efficiency and accuracy usually depend on the number of particles and noise variance used in state estimation and particle propagation. Both parameters are specified beforehand and kept fixed in the regular PF implementation which makes the tracker unstable in practice. To address these problems, we design an algorithm which adapts both the number of particles and noise variance based on tracking error and the area occupied by the particles in the image. Experiments on the $AV16.3$ dataset show the advantage of our proposed methods over the baseline PF method and an existing adaptive PF algorithm for tracking occluded speakers with a significantly reduced number of particles.**

***Index Terms*—Adaptive particle filter, audio-visual speaker tracking, particle filter.**

I. Introduction

SPEAKER tracking in smart environments has attracted an increasing amount of attention in the last decade, driven by applications such as automatic camera steering in video conferencing and individual speaker discrimination in multispeaker environments. Earlier techniques were designed to track one person in a static and controlled environment. However, theoretical and algorithmic advances together with the increasing capability in computer processing have led to the emergence of more sophisticated techniques for tracking in dynamic and less controlled (or natural) environments with multiple speakers [1]–[3]. The type of sensors used to collect the measurements is also evolving from single- to multi-modality.

Early efforts in speaker tracking often use either visual only or audio only data despite the fact that both audio and visual information are readily available in many real world scenarios. The method of video-only tracking [4], [5] is generally reliable and accurate when the targets are in the camera field of view, but limitations are introduced 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 is changed [3], [6]. Audio tracking [7]–[9] is not restricted by these limitations, however, audio data is intermittent over time and may be corrupted by background noise and room reverberations, which may introduce non-negligible tracking errors. In addition, spatial resolution (tracking resolution in the world space) of audio is in general worse than that of video. Using both audio and visual data has the potential to improve the tracking performance in the case that either modality is unavailable or both are corrupted.

A popular approach for tracking speakers with audio and video data is to use a state-space approach based on the Bayesian framework, for example, the Kalman filter (KF) for linear motion and sensor models [10], extensions of KF for the nonlinear models using the first order Taylor expansion including the decentralized Kalman filter (DKF) [11], [12] and extended Kalman filter (EKF) [13], [14], and the particle filter (PF) for nonlinear and non-Gaussian models [15]. In comparison to the KF and EKF approaches, the PF approach is more robust for nonlinear models as it can approach the Bayesian optimal estimate with a sufficiently large number of particles [15]. It has been widely employed for speaker tracking problems [16]–[18]. For example, in [16] and [17], PF is used to fuse object shapes and audio information. In [18], independent audio and video observation models are fused for simultaneous tracking and detection of multiple speakers. One challenge in using PF, however, is to choose an appropriate number of particles. An insufficient number may lead to particle impoverishment while a larger number (than required) will introduce extra computational burden. Choosing the optimal number of particles is one of the issues that affect the performance of the tracker, and none of the above works have addressed this problem as we do here.

Besides the Bayesian method, another approach for tracking is based on finite-set statistics (FISST) theory called the probability hypothesis density (PHD) filter [19]. The PHD filter is a first-moment filter which propagates the first order moment of a dynamic point process. Some applications of the PHD filter with speaker tracking are given in [20] and [21]. The main advantage of the PHD filter over Bayesian (Kalman or PF) approach is that it does not require any *a priori* knowledge

Manuscript received March 05, 2014; revised June 20, 2014 and September 12, 2014; accepted November 18, 2014. Date of publication December 04, 2014; date of current version January 15, 2015. This work was supported by the Engineering and Physical Sciences Research Council of the U.K. under Grant EP/H050000/1, 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. Lexing Xie.

The authors are with the Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: v.kilic@surrey.ac.uk; mark.barnard@surrey.ac.uk; w.wang@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.2014.2377515

of the number of targets, which is actually estimated during the tracking process. However, the PHD filter confines the propagation of the full multi-target posterior to the first order multi-target moment which corresponds to a loss of higher order cardinality information that results in erratic estimates of the number of objects in low signal-to-noise ratio (SNR) scenarios [22]. Propagating the whole multi-target posterior is computationally intractable. The cardinalized PHD (CPHD) filter additionally propagates the cardinality distribution to PHD and leads to better performance over the PHD for the estimation of instantaneous target number [22], [23] and the position of the speakers [24]. The cardinality distribution, nevertheless, makes the CPHD more computationally demanding than the PHD filter. Also, the CPHD does not provide explicit models for the spawning of new targets by prior targets.

Apart from the tracking methods mentioned above, multimodal usage in speaker tracking brings a problem of associating each measurement with an appropriate target which is known as data association. Data association methods can be classified into two main categories [25]. The first one is unique-neighbor data association such as multiple hypothesis tracking (MHT) which associates each measurement to one of the existing tracks. The second one is all-neighbors data association, such as joint probabilistic data association (JPDA) which uses all the measurements for updating the entire track estimate. MHT filter has an advantage in maintaining multiple hypotheses of the association between a target state and the measurements in the measurement set. The drawback of MHT is that the number of hypotheses grows exponentially over time [26]. JPDA approximates the posterior target distribution as separate Gaussian distributions for each target [27], [28] which results in an increased computational cost. Data association algorithms with Bayesian methods and PHD filter in target tracking applications can be found in [7], [29]–[32]. However, some researchers found that classical data association algorithms are computationally expensive, and this led them to fuse multi-modal measurements inside their proposed framework [11], [14], [16], [17], [20] as we also do here.

Among the approaches presented above, the PF framework has been chosen for tracking multiple speakers in this study. Compared to other sequential Bayesian estimation techniques, the advantage of PF lies in their flexibility with respect to the types and numbers of features they support, their robustness in the presence of noise, and the nonparametric fashion in which they represent the belief about the target state, which makes them applicable for highly nonlinear, non-Gaussian estimation problems. Here, we focus on two challenging problems associated with PF based visual tracking.

The first problem stems from the limitations of using the single modality of vision which affects the accuracy and reliability of tracker because of the limited field of view and occlusion. To address this problem, audio data is used as a second modality to improve the performance of visual tracker. Researchers have presented fusion strategies for integrating audio localization information with video tracking, [33]–[35], [2]. These strategies are performed by modifying the observation model [33], using the likelihood function composition of different sensor information [34], state association of two modalities [35], or a graphical model for characterising mutual dependencies of the two modalities [2]. These methods, however, are sensitive to the outliers in audio data, and noisy audio data can easily cause deviation in the estimation of the target position. Unlike these methods, in this paper, we propose integrating audio and visual data in the steps of the PF framework, by weighting the contribution of the audio in order to minimize the negative effect of outliers and noise coming from the audio data, rather than performing any *a priori* data fusion algorithm. One benefit of this approach is that running a data fusion algorithm is not required which would introduce extra computational cost. To the best of our knowledge, audio information has not been previously fused with visual information in a PF as we do here.

The second problem originates from the PF itself. It uses a weighted set of samples (particles) in order to approximate the filtering distributions and hence the quality of the sample based representation increases with the number of particles. It is, however, not clear how to determine the optimal number of particles to be used for a specific estimation problem. As a rule of thumb, the number of particles is chosen to be as large as possible to get accurate results which leads to an increased computational cost. A detailed analysis of this trade-off is performed by Pitt *et al.* [36] who provided practical guidelines for the estimation of the optimal number of particles in Markov chain Monte Carlo particle filter with the Metropolis Hastings sampler. It is assumed that the standard deviation of the estimated log-likelihood from the PF is around 1 and inversely proportional to the number of particles. Their results are valid for the Metropolis Hastings sampler, but for other samplers, it is not clear whether the standard deviation of the likelihood from the PF plays the same role in the estimation of the optimal number of particles. Another potential approach for this problem is variable resolution particle filter (VRPF) [37] which introduces the concept of "abstract particles" that a particle may represent an individual state or a set of similar states. The VRPF has the advantage that a limited number of particles are sufficient to represent the large portions of the state space since a single abstract particle simultaneously tracks multiple similar states. However, this method cannot answer the question of how to determine the optimal number of particles. Subsequent researchers have therefore proposed adaptive particle filtering (A-PF) approaches in [38]–[40]. The Kullback-Leibler divergence (KLD) sampling algorithm was proposed by Fox [38]. The idea behind this algorithm is to adaptively estimate the number of particles at each step to bound the approximation error introduced by the sample based representation of the PF below a specified threshold. One assumption of KLD-sampling is that a sample based representation of the PF can be used to estimate the posterior by a discrete piecewise constant distribution consisting of a set of multidimensional bins. Subsequent work [39] modified the KLD-sampling criterion to estimate the number of particles and proposed an approach for adaptive propagation of the samples. Recent work [40] uses the innovation error to modify the number of particles being used where a two-fold metric is employed to select the number of particles. The first metric is used to eliminate the particles whose distance to a neighboring particle is below a predefined threshold, and the second is a basis for setting the threshold on the innovation error to control the birth of particles. These two thresholds should be set prior to running the algorithm, but it is not mentioned how, and also the evaluation of the algorithm is limited to only a simple computer simulation which could not give an insight into the strength and weakness of the framework.

TABLE I
COMMONLY USED NOTATIONS

$\mathbf{x}$	bold lower case denotes vectors
N	the number of particles
$n = 1, ..., N$	the particle index
K	the total number of image frames
$k = 1, ..., K$	the image frame index
$\tilde{\mathbf{x}}_k$	estimated target position at the frame k
$\hat{\mathbf{x}}_k^{(n)}$	the position of the n-th particle at the frame k after incorporating DOA
$\mathbf{F}$	bold capital denotes matrices
$\|.\|_1$	ℓ_1 norm
$\mathcal{V}$	tensor
$\odot$	the element-wise product
$\otimes$	the outer product
$\oplus$	the element-wise addition

As we describe in the rest of the paper, our work differs substantially from previous works on AV multiple speaker tracking with respect to audio integration into the PF framework, and adaptive estimation of the particle number and variance of Gaussian noise. Direction of arrival (DOA) angles of the audio sources are used to relocate the particles in the propagation step and recalculate the weights of the particles in the measurement step. Audio is fused to the visual particle filter (V-PF) through modifications of the PF steps. That makes the tracker less sensitive to outliers and noise in audio data. This method is then further improved by proposing an adaptive approach to PF based on the occupied area by the particles in each frame. Our adaptive approach allows us to estimate dynamically not only the number of particles but also the noise variance which makes it different from the adaptive approaches mentioned above with the advantage that adaptive noise variance is used in the estimation of the optimal number of particles. Finally, we demonstrate the results using simulations to compare the performance of the proposed algorithms with [38] and V-PF.

The rest of this paper is organized as follows: the next section introduces the PF used in visual tracking. Section III presents our proposed audio-visual particle filter (AV-PF) algorithm. Section IV describes our proposed audio-visual adaptive particle filter (AV-A-PF) algorithm. Section V shows experimental results performed on the $AV16.3$ dataset and compares the performance of the algorithms. Closing remarks are given in Section VI.

For readability, commonly used notations in the paper are defined in Table I.

II. PARTICLE FILTERING-BASED VISUAL TRACKING

The PF is an approach for obtaining estimates of the state of a stochastic dynamical system based on observations recursively in time. It is also known as sequential Monte Carlo methods (SMC) based on simulation. It was first introduced by Gordon *et al.* [41]. The PF, which is based on sequential importance sampling and Bayesian theory, is a powerful approach for nonlinear and non-Gaussian problems.

The sampling importance resampling (SIR) is a generalization of the PF framework which can be used in visual tracking to track the position of the speaker face (x_1, x_2) in five steps. The particles are initialized as $\mathbf{x}_0^{(n)} \sim p(\mathbf{x}_0)$, $w_0^{(n)} = \frac{1}{N}$ for $n = 1, \ldots, N$ in the first step of V-PF. Here N is the number of particles and $w_0^{(n)}$ are the initial weights of the particles. The state vector is defined as $\mathbf{x} = [\, x_1 \quad \dot{x}_1 \quad x_2 \quad \dot{x}_2 \quad s \,]^T$, where x_1 and x_2 are the horizontal and vertical positions of the rectangle centred around the face that we wish to track, $\dot{x}_1$ is the horizontal velocity, $\dot{x}_2$ is the vertical velocity and s is the scale of the rectangle centred around (x_1, x_2). In the second step, particle propagation is employed by a dynamic model

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

where $\mathbf{x}_k^{(n)}$ is the state of the nth particle at time frame $k = 1, \ldots, K$ and $\mathbf{q}_k^{(n)}$ is the zero-mean Gaussian noise with covariance $\mathbf{Q}$, $\mathbf{q}_k^{(n)} \sim \mathcal{N}(0, \mathbf{Q})$ for each particle and $\mathbf{F}$ is the linear motion model

$$\mathbf{F} = \begin{bmatrix} 1 & T & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & T & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_s^2 \end{bmatrix}$$

where T is the period between two adjacent frames, σ_s^2 is the variance of the scale and σ^2 is the variance for both the position and the velocity. The third step is the weighting step and the particles are weighted by the observation model

$$w_k^{(n)} = e^{-\lambda (D^{(n)})^2} \tag{2}$$

where λ is the design parameter and $D^{(n)}$ is the Bhattacharyya distance

$$D^{(n)} = \sqrt{1 - \sum_{u=1}^{U} \sqrt{r(u) q^{(n)}(u)}} \tag{3}$$

where U is the number of histogram bins, $r(u)$ is the Hue histogram of the reference image determined by the user in the initialization step, and $q^{(n)}(u)$ is the Hue histogram extracted from the rectangle centred on the position of the nth particle. The RGB or HSV colour model is commonly used in the literature [42]. In our study, HSV is chosen since it is observed to be more robust to illumination variation. Before the fourth step, normalization is applied to ensure that $\sum_{n=1}^{N} w_k^{(n)} = 1$. In the fourth step, the position of the speaker is estimated by

$$\tilde{\mathbf{x}}_k = \sum_{n=1}^{N} w_k^{(n)} \mathbf{x}_k^{(n)} \tag{4}$$

As a last step, the particles $\mathbf{x}_k^{(n)}$ are resampled to remove the particles with very small weights and duplicate particles with large weights, so a new particle set drawn from $\{\mathbf{x}_k^{(n)}, w_k^{(n)}\}_{n=1}^{N}$ is generated. Then it returns to the second step and continues recursively. The pseudo code of V-PF algorithm is given in Algorithm 1.

Fig. 1. V-PF fails after occlusion. The first row shows Sequence 11 camera #1 of the $AV16.3$ dataset where a single speaker disappears for a while and reenters to the scene. The second row shows the propagation of the particles to detect the speaker.

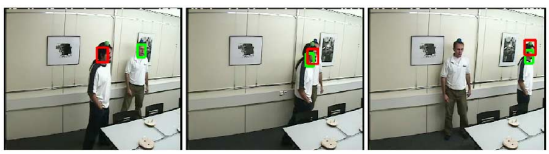

Fig. 2. Sequence 24 camera #1 of the $AV16.3$ dataset shows multiple speakers occluding each other and the visual tracker fails after occlusion.

Algorithm 1: Visual particle filter (V-PF) tracking algorithm.

Initialize: N, σ^2, U, T, $\mathbf{F}$, λ, $r(u)$, $\mathbf{x}_0^{(n)}$, $w_0^{(n)}$, k
while $k < K$ **do**
  Propagate particles: $\mathbf{x}_k^{(n)} = \mathbf{F}\mathbf{x}_{k-1}^{(n)} + \mathbf{q}_k^{(n)}$
  Calculate $D^{(n)}$ using equation (3), for $n = 1 \dots N$
  Weighting: $w_k^{(n)} = e^{-\lambda (D^{(n)})^2}$, for $n = 1 \dots N$
  Estimate target position $\tilde{\mathbf{x}}_k = \sum_{n=1}^{N} w_k^{(n)} \mathbf{x}_k^{(n)}$
  Resampling: Generate $\mathbf{x}_k^{(n)}$ from the set $\{\mathbf{x}_k^{(n)}, w_k^{(n)}\}_{n=1}^{N}$
  $k = k + 1$
end

Although the V-PF algorithm works well in regular conditions, it fails in challenging situations like occlusion. This case is depicted in Fig. 1 and Fig. 2 using sequences recorded by calibrated cameras in $AV16.3$ dataset described in Section V-A. Fig. 1 shows an occlusion case where speaker re-appears in the scene after going out for a while, and another occlusion case is shown in Fig. 2 where two speakers occlude each other. The visual tracker has no visual cues during the occlusion which causes losing the speaker. Even when the speaker becomes visible again after the occlusion, the tracker is unable to detect the speaker as it is depicted in the first row of Fig. 1. In the second row of Fig. 1, the particles of the tracker, shown as red spots, are propagated to detect the face of the speaker. Once the tracker loses the speaker, the particles focus on objects similar to the speaker, causing divergence from the speaker. To address this problem, several methods could be used, such as occlusions map [3] and BraMBLe tracker [6]. Here we present an alternative method by introducing audio information, as discussed next.

III. Proposed Particle Filter-based Audio Constraint Visual Tracking Algorithm

In this section, we present a new method to enhance the visual tracker described above by introducing audio information.

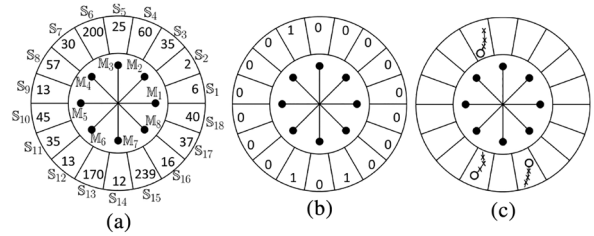

Fig. 3. Sector-based activeness measure is depicted in (a) , sector-based detection in (b), and points-based localization in (c).

Despite the fact that a variety of audio information could be used such as sound source localization (SSL) and time delay estimation (TDE), as a proof of concept, the DOA angle is used here which is more feasible in audio processing from circular microphone array in an indoor environment, employed for collecting the dataset used in our experiments. In the literature, many methods are proposed such as the coherent signal subspace (CSS) [43] and the MUSIC algorithm [44]. Our proposed AV tracking algorithm is designed to handle the noise within the DOA estimates. Therefore, the choice of DOA estimation algorithms is not crucial. In this study, we used sam-spare-mean (SSM) method [45] to estimate the DOA information which is incorporated to improve the tracking performance and robustness of the visual tracker. The SSM method is a two-step method. The first step consists of a sector-based combined detection and localization. The space around a circular microphone array of eight microphones $\{\mathbb{M}_1 \cdots \mathbb{M}_8\}$ is divided into 18 sectors $\{\mathbb{S}_1 \cdots \mathbb{S}_{18}\}$ in Fig. 3(a) and for each sector an "activeness" measure is evaluated at each time frame [46]. Then, in Fig. 3(b) this measure of activeness is compared to a threshold in order to give a binary decision of whether there is an active source in that sector. The second step is a point-based search conducted in each of the sectors labelled as having at least one active source in Fig. 3(c). The parametric approach [46] is then used for localization, and the location parameters are optimized with respect to a cost function such as SRP-PHAT [47]. Due to space constraint, more details on the derivation of the DOA angle are omitted here and can be found in [48].

The DOA estimates given by the SSM method can be noisy for reverberant audio measurements. To mitigate the noise effect, we apply a third order AR model [49] to improve the estimate of the azimuth.

$$\theta_k = \sum_{i=1}^{3} \varphi_i \theta_{k-i} + \varepsilon_k \tag{5}$$

where θ_k is the DOA (azimuth) angle (in degrees) of the speaker estimated from the audio frame that is synchronized with image frame k, φ_i are the parameters of the model and ε_k is white noise. Note that, to estimate the DOA angles, it is not necessary for the microphone array to appear in the field of view the cameras, as the DOA is estimated from the acoustic recordings acquired by the microphone arrays which have a listening range of 360 degrees no matter whether the cameras are presented in the room.

The V-PF approach described in Section II can now be enhanced by the DOAs information discussed above. Here, we assume that the calibration information of the microphone array, such as its position, is available. If the calibration information is not available, the positions of the microphone arrays could be estimated via microphone self-calibration [50] or combined

microphone and camera calibrations [51], which is however beyond the scope of this study. The idea behind our approach is to relocate the distributed particles around the DOA line and then re-calculate the weights of the relocated particles according to their distance to the DOA line [52]. The DOA line can be drawn as follows. First, the 3-D position of the speaker's head (A, B_k, C) is determined based on the estimated DOA angle and the following assumptions: (1) A is the distance from the centre of the microphone array to the wall in metres (which is 1.75 metres in our experiments), (2) C is the estimated height of the speaker, typically chosen as 1.80 metres in our experiments. Then B_k is calculated using the standard trigonometric identity as

$$B_k = \tan\left(\theta_k \times \frac{\pi}{180}\right) \cdot A \tag{6}$$

The 3-D coordinate (A, B_k, C) is then projected to the image frame to obtain the 2-D coordinate (a_k, b_k) using the calibration matrix, formed from the calibration information of the microphone arrays and cameras available in the dataset, e.g. the 3-D coordinates of the center of the two microphone arrays (0, 0, 0), and the three cameras positions, $(-1.56, 2.02, 1.40)$, $(1.52, -2.25, 1.13)$ and $(-0.25, -3.03, 1.26)$ (unit in meters), for camera #1, #2, and #3, respectively. The DOA line is drawn from (a_k, b_k) to the 2-D coordinate of the centre of the microphone array which is estimated only once using the same calibration matrix at the initialization step, since all the cameras are stationary and the positions of the microphone arrays are always constant for all the camera views.

When the particles are propagated, we want to concentrate on particles located around the DOA line. Concentrating around DOA line is likely to increase the possibility of speaker detection by the particles since the DOA indicates the approximate direction of the sound emanating from the speaker. If the location of particles is assumed to be initially distributed in a circular area, then after relocation, it is expected to be elliptical instead of being exactly on the DOA line in order to avoid deviation in the detection in the case of noisy DOAs measurements. To get elliptical distribution, the moving distance of the particles should be proportional to their initial distances to the DOA line which allows the farthest particle to move more than the closest particle thus maintaining the relative distance to the DOA line. To this end, perpendicular Euclidean distances $\mathbf{d}_k = [\, d_k^{(1)} \quad \dots \quad d_k^{(N)} \,]^T$ of the particles to the DOA line are first calculated. These distances are then normalized to obtain distance coefficients to be used to derive the movement distances $\hat{\mathbf{d}}_k$ as follows:

$$\hat{\mathbf{d}}_k = \frac{\mathbf{d}_k}{\|\mathbf{d}_k\|_1} \odot \mathbf{d}_k \tag{7}$$

where $\hat{\mathbf{d}}_k = [\, \hat{d}_k^{(1)} \quad \dots \quad \hat{d}_k^{(N)} \,]^T$ and $\odot$ is the element-wise product and $\|.\|_1$ is the ℓ_1 norm. Then $\hat{\mathbf{d}}_k$ is used to guide how much the particles should be moved towards the DOA line. This information is then used to relocate the particle distribution during the propagation step in (8).

The noise within the audio measurements can affect the reliability and accuracy of the DOAs. To deal with these effects, the impact of audio to the calculation of particle propagation and importance weighting is controlled by γ_k, which is calculated as the Bhattacharyya distance to measure the similarity between $q(u)$, i.e the image patch centred on the estimated position, and the reference image patch $r(u)$, by substituting $q(u)$ for $q^{(n)}(u)$ in (3). The dynamic model given in (1) is then revised to

$$\hat{\mathbf{x}}_k^{(n)} = \mathbf{x}_k^{(n)} \oplus \hat{d}_k^{(n)} \mathbf{h}_k \gamma_k \tag{8}$$

where $\oplus$ is the element-wise addition and $\mathbf{h}_k = [\, cos(\theta_k) \quad 0 \quad sin(\theta_k) \quad 0 \quad 0 \,]^T$. The movement distance of each particle $\hat{d}_k^{(n)}$ is weighted by γ_k and this is multiplied by $\mathbf{h}_k$ to update only position (x_1, x_2) of the particle state vector $[\, x_1 \quad \dot{x}_1 \quad x_2 \quad \dot{x}_2 \quad s \,]^T$ in order to provide the perpendicular movement to the DOA line. Since the positions of the particles are changed, the importance weights are also revised by multiplying them with the inverse of the distance coefficients calculated in the previous step to make sure that the particles that are close to the DOA line in terms of the Euclidean distance still have high importance weights

$$\hat{w}_k^{(n)} = (e^{-\lambda (D^{(n)})^2}) \frac{\|\mathbf{d}_k\|_1}{d_k^{(n)}} \tag{9}$$

The weights are then normalized to ensure that $\sum_{n=1}^{N} \hat{w}_k^{(n)} = 1$. The fourth and fifth steps of the PF algorithm are performed in the same way as in Algorithm 1. Position estimation follows the weighting step and it is calculated using (4) and denoted as $\tilde{\mathbf{x}}_k^{av}$. Before the resampling step, to prevent the tracker to be deceived by noise in audio, γ_k is calculated again with $\tilde{\mathbf{x}}_k^{av}$ and denoted as γ_k^{av}. If γ_k^{av} is smaller than γ_k, the AV tracker results are used in the next step and iteration. Otherwise, audio is assumed to be noisy and the visual-only tracker results are used in the next step and iteration. Then the resampling step is performed to generate the new particles $\mathbf{x}_k^{(n)}$ from the set $\{\mathbf{x}_k^{(n)}, w_k^{(n)}\}_{n=1}^{N}$. The pseudo code of the proposed AV-PF algorithm is depicted in Algorithm 2.

Algorithm 2: Proposed AV-PF algorithm.

Initialize: N, σ^2, U, T, $\mathbf{F}$, λ, $r(u)$, $\mathbf{x}_0^{(n)}$, $w_0^{(n)}$, k
while $k < K$ **do**
  Propagate particles: $\mathbf{x}_k^{(n)} = \mathbf{F}\mathbf{x}_{k-1}^{(n)} + \mathbf{q}_k^{(n)}$
  Calculate $D^{(n)}$ using equation (3), for $n = 1 \dots N$
  Calculate weights: $w_k^{(n)} = e^{-\lambda (D^{(n)})^2}$, for $n = 1 \dots N$
  Estimate the target position $\tilde{\mathbf{x}}_k$ using equation (4)
  Calculate γ_k using equation (3)
  Get corresponding DOA angle θ_k
  Calculate distances $\mathbf{d}_k = [\, d_k^{(1)} \quad \dots \quad d_k^{(N)} \,]^T$
  Find movement distances: $\hat{\mathbf{d}}_k = \frac{\mathbf{d}_k \odot \mathbf{d}_k}{\|\mathbf{d}_k\|_1}$
  Re-propagate particles: $\hat{\mathbf{x}}_k^{(n)} = \mathbf{x}_k^{(n)} \oplus \hat{d}_k^{(n)} \mathbf{h}_k \gamma_k$
  Re-weighting: $\hat{w}_k^{(n)} = (e^{-\lambda (D^{(n)})^2}) \frac{\|\mathbf{d}_k\|_1}{d_k^{(n)}}$
  Re-estimate target position $\tilde{\mathbf{x}}_k^{av}$ using equation (4)
  Calculate γ_k^{av} using equation (3)
  if $\gamma_k^{av} < \gamma_k$ **then**
    $\mathbf{x}_k^{(n)} = \hat{\mathbf{x}}_k^{(n)}$, $w_k^{(n)} = \hat{w}_k^{(n)}$, $\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_k^{av}$
  end
  Resampling: Generate $\mathbf{x}_k^{(n)}$ from the set $\{\mathbf{x}_k^{(n)}, w_k^{(n)}\}_{n=1}^{N}$
  $k = k + 1$
end

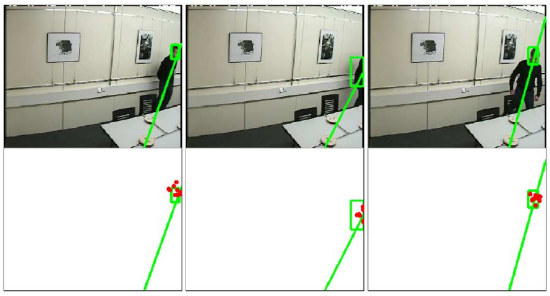

Fig. 4. The first row shows Sequence 11 camera #1 of the $AV16.3$ dataset where the single speaker disappears for a while and reenters the scene. After occlusion, the AV tracker continues tracking. The second row shows the distribution of the particles which are relocated by the DOA line.

With our proposed modifications in (8) and (9), the tracking algorithm can preserve the position of the face even if the visual tracker is lost, due to the use of the DOAs as depicted in Fig. 4. Contrary to the visual tracker in Fig. 1, the AV tracker continues tracking after the speaker comes back to the camera view in the first row of Fig. 4. The second row shows how the particles are distributed around the DOA line. Concentrating particles around the DOA line increases the efficiency of the particles in terms of speaker detection since all particles converge to the potential location of the speaker. This allows us to use a smaller number of particles than required in visual-only PF.

In the $AV16.3$ dataset that we used in our experiments, the speakers are talking continuously in most of the time in the video sequence which therefore provides the advantage of using DOA information to improve visual tracking. In the case of missing audio clue, the DOA is estimated by interpolation, based on those obtained from the previous frames where the DOAs may be available. If the gap of the missing audio clue is large, the accuracy of such interpolation will be limited. However, by making small changes in the proposed algorithm (details are omitted due to space constraints), the audio-visual tracker can be reduced to visual-only tracker when the DOA information is missing.

IV. Improved AV Tracking With Adaptive Particle Filter

The limitations of the baseline PF approach using a fixed number of particles have been discussed in Section I. To address these limitations, we propose a new adaptive approach to estimate the optimal number of particles at each iteration.

Fox [38] proposed an adaptive approach called KLD-sampling where the number of particles is estimated adaptively by bounding the tracking error of the PF. It uses the Kullback-Leibler (KL) divergence between the empirical distribution and the true posterior distribution, known as nonparametric maximum likelihood estimate, to measure the error. One assumption in this approach is that the true posterior can be represented by a discrete piecewise constant distribution consisting of a set of multidimensional bins. However, there is no certain way to estimate the size of these bins, and incorrect determination may cause deviation in the estimation of N. Also, it does not mention anything about the second fixed parameter of the PF, i.e. noise variance σ^2 whose selection affects the distribution of the particles, causing the tracker to become potentially unstable.

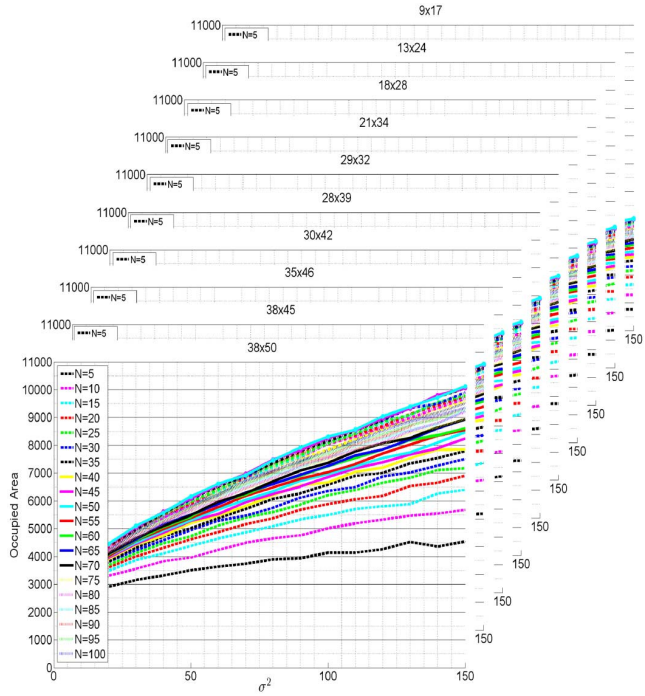

Fig. 5. Ten mapping tables for different L are created to observe the relation between N, S and σ^2.

In this study, we aim to design a new adaptive approach which addresses the problems in the KLD-sampling algorithm. More specifically we adapt both N and σ^2 dynamically in a simple way which is easily applicable to any implementation. The particles search a rectangular area to detect the face of the speaker before their weights are allocated. The accuracy of the speaker detection partly depends on the size of the area searched. We use this relationship and build our proposed algorithm on the area occupied by the rectangles centred on the positions of the particles [53]. The total area, S, occupied by the rectangles can be defined as

$$S = f(N, \sigma^2, L) \tag{10}$$

where L is the area of each rectangle. The value of S depends on the number of particles, the area of rectangle centred around each particle, and the overlap between the rectangles. The overlap is highly related to the distance between the particles, namely σ^2 which affects the distribution of the particles. One way to formulate the calculation of S is to analyse the relationships between S, N, L and σ^2 using mapping tables. These mapping tables are created by distributing N particles with the variance σ^2 and calculating the area of L pixels occupied by the particles. For each mapping table, N is varied from 5 to 100 with a step size of 5, and σ^2 is varied from 10 to 150 with a step size 10. For each point (for example, $N = 10$, $\sigma^2 = 50$), it is repeated 100 times and the average of the occupied area S is estimated. Therefore, the relationships between S, N and σ^2 are observed in one mapping table for a particular L. Then this process is repeated for ten different L s as illustrated in Fig. 5.

An illustration of the occupied area estimation is presented in Fig. 6. Based on the particle distribution, rectangles are drawn centred on the position of the particles. Since overlaps between

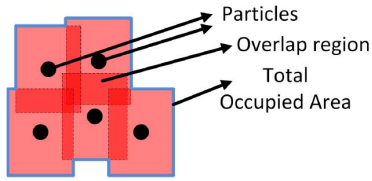

Fig. 6. The area inside the blue line indicates the total occupied area by five particles.

rectangles are inevitable, the total occupied area, S, is estimated by counting the number of pixels inside the blue line in Fig. 6.

For adaptive estimation of N, we need to describe 20 different lines in each 10 different mapping tables using a single formula. However, the behaviour of the lines in the mapping tables is nonlinear and this makes the problem intractable. As a solution, a curve fitting process is applied to linearise the nonlinear relation as shown in Fig. 5. Based on the goodness-of-fit test results, a polynomial model is chosen among the several candidate curve fitting methods. A pth order polynomial model is represented by $p+1$ coefficients. In our mapping tables, the occupied area, S, depends on three variables: σ^2, N and L. Therefore, the number of polynomial coefficients grows with the power of three. Clearly, there is a trade-off between the order of the model and the goodness-of-fit as measured in terms of the sum of squares due to error (SSE). A higher order leads to a lower SSE, but it requires a higher number of polynomial coefficients. As a trade-off, the order of the polynomial model is set to 2. Let us denote $\ell = [L^2 \quad L \quad 1]^T$, $\mathbf{n} = [N^2 \quad N \quad 1]^T$ and $\mathbf{m} = [(\sigma^2)^2 \quad \sigma^2 \quad 1]^T$. These three vectors form a tensor $\mathcal{V} = \ell \otimes \mathbf{n} \otimes \mathbf{m}$ where $\otimes$ is the outer product. Then, the total area, S, can be expressed as

$$S = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{m=1}^{3} c_{i,j,m} \cdot v_{i,j,m} \tag{11}$$

where $v_{i,j,m}$ is the element of the tensor $\mathcal{V}$ and $c_{i,j,m}$ is the element of tensor $\mathcal{C}$ containing the coefficients determined by the second order polynomial model fitting. After rearranging we get

$$\begin{aligned} S = {} & [(c_{1,1,1}L^2 + c_{2,1,1}L + c_{3,1,1})(\sigma^2)^2 + (c_{1,1,2}L^2 + c_{2,1,2}L \\ & + c_{3,1,2})\sigma^2 + (c_{1,1,3}L^2 + c_{2,1,3}L + c_{3,1,3})]N^2 + [(c_{1,2,1}L^2 \\ & + c_{2,2,1}L + c_{3,2,1})(\sigma^2)^2 + (c_{1,2,2}L^2 + c_{2,2,2}L + c_{3,2,2})\sigma^2 \\ & + (c_{1,2,3}L^2 + c_{2,2,3}L + c_{3,2,3})]N + [(c_{1,3,1}L^2 + c_{2,3,1}L \\ & + c_{3,3,1})(\sigma^2)^2 + (c_{1,3,2}L^2 + c_{2,3,2}L + c_{3,3,2})\sigma^2 + (c_{1,3,3}L^2 \\ & + c_{2,3,3}L + c_{3,3,3})] \end{aligned} \tag{12}$$

Equation (11) has 27 coefficients calculated by the curve fitting process and given in Table II. Then equation (12) can be simplified to

$$0 = \Upsilon N^2 + \Psi N + \Omega \tag{13}$$

where $\Upsilon = (c_{1,1,1}L^2 + c_{2,1,1}L + c_{3,1,1})(\sigma^2)^2 + (c_{1,1,2}L^2 + c_{2,1,2}L + c_{3,1,2})\sigma^2 + (c_{1,1,3}L^2 + c_{2,1,3}L + c_{3,1,3})$ $\Psi = (c_{1,2,1}L^2 + c_{2,2,1}L + c_{3,2,1})(\sigma^2)^2 + (c_{1,2,2}L^2 + c_{2,2,2}L + c_{3,2,2})\sigma^2 + (c_{1,2,3}L^2 + c_{2,2,3}L + c_{3,2,3})$ and

TABLE II
CURVE FITTING COEFFICIENTS

$c_{1,1,1}$	-2.14×10^{-12}	$c_{3,2,2}$	3.93×10^{-1}
$c_{2,1,1}$	6.42×10^{-9}	$c_{1,3,2}$	-2.71×10^{-6}
$c_{3,1,1}$	1.73×10^{-6}	$c_{2,3,2}$	1.41×10^{-2}
$c_{1,2,1}$	2.64×10^{-10}	$c_{3,3,2}$	55.02×10^{-1}
$c_{2,2,1}$	-9.5×10^{-7}	$c_{1,1,3}$	7.59×10^{-9}
$c_{3,2,1}$	-2.86×10^{-4}	$c_{2,1,3}$	-4.59×10^{-5}
$c_{1,3,1}$	-3.1×10^{-9}	$c_{3,1,3}$	-1.77×10^{-2}
$c_{2,3,1}$	-9.6×10^{-6}	$c_{1,2,3}$	-1.39×10^{-6}
$c_{3,3,1}$	-2.89×10^{-2}	$c_{2,2,3}$	7.86×10^{-3}
$c_{1,1,2}$	4.52×10^{-10}	$c_{3,2,3}$	29.04×10^{-1}
$c_{2,1,2}$	-1.99×10^{-6}	$c_{1,3,3}$	-7.44×10^{-5}
$c_{3,1,2}$	-2.14×10^{-3}	$c_{2,3,3}$	13.98×10^{-1}
$c_{1,2,2}$	-6.45×10^{-8}	$c_{3,3,3}$	165.18
$c_{2,2,2}$	3.17×10^{-4}		

$\Omega = (c_{1,3,1}L^2 + c_{2,3,1}L + c_{3,3,1})(\sigma^2)^2 + (c_{1,3,2}L^2 + c_{2,3,2}L + c_{3,3,2})\sigma^2 + (c_{1,3,3}L^2 + c_{2,3,3}L + c_{3,3,3}) - S$.

From equation (13) N can be readily found as

$$N = \frac{-\Psi + \sqrt{\Psi^2 - 4\Upsilon\Omega}}{2\Upsilon} \tag{14}$$

Note that L is estimated in every frame after the face of the speaker is detected. In practice, N is implicitly bounded by the choice (or calculation) of σ^2, L and S which usually take a limited range of values. In equation (14), S and σ^2 are unknown parameters that need to be estimated. To this end, we propose an iterative method where the values of S and σ^2 in step k are derived from the initial values confined by $\bar{\gamma}_k$ which is the difference between γ_k and γ_{k-1}. In other words, the calculation of S and σ^2 is linked to the difference of γ in successive frames. We propose to use a statistical model to establish that link. Many distribution functions could be employed. In our case, however, we have several requirements: (1) the input parameter should change between 0 to 1 (to match with the range of γ value); (2) the function may be controlled by at most two parameters (for simplicity); (3) the output of the function should be in the range of 0 to 1 to point out alteration ratio. To meet these requirements, a cumulative beta distribution (CBD) function appears to be the best choice and therefore it is used to model the link between $\bar{\gamma}_k$ and S, as well as σ^2. The CBD function is depicted in Fig. 7 and given in (15).

$$I_{\bar{\gamma}}(\alpha,\beta) = \sum_{j=\alpha}^{\alpha+\beta-1} \binom{\alpha+\beta-1}{j} \bar{\gamma}^j (1-\bar{\gamma})^{(\alpha+\beta-1-j)} \tag{15}$$

It needs two control parameters (α and β) and both input and output values change between 0 to 1. Then, S and σ^2 at time k are defined as

$$\begin{aligned} S_k &= S_0 * (\rho_S + sign(\bar{\gamma}_k) * I_{|\bar{\gamma}_k|}(\alpha_S, \beta_S)) \\ \sigma_k^2 &= \sigma_0^2 * (\rho_{\sigma^2} + sign(\bar{\gamma}_k) * I_{|\bar{\gamma}_k|}(\alpha_{\sigma^2}, \beta_{\sigma^2})) \end{aligned} \tag{16}$$

where α_S and β_S are the control parameters of the CBD for modelling S_k, and α_{σ^2} and β_{σ^2} are the control parameters of the CBD for modelling σ_k^2. S_0 and σ_0^2 are the initial values of S and σ^2. Absolute value of $\bar{\gamma}_k$ is used in the CBD function, because input values of CBD range between 0 to 1 and $\bar{\gamma}_k$ may be positive or negative depending on the change of γ in successive

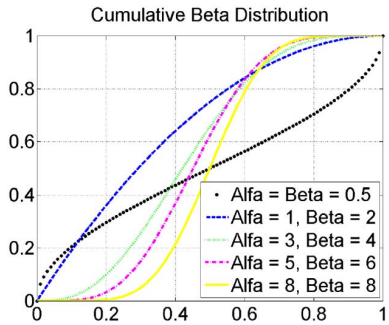

Fig. 7. Cumulative beta distribution function for different α and β values.

frames. The output of CBD is multiplied with the sign of $\bar{\gamma}_k$ to make the change of S_k and σ^2 dependent on the change in $\bar{\gamma}_k$.

The proposed AV-A-PF algorithm is an improved version of our proposed AV-PF algorithm explained in Section III. At every iteration, after the comparison of γ_k^{av} with γ_k, S_k and σ_k^2 values are updated using (16) in order to find the optimal N_k by (14). The last step of the PF algorithm is resampling and since the N_k value has just been changed, this step is also modified for the new N_k. If N_k is decreased, the particles with the smallest weights are removed. The particles with largest weights are duplicated if N_k is increased before the resampling step is performed. The pseudo code of the proposed AV-A-PF algorithm is depicted in Algorithm 3.

Algorithm 3: Proposed AV-A-PF Algorithm.

Initialize: N_0, σ_0^2, S_0, U, T, $\mathbf{F}$, λ, $r(u)$, $\mathbf{x}_0^{(n)}$, $w_0^{(n)}$, k
while $k < K$ **do**
  // AV Particle Filter - Section III.
  Calculate $\mathbf{x}_k^{(n)}$, $w_k^{(n)}$, $\tilde{\mathbf{x}}_k$ and γ_k using equation (1), (2), (4) and (3), respectively.
  Find movement distances by equation (7)
  Calculate $\hat{\mathbf{x}}_k^{(n)}$ and $\hat{w}_k^{(n)}$ using equation (8) and (9)
  Re-estimate target position: $\tilde{\mathbf{x}}_k^{av} = \sum_{n=1}^{N_k} \hat{w}_k^{(n)} \hat{\mathbf{x}}_k^{(n)}$
  // Adaptive approach modifications - Section IV
  Calculate γ_k^{av} using equation (3)
  if $\gamma_k^{av} < \gamma_k$ **then**
    $\mathbf{x}_k^{(n)} = \hat{\mathbf{x}}_k^{(n)}$, $w_k^{(n)} = \hat{w}_k^{(n)}$, $\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_k^{av}$, $\gamma_k = \gamma_k^{av}$
  end
  Calculate $\bar{\gamma}_k$ value: $\bar{\gamma}_k = \gamma_k - \gamma_{k-1}$
  Calculate new S value:
  $S_k = S_0 * (\rho_S + sign(\bar{\gamma}_k) * I_{|\bar{\gamma}_k|}(\alpha_S, \beta_S))$
  Calculate new σ^2 value:
  $\sigma_k^2 = \sigma_0^2 * (\rho_{\sigma^2} + sign(\bar{\gamma}_k) * I_{|\bar{\gamma}_k|}(\alpha_{\sigma^2}, \beta_{\sigma^2}))$
  Estimate optimal N using equation (14)
  Resampling: Generate $\mathbf{x}_k^{(n)}$ from the set $\{\mathbf{x}_k^{(n)}, w_k^{(n)}\}_{n=1}^{N_k}$
  $k = k + 1$
end

V. Experimental Evaluations

In this section, the proposed and baseline algorithms are evaluated on the $AV16.3$ dataset [54] and the results are presented in plots and tables. First, the experimental setup is described and the evaluation metrics are discussed. Then, comparative results between V-PF and our proposed AV-PF are given and discussed. Last, the performance of our adaptive algorithm AV-A-PF is compared with our AV-PF algorithm and the baseline KLD-sampling algorithm.

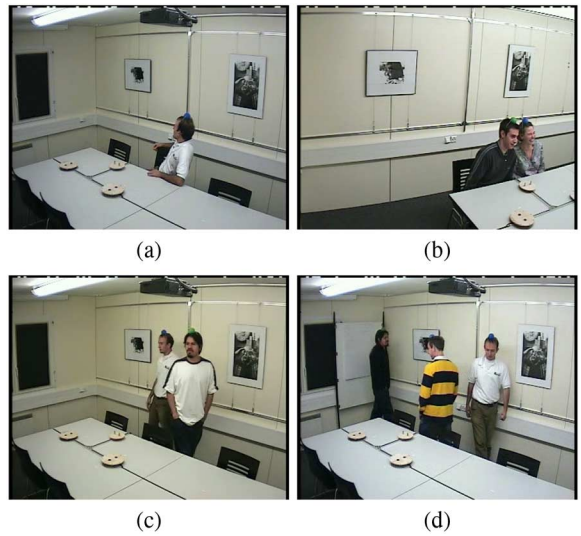

Fig. 8. Some challenging cases from $AV16.3$. The case of face rotation is shown in (a), contiguous faces in (b), and multispeaker occlusion in (c) and (d).

A. Setup

In order to perform a quantitative evaluation of the proposed algorithms, both audio and video sequences are required, together with the calibration information of the cameras and microphone arrays (circular arrays). Apart from "$AV16.3$", we have also explored the suitability of several other publicly available audio-visual datasets, such as "CLEAR" [55], "AMI" [56] and "SPEVI" [57], and concluded that only the $AV16.3$ dataset is suitable for the evaluation of our proposed methods. It complies with our requirements in terms of having circular microphone arrays with calibration information, mostly talking speakers, and challenging scenarios such as occlusion and rapid movements of the speakers. The other datasets do not fit at least one requirement of this study. For example, in "CLEAR" and "AMI", the speakers are mostly static or with small movements. In "SPEVI" and "CLEAR", the audio signals were acquired with linear microphone arrays. In addition, none of the three datasets contains the calibration information required for the quantitative evaluations of the proposed algorithm.

The corpus $AV16.3$ has many sequences for different scenarios where subjects 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 independently from each other. The audio signals were recorded at 16 kHz and the concurrent video sequences were recorded at 25 Hz. They were then synchronized before being used in our system. Each video frame is a color image of 288 $\times$ 360 pixels. Some sequences are annotated to get the ground truth speaker position which allows us to measure the accuracy of each tracker and compare the performance of the algorithms. To analyze the performance of the compared algorithms, several metrics are employed. The first one is the mean absolute error (MAE) which is estimated as the Euclidean distance in pixels between the estimated and the ground truth

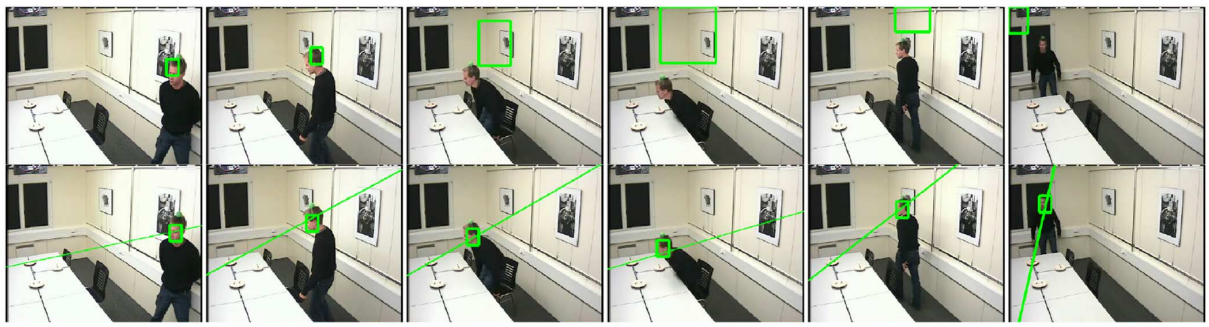

Fig. 9. Frames are taken from Sequence 11 camera #3. The first row shows the V-PF and the second row shows our proposed AV-PF tracking.

positions, then divided by the number of frames. This metric is chosen because of its simplicity and explicit output for the performance comparison. The algorithms are also evaluated by two other metrics. One is the multiple object tracking (MOT) metric proposed in [58], together with its quantities, MOT precision (MOTP) and MOT accuracy (MOTA). The MOTP measures the precision of the tracking system by comparing it with a threshold value pre-defined in terms of the Euclidean distance (either in pixels [59] or meters [58]). On the other hand, the MOTA measures the tracking configuration errors, consisting of the false positives (i.e., the case where the error is greater than the threshold value), false negatives (if the speaker is not tracked with the accuracy measured by the threshold) and mismatches (when the speaker identity is switched). The last metric is the trajectory-based measures (TBM) proposed in [60] and [61] which measures the performance on the basis of trajectory. According to their definitions, a trajectory can be categorized as mostly tracked (MT) or mostly lost (ML) if, respectively, at least 80% or less than 20% of its ground truth (GT) trajectory is covered by the tracker. Otherwise, it is considered as partially tracked (PT). Additionally, track fragmentation (Frag) is the total number of times that GT is interrupted in tracking result, and identity switches (IDS) measures the total number of times that a tracked trajectory changes its matched GT identity. We have evaluated our proposed algorithms and the baseline algorithms using both MOT and TBM metrics, and because of the space constraints, only overall average results are given in Sections V-B and V-C.

The speakers wear a coloured balls in particular sequences which are only used for annotation, but not for tracking in our system. In the experiments with the AV-PF algorithm, the number of particles, N, is selected to be 10. The covariance matrix $\mathbf{Q}$ is a diagonal matrix with $\sigma^2 = 50$, and this is used as the variance for both the position and velocity. For the AV-A-PF algorithm, N and σ^2 are estimated dynamically. T is the period between frames and equals 0.04 seconds and λ in (2) is chosen as 150. The number of bins used for Hue histogram is 8. The scale factors ρ_S and ρ_{σ^2} are set to 1. Both α_S and β_S values are chosen as 8, and α_{σ^2} and β_{σ^2} are chosen 0.5 for CBD functions. These α and β values are intuitively chosen based on expected response of the CBD function with respect to the error change (see Fig. 7). The value of S_0 is taken as 2000 and σ_0^2 is 50 in the simulations. These initial values are found to be appropriate based on cross-validation. In our work, we have

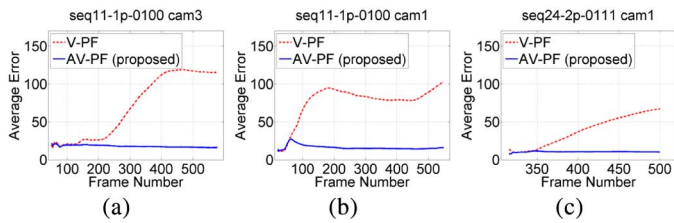

Fig. 10. In (a), (b), and (c), the performance of algorithms is given for Sequence 11 camera #3, Sequence 11 camera #1, and Sequence 24 camera #1, respectively.

used annotated DOAs as *a priori* to avoid mis-correspondence of person-ID after occlusion. Such information may not be available in a practical tracking system, and the person-IDs would have to be modelled and adapted during tracking using methods such as in [2] and [3].

Ten different sequences (3 single speaker, 5 two speakers and 2 three speakers) with three different camera angles from $AV16.3$ corpus have been used to perform the experiments. Some frames from $AV16.3$ are shown in Fig. 8. These selected sequences cover many challenging situations such as rotation of head [Fig. 8(a)], contiguous faces [Fig. 8(b)], occlusions [Fig. 8(c) and (d)] which make tracking much more difficult than an ordinary case.

B. Visual PF Versus Audio-Visual PF

The V-PF and AV-PF algorithms are run in thirty experiments (10 sequences with 3 different camera angles) in order to compare their performance. One of the single speaker experiments is Sequence 11 camera #3 in which the speaker is making random motions as illustrated in Fig. 9. Here, the speaker moves around the table and makes rapid and sudden movements. In the first row, the performance of the V-PF algorithm is shown. At the beginning, the tracker follows the speaker with small errors, but the error increases in challenging situations and eventually the tracker fails. In the second row, the results of the proposed AV-PF algorithm are given and here the tracker successfully follows the speaker with the assistance of the audio line. The plot in Fig. 10(a) shows the tracking error for this sequence. Here, the error at frame k is given as the average of the errors from frame 1 to k. This representation is chosen instead of plotting error on corresponding frame k, which would give oscillating graph since errors may change abruptly in subsequent frames.

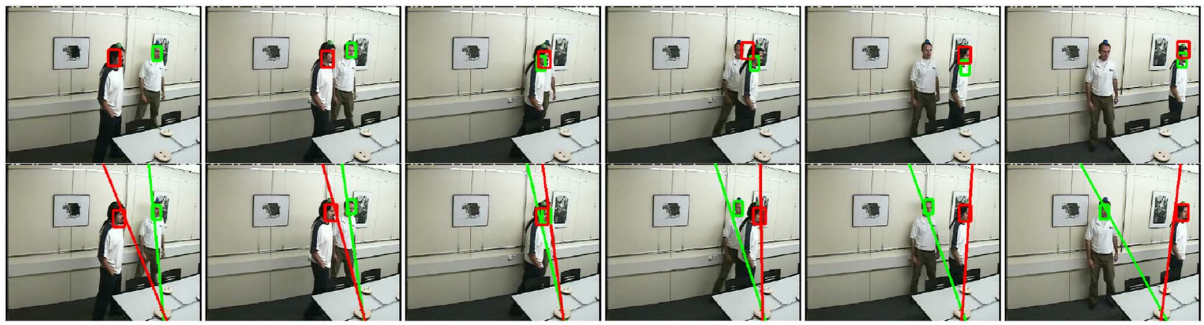

Fig. 11. Sequence 24 camera #1: Multiple speakers with occlusions. V-PF in the first row cannot track the speaker after occlusion. On the contrary, the proposed AV-PF algorithm keeps tracking.

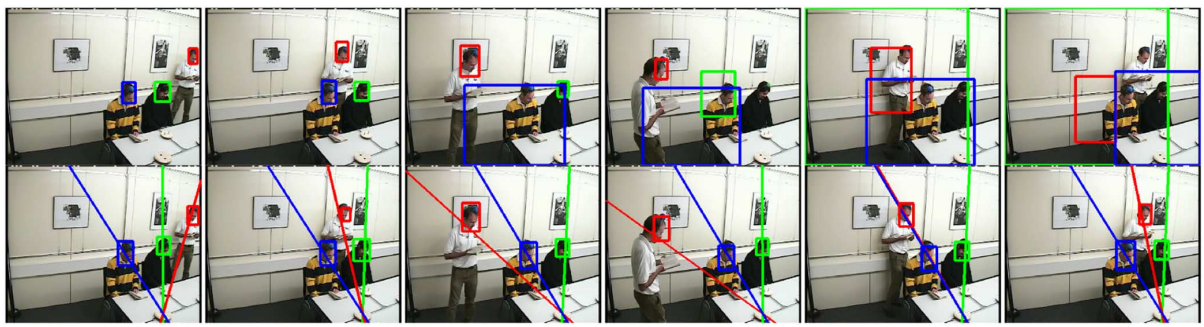

Fig. 12. Sequence 40 camera #1: Three speakers with occlusions. V-PF performance is shown in the first row, and AV-PF shows better performance in the second row.

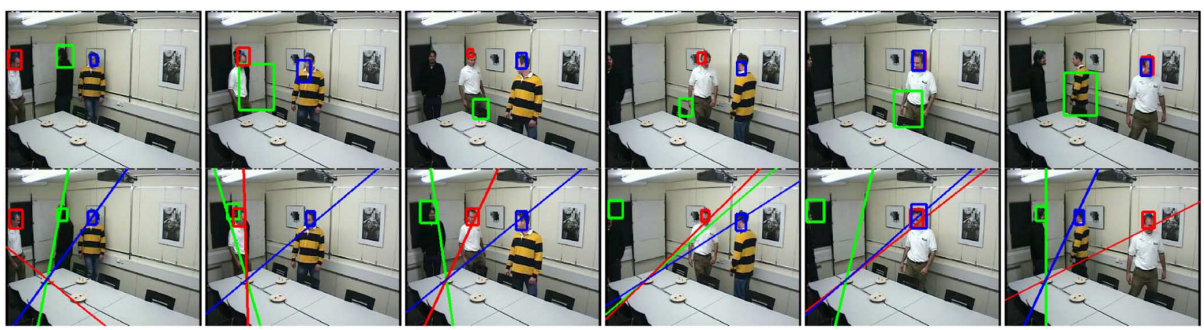

Fig. 13. Sequence 45 camera #2: Multiple speakers with occlusions. Occlusions occur multiple times. V-PF fails in the first row, but AV-PF continues to tracking in the second row.

On the other hand, plotting average error at each frame k gives smooth graph which can be interpreted easily and the overall performance of each tracker can be compared clearly.

The experiment for Sequence 11 camera #1 is given in Figs. 1 and 4 shown in Section II and III respectively where the speaker comes back after disappearing for a while. In Fig. 1, the V-PF approach results are given in the first row and as seen from the frames, when the speaker re-appears, the tracker fails to track the face. Contrary to the V-PF, tracking resumes with the reappearance of the speaker in our proposed AV-PF algorithm as shown in Fig. 4. The plot in Fig. 10(b) shows the tracking error for this sequence. After occlusion, the V-PF algorithm lost tracking, but our proposed AV-PF algorithm continued tracking.

Fig. 11 shows the result for multispeaker occlusion case, Sequence 24 camera #1 where one speaker is occluded by the other. After the occlusion, our proposed AV-PF algorithm (in the second row) resumes tracking. The average error of the two speakers for this sequence is shown in Fig. 10(c), and after the 350th frame the V-PF fails, but our proposed AV-PF algorithm continues tracking with small errors.

These two algorithms are also tested on the case of three speakers with two sequences, Sequence 40 and Sequence 45, respectively. The results for Sequence 40 camera #1 are illustrated in Fig. 12. Even this sequence is not challenging, V-PF fails to track all three speakers. Sequence 45 is the most challenging sequence in this corpus where all the speakers walk and occlude each other many times as shown in Fig. 13. The V-PF fails as expected. Unlike the V-PF, the proposed AV-PF algorithm successfully tracks the speakers both on Sequence 40 and Sequence 45. The error plots for three speaker experiments are given in Fig. 14. Both plots show that the proposed AV-PF approach has more stable performance than the V-PF.

Because of space constraints, we are not able to show all the frames for 30 experiments. All experiments are repeated 10 times and the results are given in Tables III and IV.

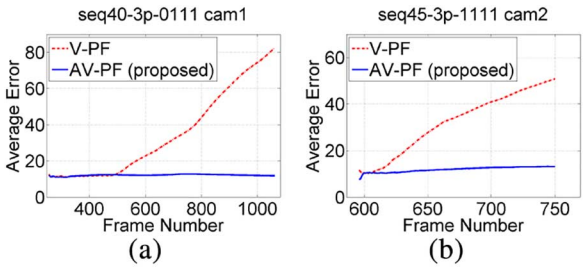

Fig. 14. In (a) and (b), the performance of the V-PF and AV-PF algorithms is given for Sequence 40 camera #1 and Sequence 45 camera #2, respectively.

TABLE III
EXPERIMENTAL RESULTS FOR V-PF AND PROPOSED AV-PF

		MAE	
		V-PF	AV-PF
seq08-1p-0100	cam1	109.36	11.74
	cam2	120.92	9.05
	cam3	83.81	10.91
seq11-1p-0100	cam1	81.15	22.27
	cam2	117.58	17.05
	cam3	130.50	18.29
seq12-1p-0100	cam1	145.01	16.54
	cam2	170.86	18.79
	cam3	157.25	12.77
seq18-2p-0101	cam1	115.26	15.24
	cam2	110.14	14.33
	cam3	115.54	19.25
seq19-2p-0101	cam1	62.51	12.67
	cam2	61.95	11.41
	cam3	56.49	12.91
seq24-2p-0111	cam1	73.34	10.66
	cam2	51.35	9.57
	cam3	44.20	10.44
seq25-2p-0111	cam1	33.79	16.17
	cam2	20.51	8.77
	cam3	33.52	10.09
seq30-2p-1101	cam1	33.72	15.21
	cam2	19.23	9.48
	cam3	26.35	10.86
seq40-3p-0111	cam1	75.69	12.88
	cam2	72.60	18.99
	cam3	78.94	16.05
seq45-3p-1111	cam1	73.81	17.19
	cam2	41.13	19.59
	cam3	71.58	20.94
Average		**79.60**	**14.34**

The V-PF and AV-PF algorithms are compared according to MAE in Table III, and using TBM and MOT metrics in Table IV. From these tables, it can be observed that the proposed AV-PF algorithm is consistently better than the V-PF algorithm.

Another experiment has been performed in order to see the effects of particle numbers on the performance of the algorithms. The numbers of particles are selected as: 10, 20, 30, 40, 50, 75, 100, 150 and 200 for all the three sequences. The results for Sequence 11 camera #1, Sequence 24 camera #1, and Sequence 11 camera #3 are shown in Fig. 15(a), (b) and (c), respectively. In the case of occlusion, the V-PF fails even if it has high numbers of particles as seen in Sequence 11 camera #1 and Sequence 24 camera #1. Sequence 11 camera #3 features a person making a variety of rapid movements, despite the fact that no occlusion is involved. The V-PF has almost the same performance as our proposed AV-PF algorithm when it has a larger number

TABLE IV
EXPERIMENTAL RESULTS WITH TBM AND MOT METRICS FOR V-PF AND PROPOSED AV-PF

Method	GT	MT	PT	ML	Frag	IDS	MOTA	MOTP	MAE
V-PF	57	43.0%	48.6%	8.4%	337	159	20.5%	14.6	79.6
AV-PF	57	92.4%	7.6%	0.0%	304	70	90.5%	12.7	14.3

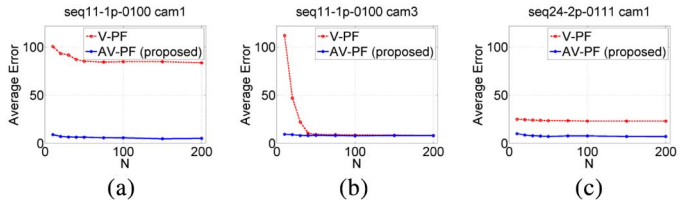

Fig. 15. Average error comparison between the V-PF and AV-PF algorithms for a variable number of particles N.

of particles as seen in Fig. 15(b). However, when the number of particles is reduced significantly, e.g. when $N = 10$, the tracking errors increase dramatically in the V-PF while our proposed AV-PF tracking algorithm continues to show excellent performance.

We have also compared our proposed AV-PF algorithm with one of the non-PF approaches, i.e. mean-shift tracking [62] which is a nonparametric statistical method. DOA is also fused in the same way as in our approach. As an example, here we show the results for Sequence 12 camera #3. Using the AV mean shift tracking, we got MAE $= 33.6$, MOTP $= 22.6$ and MOTA MOTP $= 60.3\%$. For our proposed approach, these values are 12.8, 12.3 and 97.9% respectively. It is clear that the PF approach outperforms the mean shift tracking algorithm.

C. Audio-Visual PF Versus Audio-Visual A-PF

The results in Fig. 15 show that the AV tracker is better than the visual-only tracker in handling occlusions even with a small number of particles. Here we demonstrate that we can further reduce the tracking errors by using our proposed adaptive approach, i.e. the AV-A-PF algorithm, as explained in Section IV. The AV-A-PF algorithm is also tested on $AV16.3$ and its performance is compared with the baseline algorithm, i.e. KLD-sampling [38]. Since the adaptive approach is based on our proposed AV-PF algorithm, the KLD-sampling is also combined with our proposed AV-PF algorithm in order to make a fair comparison between these two approaches.

To see the advantage of the A-PF, we perform an experiment to compare the proposed AV-A-PF algorithm with the use of a fixed number of particles (AV-PF). Firstly, the AV-A-PF algorithm is run on Sequence 12 camera #1 and we reach an average $\gamma = 0.27$ with an average $N = 15$ in 138.17 seconds. Then, the AV-PF is run with $N = 15$ and γ goes up to 0.35 in 135.87 seconds. The 30% γ difference shows that the AV-A-PF approach is better than the fixed number AV-PF approach with a more accurate tracking result. On the other hand, adaptive estimation of the particle numbers took around extra 2 seconds computational cost. However when we increase N up to 22 to get $\gamma = 0.27$ in the fixed AV-PF, the computational cost became 149.35 seconds. Here, experiments are implemented in

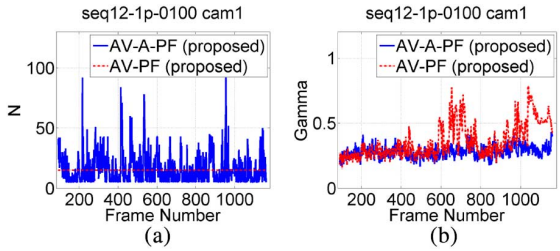

Fig. 16. The average N for the AV-A-PF and the fixed N for AV-PF is 15 in (a). The average γ for the proposed adaptive and fixed PF is 0.27 and 0.35, respectively, in (b).

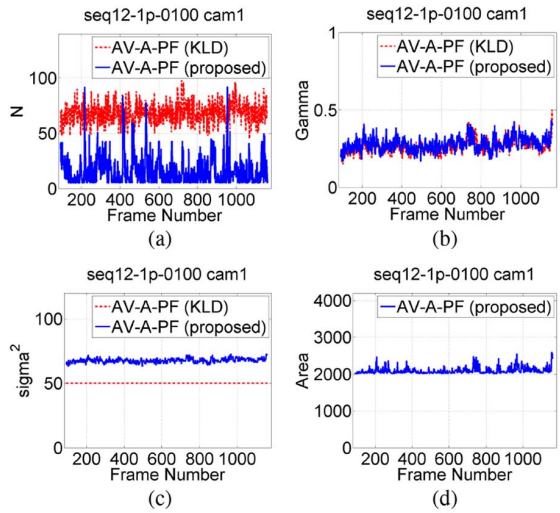

Fig. 17. The average N for the AV-A-PF algorithm and KLD-sampling is 15 and 68, respectively in (a). The average γ for AV-A-PF algorithm and KLD-sampling is 0.27 and 0.26 respectively in (b). In (c) $\sigma^2 = 50$ is used for KLD-sampling and while a changing σ^2 is for AV-A-PF algorithm with an average σ^2 equal to 67.61. In (d) change of S is given by time and the average S is 2083.

Intel core $i7$ 2.2 GHz processor with 8 GB memory under Windows 7 operating system. Adaptive estimation of N and variance adds slight computational cost, but it is reasonable when compared with fixed N usage to reach the same accuracy. It shows that the adaptive approach is beneficial both in terms of accuracy and the computational cost. The plot for this experiment is shown in Fig. 16. In Fig. 16(a), N is changing with time for AV-A-PF algorithm, while it is fixed for AV-PF and Fig. 16(b) shows γ for both approaches.

The KLD-sampling algorithm is also tested on the same sequence and compared with the AV-A-PF algorithm and the results are given in Fig. 17. The KLD-sampling algorithm needs an average of 68 particles to reach almost the same value, $\gamma = 0.26$. Fig. 17(a) and Fig. 17(b) show the effect of changing N and γ respectively. σ^2 is set to 50 in the KLD-sampling algorithm. Since σ^2 is adaptive in our proposed approach, the average σ^2 is found to be 67.61 as seen in Fig. 17(c). The effect of changing S is shown in Fig. 17(d) which is a parameter specific to our proposed approach.

In another experiment, we used three multispeaker sequences with speakers occluding each other. The results of these experiments are shown in Fig. 18. The AV-A-PF used an average of 27, 12 and 17 particles for Sequence 24 camera #2, Sequence 40 camera #1 and Sequence 45 camera #2, respectively. However,

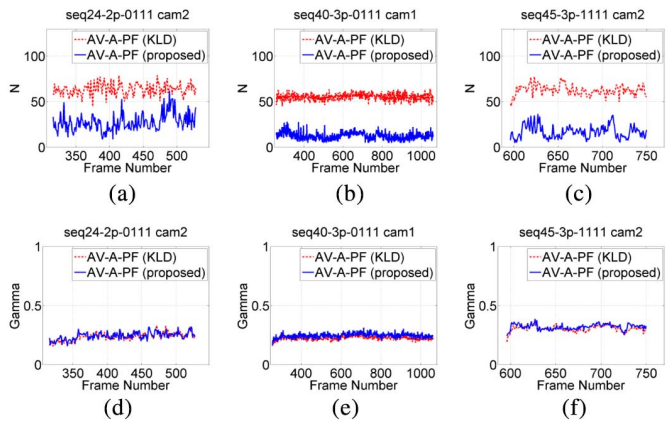

Fig. 18. Multi-person tracking. In (a), (b), and (c) the average N is given for both sampling algorithms. For Sequence 24 camera #2, the average N for the AV-A-PF and KLD-sampling is 27 and 64, respectively. For Sequence 40 camera #1, the average N for the AV-A-PF and KLD-sampling is 12 and 56, respectively. For Sequence 45 camera #2, the average N for the AV-A-PF and KLD-sampling is 17 and 63, respectively. The average γ is shown in (c), (d), and (e) for both algorithms. For Sequence 24 camera #2, it is 0.24 for both. For Sequence 40 camera #1, the average γ for the AV-A-PF and KLD-sampling is 0.24 and 0.22, respectively. For Sequence 45 camera #2, the average γ for the AV-A-PF and KLD-sampling is 0.32 and 0.30, respectively.

the KLD-sampling used 64, 56 and 63 particles for the same sequences. The difference in γ values is quite small despite the big difference in N. For Sequence 24 camera #2, it is 0.24 for both. For Sequence 40 camera #1, the average γ for the AV-A-PF and KLD-sampling is 0.24 and 0.22, respectively. For Sequence 45 camera #2, the average γ for the AV-A-PF and KLD-sampling is 0.32 and 0.30, respectively.

KLD-sampling is a popular approach in the literature, but one of the limitations of this approach is having only one adaptive parameter, N. Another limitation is that it needs a parameter, the bin size Δ, which also affects the performance of the algorithm. Generally, KLD-sampling shows better performance in the area of robotics in which tracking is done in a vast area with a large number of N (over 1000). In our adaptive approach, we have used the σ^2 value to find the optimal value for N. The errors can be reduced by adapting σ^2 without changing N. The mapping table also simplifies the calculation of N. These make AV-A-PF algorithm simple and efficient.

We performed KLD-sampling and AV-A-PF algorithm over 30 experiments. The results for all the experiments are given in Tables V and VI. Here also, all experiments are repeated 10 times and the averages of these results are shown in the tables.

Overall, our proposed AV-A-PF approach shows almost the same performance as the KLD-sampling despite it uses a much smaller N, as shown in Tables V and VI. The average of the estimated N by the KLD-sampling method for all sequences is around 53, while our proposed AV-A-PF algorithm gives an estimate of N at around 17. To examine whether the difference in N between these two methods is statistically significant, we have performed one-way ANOVA based F-test [63]. We obtained $F = 559.19$, p-value $= 1.8 \times 10^{-31}$ and the degree of freedom (1, 58). Using the degree of freedom value, the critical value F_{crit} is found to be 4.01 from the F-distribution table given in [63] which is the number that the test statistic must overcome to reject the test. The p-value (or probability value) is

TABLE V
EXPERIMENTAL RESULTS FOR KLD-SAMPLING AND PROPOSED AV-A-PF

		KLD		AV-A-PF	
		N	MAE	N	MAE
seq08-1p-0100	cam1	67.69	10.40	14.22	10.75
	cam2	48.58	7.27	20.19	7.33
	cam3	41.59	9.33	16.80	9.85
seq11-1p-0100	cam1	50.84	18.99	13.65	14.66
	cam2	53.55	15.73	20.80	14.01
	cam3	41.64	14.85	17.58	13.96
seq12-1p-0100	cam1	67.48	12.08	16.47	12.49
	cam2	56.48	9.66	18.15	10.81
	cam3	42.77	10.32	12.54	11.86
seq18-2p-0101	cam1	47.40	14.03	13.41	14.31
	cam2	52.67	11.30	11.96	11.66
	cam3	49.87	15.37	12.52	15.80
seq19-2p-0101	cam1	52.70	11.87	16.95	11.88
	cam2	57.65	9.49	24.52	9.62
	cam3	57.19	11.96	21.11	12.08
seq24-2p-0111	cam1	53.98	9.39	17.08	9.95
	cam2	62.60	8.58	27.80	8.85
	cam3	55.82	9.70	17.89	10.02
seq25-2p-0111	cam1	43.28	15.75	12.25	14.78
	cam2	59.52	7.39	12.16	7.70
	cam3	57.84	8.67	15.77	8.93
seq30-2p-1101	cam1	60.68	14.07	18.42	13.84
	cam2	50.22	8.57	14.04	8.85
	cam3	46.40	9.92	11.82	10.30
seq40-3p-0111	cam1	55.92	12.11	12.71	12.38
	cam2	49.19	11.58	12.86	12.04
	cam3	44.79	10.85	14.40	11.30
seq45-3p-1111	cam1	52.33	16.42	21.92	16.35
	cam2	62.59	14.24	17.75	17.22
	cam3	59.26	14.52	27.89	13.84
Average		**53.42**	**11.81**	**16.85**	**11.91**

TABLE VI
EXPERIMENTAL RESULTS WITH TBM AND MOT METRICS FOR KLD-SAMPLING AND PROPOSED AV-A-PF

Method	**GT**	**MT**	**PT**	**ML**	**Frag**	**IDS**	**MOTA**	**MOTP**	**MAE**	**N**
KLD	57	96.7%	3.3%	0.0%	185	39	94.7%	11.5	11.8	53.4
AV-A-PF	57	96.2%	3.8%	0.0%	234	44	94.4%	11.6	11.9	16.9

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 has been carried out at 5% significance level. According to this test, the results are accepted as statistically significant if $F < F_{crit}$ and p-value is less than 0.05 (for a 5% significance level). From the test results, we can observe that the difference in N between the two methods is indeed statistically significant.

VI. CONCLUSION

We have presented a new audio-visual tracking algorithm in which audio information has been used to modify particle propagation and the weights assigned to the particles. Our proposed algorithm has been tested on both single and multiple speaker sequences and showed significantly improved tracking performance over the V-PF approach for the scenarios where the speaker is either occluded by other speakers or out of the range of the camera view. We demonstrate that by using audio information we can significantly reduce the number of particles, whilst maintaining good tracking performance. This approach has the potential for handling weight degeneracy and particle impoverishment problems due to the significant reduction in the number of particles being used in tracking.

As an enhanced version of our proposed algorithm, we have presented a new adaptive PF algorithm which uses audio and visual information to adapt the number of particles and noise variance dynamically. Our proposed AV-A-PF algorithm has also been tested on both single and multiple speaker sequences and compared with a fixed particle filter and an existing A-PF algorithm. The experiments demonstrate that the proposed algorithm can effectively track moving speakers and increase robustness in tracking in the sense that it reduces the number of particles without increasing errors.

Despite the fact that our proposed algorithms offer advantages in speaker tracking and the estimation of the optimal number of particles, there are also some constraints and limitations associated with them that we want to point out. The first one is about the audio detection and localization algorithm which assumes that the microphone array is circular. Secondly, the audio information used in tracking is DOA, and as a result, the calibration information is required when projecting the DOA into the 2-D image plane. Third, we assume that the speaker to be tracked is active, from which the DOA information can be obtained. These assumptions or constraints may limit its generalization capability for other scenarios or datasets. However, with some modifications to the proposed algorithm, the proposed method could also be used in these cases. For example, if the audio localization algorithm used in the proposed tracking system is replaced by a linear microphone array based localization method together with the microphone calibration information, then the proposed system can also be applied to "CLEAR", "AMI" or "SPEVI" datasets. If the calibration information of the microphones is not available in the dataset, the proposed system could still be used, provided that the calibration information can be derived by a reliable self-calibration algorithm.

In conclusion, we first reduced the number of particles needed for tracking by combining audio information with V-PF, and then we converted AV-PF to AV-A-PF to increase the accuracy and robustness. The limitations associated with the proposed algorithms could be interesting directions for future work.

ACKNOWLEDGMENT

The authors wish to thank the associate editor and the anonymous reviewers for their contributions to improving the quality of this paper.

REFERENCES

[1] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer, "Audiovisual person tracking with a mobile robot," in *Proc. Int. Conf. Intell. Auton. Syst.*, 2004, pp. 898–906.
[2] M. J. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 828–836, 2003.
[3] O. Lanz, "Approximate Bayesian multibody tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1436–1449, Sep. 2006.
[4] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 125–141, 2008.
[5] M. M. Trivedi, H. S. Kohsia, and I. Mikic, "Dynamic context capture and distributed video arrays for intelligent spaces," *IEEE Trans. Syst., Man, Cybern. A: Syst., Humans*, vol. 35, no. 1, pp. 145–163, Jan. 2005.

[6] M. Isard and J. MacCormick, "Bramble: A Bayesian multiple-blob tracker," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, 2001, vol. 2, pp. 34–41.

[7] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 520–529, Sep. 2004.

[8] D. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 4, pp. 728–739, May 2008.

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

[10] Y. Bar-Shalom and T. Fortmann, *Tracking and Data Association*. Waltham, MA, USA: Academic Press, 1998.

[11] F. Talantzis, A. Pnevmatikakis, and L. C. Polymenakos, "Real time audio-visual person tracking," in *Proc. 8th IEEE Workshop Multimedia Signal Process.*, Oct. 2006, pp. 243–247.

[12] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 22–31, Jan. 2001.

[13] L. A. McGee and S. F. Schmidt, "Discovery of the Kalman filter as a practical tool for aerospace and industry," NASA, Moffett Field, CA, USA, NASATM-86847, 1985.

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

[15] 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.

[16] J. Vermaak, M. Gangnet, A. Blake, and A. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, 2001, vol. 1, pp. 741–746.

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

[18] N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 5, pp. V–881–V–884.

[19] R. P. S. Mahler, "Statistics 101 for multisensor, multitarget data fusion," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 19, no. 1, pp. 53–64, Jan. 2004.

[20] Q. Nguyen and J. Choi, "Localization and tracking for simultaneous speakers based on time-frequency method and probability hypothesis density filter," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2011, pp. 2866–2871.

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

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

[23] R. Mahler, "PHD filters of higher order in target number," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 4, pp. 1523–1543, Oct. 2007.

[24] N. T. Pham, W. Huang, and S. H. Ong, "Tracking multiple speakers using CPHD filter," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 529–532.

[25] S. Blackrnan and A. House, *Design and Analysis of Modern Tracking Systems*. Boston, MA, USA: Artech House, 1999.

[26] K. Panta, B. N. Vo, S. Singh, and A. Doucet, "Probability hypothesis density filter versus multiple hypothesis tracking," in *Proc. SPIE*, vol. 5429, pp. 284–295.

[27] S. Sarkka, A. Vehtari, and J. Lampinen, "Rao-blackwellized Monte Carlo data association for multiple target tracking," in *Proc. 7th Int. Conf. Inf. Fusion*, 2004, pp. 583–590.

[28] Y. Bar-Shalom and X.-R. Li, *Multitarget-Multisensor Tracking: Principles and Techniques*. Storrs, CT, USA: YBS Publishing, 1995.

[29] R. Chakravorty and S. Challa, "Multitarget tracking algorithm-joint IPDA and Gaussian mixture PHD filter," in *Proc. 12th Int. Conf. Inf. Fusion*, 2009, pp. 316–323.

[30] M. Jaward, L. Mihaylova, N. Canagaraja, and D. Bull, "Multiple object tracking using particle filters," in *Proc. IEEE Aerosp. Conf.*, 2006, pp. 8–15.

[31] K. Kim and L. S. Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," in *Proc. Eur. Conf. Comput. Vis.*, Graz, Austria, 2006, pp. 98–109.

[32] Y. Wang, Z. Jing, and S. Hu, "Data association for PHD filter based on MHT," in *Proc. 11th Int. Conf. Inf. Fusion*, 2008, pp. 1–8.

[33] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audio-visual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, pp. 601–616, Feb. 2007.

[34] 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. Select. Topics Signal Process.*, vol. 4, pp. 882–894, Oct. 2010.

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

[36] M. K. Pitt, R. d. S. Silva, P. Giordani, and R. Kohn, "On some properties of markov chain monte carlo simulation methods based on the particle filter," *J. Econometrics*, vol. 171, no. 2, pp. 134–151, 2012.

[37] V. Verma, S. Thrun, and R. Simmons, "Variable resolution particle filter," in *Proc. Int. Joint Conf. Artificial Intell.*, 2003, pp. 976–984.

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

[39] A. Soto, "Self adaptive particle filter," in *Proc. 19th Int. Joint Conf. Artificial Intell.*, 2005, pp. 1398–1403.

[40] P. Closas and C. Fernandez-Prades, "Particle filtering with adaptive number of particles," in *Proc. IEEE Aerosp. Conf.*, Mar. 2011, pp. 1–7.

[41] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *Proc. IEE-F Radar Signal Process.*, vol. 140, no. 2, pp. 107–113, 1993.

[42] L. Sigal, S. Sclaroff, and V. Athitsos, "Estimation and prediction of evolving color distributions for skin segmentation under varying illumination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2000, vol. 2, pp. 152–159.

[43] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wideband sources," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-33, no. 4, pp. 823–831, Aug. 1985.

[44] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.

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

[46] G. Lathoud, J. Bourgeois, and J. Freudenberger, "Sector-based detection for hands-free speech enhancement in cars," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 169–184, 2006.

[47] J. DiBiase, "A high-accuracy, low-latency technique for talker localisation in reverberant environments," Ph.D. dissertation, Brown Univ., Providence, RI, USA, 2000.

[48] G. Lathoud, "Spatio-temporal analysis of spontaneous speech with microphone arrays," Ph.D. thesis, EPFL Univ., Lausanne, Switzerland, 2006.

[49] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration—A Statistical Model Based Approach*. Berlin, Germany: Springer-Verlag, 1998.

[50] M. Crocco, A. D. Bue, M. Bustreo, and V. Murino, "A closed form solution to the microphone position self-calibration problem," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 2597–2600.

[51] M. Legg and S. Bradley, "A combined microphone and camera calibration technique with application to acoustic imaging," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 4028–4039, Oct. 2013.

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

[53] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Adaptive particle filtering approach to audio-visual tracking," in *Proc. IEEE 21st Eur. Signal Process. Conf.*, Sep. 2013, pp. 1–5.

[54] G. Lathoud, J. M. Odobez, and D. Gatica-Perez, S. Bengio and H. Bourlard, Eds., "AV16.3: An audiovisual corpus for speaker localization and tracking," in *Proc. 2004 Mach. Learn. Med. Imag. Workshop*, 2005, pp. 182–195.

[55] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo, "The clear 2007 evaluation," in *Multimodal Technologies for Perception of Humans*. New York, NY, USA: Springer, 2008, pp. 3–34.
[56] J. Carletta, S. Ashby, and S. Bourban *et al.*, "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*. New York, NY, USA: Springer, 2006, pp. 28–39.
[57] M. Taj, School of Electron. Eng. and Comput. Sci., Queen Mary Univ. of London, London, U.K., *Surveillance performance evaluation initiative (SPEVI) audiovisual people dataset*, 2007 [Online]. Available: http://www.eecs.qmul.ac.uk/~andrea/spevi.html, accessed Aug. 24, 2014.
[58] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP J. Image Video Process.*, vol. 2008, 2008.
[59] A. Sanin, C. Sanderson, and B. C. Lovell, "Shadow detection: A survey and comparative evaluation of recent methods," *Pattern Recog.*, vol. 45, no. 4, pp. 1684–1695, 2012.
[60] B. Wu and R. Nevatia, "Tracking of multiple, partially occluded humans based on static body part detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 1, pp. 951–958.
[61] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *Proc. IEEE Comput. Society Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 2953–2960.
[62] 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.
[63] 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, the M.Sc. degree in electronics engineering from Istanbul Technical University, Institute of Science and Technology, Istanbul, Turkey, and is working toward the Ph.D. degree from the Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford, U.K.

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.

Whilst completing his Ph.D., he worked with the IDIAP Research Institute as a Research Assistant. In 2006, he joined the Machine Vision Group, University of Oulu, Oulu, Finland, where he was Post-Doctoral 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. degree, the M.E. degree, and the Ph.D. degree from 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 Senior Lecturer and a Co-Director of the Machine Audition Lab. He has authored or coauthored over 130 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 IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Josef Kittler (M'74–LM'12) is Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing, Department of 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 published the textbook *Pattern Recognition: A Statistical Approach* (Englewood Cliffs, NJ, USA: Prentice-Hall, 1982) and over 170 journal papers.

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