Multiple Speaker Tracking in Spatial Audio via PHD Filtering and Depth-Audio Fusion

Qingju Liu, Wenwu Wang, *Senior Member, IEEE,* Teófilo deCampos, Philip J.B. Jackson, *Member, IEEE* and Adrian Hilton, *Member, IEEE*

***Abstract*—In object-based spatial audio system, positions of the audio objects (e.g. speakers/talkers or voices) presented in the sound scene are required as important metadata attributes for object acquisition and reproduction. Binaural microphones are often used as a physical device to mimic human hearing and to monitor and analyse the scene, including localisation and tracking of multiple speakers. The binaural audio tracker, however, is usually prone to the errors caused by room reverberation and background noise. To address this limitation, we present a multimodal tracking method by fusing the binaural audio with depth information (from a depth sensor, e.g., Kinect). More specifically, the PHD filtering framework is first applied to the depth stream, and a novel clutter intensity model is proposed to improve the robustness of the PHD filter when an object is occluded either by other objects or due to the limited field of view of the depth sensor. To compensate mis-detections in the depth stream, a novel gap filling technique is presented to map audio azimuths obtained from the binaural audio tracker to 3D positions, using speaker-dependent spatial constraints learned from the depth stream. With our proposed method, both the errors in the binaural tracker and the mis-detections in the depth tracker can be significantly reduced. Real-room recordings are used to show the improved performance of the proposed method in removing outliers and reducing mis-detections.**

***Index Terms*—Multi-person tracking, spatial audio, binaural microphones, depth sensor, depth and audio, PHD filtering**

I. Introduction

Object-based spatial audio [1]–[3] is becoming a trend for future spatial audio production and reproduction, which provides the opportunity to deliver an immersive and interactive listening experience. In object-based spatial audio, the original sound scene is represented by a number of audio objects, which can be transmitted and rendered at the reproduction stage; each audio object contains metadata describing important attributes and properties such as positions/trajectories of the sound source. As a result, in a common indoor sound scene involving several speakers, multi-person tracking needs to be performed for metadata extraction.

To monitor and evaluate the producer-generated spatial audio content as well as the rendered sound scene from a listener's perspective, binaural recordings are often collected via a binaural microphone, which is formed of two microphones located in the ears of a dummy head. The binaural audio provides geometrical information of the source location and thus can be exploited for metadata extraction. However, the tracking performance degrades in adverse acoustic environments involving multiple talkers due to the presence of competing speech, room reverberation and background noise [4]. To address this problem, extra information obtained from other sensors/modalities can be utilised, including colour (RGB) images [5]–[8], audio streams [4], [9]–[11], wireless network and mobile technologies [12], thermal infra-red sensors [13] and recently depth sensors such as stereo cameras [14], laser range finders [15] and RGB-depth (RGB-D) sensors [16]–[19]. Cross-modality tracking has attracted much attention in the last decade, mostly in the audio-visual domain [20]–[25], the RGB-D domain [26]–[29], and the RGB-D and thermal domain [30].

For our specific spatial audio applications in living room environments, a Kinect depth sensor is used for person tracking for the following reasons. Firstly, Kinect skeletal tracking offers the state-of-the-art real time performance in indoor environments. For a successfully tracked person, an average 3D position error of 1.3 mm is observed at the range of 1.2 m, which grows within the optimal depth range and reaches 6.9 mm error at 3.5 m [31]. Errors of $1-7$ mm are tolerable for listener-centred spatial audio applications. Secondly, it can cope with poor illumination conditions where dimmed light is often the only light source. Thirdly, Kinect devices are portable and easy to setup.

The Kinect depth tracker, however, also suffers from several limitations. Firstly, in the tracking results, there are outliers inconsistent with the trajectory of a target. In addition, identities (IDs) of the tracked persons often get swapped with each other in the presence of occlusions or noisy measurements. Secondly, the depth sensor may fail to track a target and thus yield mis-detections, especially when occlusions occur. To address the above limitations, we propose a systematic method based on modified probability hypothesis density (PHD) filtering [32] and depth-audio fusion, leading to the following main contributions.

PHD filtering is utilised in our system to mitigate outliers in the depth tracker, which is a state-space approach for handling multi-target tracking with unknown and varying number of targets. Sequential Monte Carlo (SMC)-PHD [33], [34] is used here, which avoids the prior Gaussian birth models in the closed-form Gaussian mixture (GM)-PHD [35]. We modify the SMC-PHD algorithm by integrating a novel clutter intensity model, which is measurement-driven and takes the depth sensor's limitations into account, i.e. the depth sensor's limited field of view (FOV). Moreover, the occlusion problem is also considered in the intensity model to prevent the SMC-

The authors are with the Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH, UK. e-mails: {q.liu, w.wang, t.decampos, p.jackson, a.hilton}@surrey.ac.uk.

PHD from converging to outliers erroneously associated with occluded persons.

To compensate mis-detected frames, we incorporate complementary information from the concurrent binaural audio stream, which is robust to factors that affect the depth sensor such as illumination conditions. The depth and audio streams are fused together by imposing trajectory constraints on the audio azimuth estimation, learned from the depth stream.

The remainder of the paper is organised as follows. We briefly introduce the overall proposed system and justify our specific system setup in Section II. Existing technologies and related background knowledge at different stages of the proposed tracking system are presented in Section III. Section IV describes in detail our contributions, including the PHD modification, and depth-audio fusion. Experimental results are presented and analysed in Section V. Finally, conclusions and insights for future work are given in Section VI.

II. Proposed System

A. System Overview

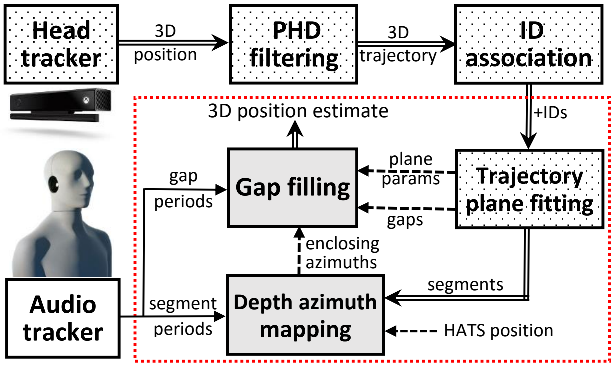


Fig. 1: Flow of the proposed depth-audio person tracking method. The single and double solid lines denote 1D and 3D data flow respectively, while the dashed lines denote model parameters. Processing of depth, audio and depth-audio modalities is represented by boxes with dots crosshatch, white and solid grey respectively.

The diagram of the proposed system is shown in Fig. 1. The *head tracker* is performed by the Kinect skeletal tracker directly. However, in an adverse environment that involves a high degree of occlusions, inconsistency is observed in the depth head tracking results both spatially (clutters) and temporally (mis-detections). Here, clutters are used to denominate failure cases of outlying position detections. The *PHD filtering* method is then applied to the head tracking results to remove these outliers/clutters, which is followed by an *ID association* scheme on a frame-by-frame basis, to address the remaining problem that IDs of detected persons may get swapped or assigned with new values. As introduced previously, mis-detections (gaps) exist in the depth tracking results when occlusions occur, which can be observed between consecutive depth detections (segments).

In parallel with the depth tracking, the *audio tracker* is applied to the binaural recordings collected by a head-and-torso simulator (HATS, the manikin as shown in the bottom left of Fig. 1). Audio azimuths of active speakers relative to the HATS can be estimated via the audio tracker, which are used to compensate mis-detections in the depth stream via depth-audio fusion.

Depth-audio fusion, as highlighted in the dashed box, contains three steps. Firstly, during the time periods when the depth tracker successfully tracks the targets and yields consecutive depth detections, i.e. segment periods, trajectory constraints for each detected target are learned via *trajectory plane fitting* techniques, with the assumption that head positions from the depth tracker lie on a plane. Secondly, 3D positions associated with each target obtained from the depth tracker, can be mapped to 1D depth azimuths relative to the HATS via *depth azimuth mapping*. Thirdly, during each time period when the depth tracker fails to track a target, i.e. the gap period due to mis-detections, audio azimuths are extended to 3D locations using a proposed *gap filling* technique, where the learned trajectory constraints and depth azimuths enclosing this gap are enforced. Since there are no valid depth detections during the periods of mis-detections, commonly used statistical modelling of bimodal features is not a good choice here.

In the proposed system, our main contributions lie in the modified PHD filtering with an adaptive clutter intensity model, as well as the depth-audio fusion that contains trajectory plane fitting, depth azimuth mapping and gap filling.

B. Setup Justification

In our system, we use a HATS as the binaural microphone to mimic human listening, to monitor and evaluate the sound scene from the listener's point of view. The HATS is located in the centre of the living room, such that 360 degree acoustic scene can be captured. Binaural recordings in the centre of the original sound scene provide immersive spatial perception of surrounding audio objects. The Kinect depth sensor is located at the edge of the sound scene, to ensure its optimal range [17] covers as much the living room as possible. This is a common Kinect setup. For instance, the Kinect sensor in Xbox 360 game console is suggested to be placed as a set top box, i.e., above the TV screen, near the edge of the room [36].

This particular HATS and Kinect setup, as illustrated in Fig. 2, enables range images to cover the indoor environment to a large degree, as well as a whole view of the acoustic scene comparable to that of the listener. In addition, since they are located at different positions with different and complimentary view angles, occlusions in one sensor can be compensated by the other. Moreover, the geometric mapping between the HATS's angular coordinate system and the Kinect's Cartesian coordinate system is straightforward, since the head position and orientation of the HATS can be directly obtained from Kinect. If the HATS and the Kinect were located at the same position, 1) the captured data would suffer from either a limited range data or a partially-reduced view of the acoustic scene; 2) their "shared view" would result in the concurrence of depth occlusions and audio occlusions. Similar setups with two modalities of sensors being located at two different positions has been employed for person tracking, e.g. a visual sensor and a compact microphone pair/array [22]–[25].

Traditional audio sensing for person tracking often involves a microphone array with a number of (distributed)

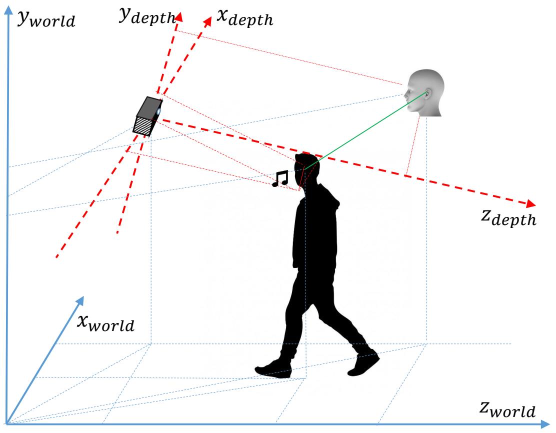


Fig. 2: Illustration of the Kinect depth sensor and the HATS in world coordinates. The two sensors face each other, standing at a distance with different height. The Kinect coordinates (dashed) are different from the world coordinates (solid). The depth sensor tracks a person with 3D position while the HATS gives the 1D azimuth of a talker. From Kinect's perspective, the HATS is also detected as a person.

microphones [4], [9]–[11]. Yet, besides issues of calibration and synchronisation, a microphone array does not provide binaural cues (i.e. interaural level difference and interaural phase difference) as would by the HATS as required for human-centred spatial audio production. Kinect also has a built-in 4-microphone array. However, its shared position/view with the depth sensor results in the aforementioned concurrent occlusion problem. In addition, errors mapped from the audio azimuth increase proportionally with the distance from the target to the audio sensor, while an audio object might be a few metres away. For instance, a 5 degree azimuth error yielded by the Kinect built-in microphone array [37] maps to 35 cm error for a target that is 4 m away.

Other modalities of sensing could also be considered, such as the commercially-used visual marker tracking [5] and low-cost camera tracking [6]–[8]. However, these methods are prone to the errors caused by challenging illumination conditions in indoor environments. In addition, multiple Kinect sensors [38], [39] could be used, which nevertheless involve increased complexity in calibration and synchronisation [40], [41]. Different from these alternatives, in our work, the use of binaural and Kinect sensing offers the advantage in capturing the listener centred spatial audio scene, and the convenience in object localisation for metadata extraction from the scene.

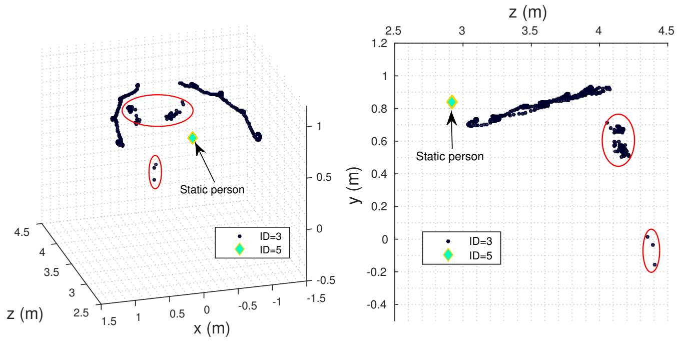
The proposed system uses the binaural audio to compensate mis-detections in the depth tracker. If no valid audio cues are available such as in a silence period, the system degenerates to depth-only tracking.

III. Background

In this section, we provide the background knowledge to help understanding the whole system in Fig. 1, including the classic PHD filtering, ID association and the audio tracker.

A. SMC-PHD Filtering

The Kinect skeletal tracker is limited by occlusions and FOV constraints, introducing outliers in the 3D head tracking results, as highlighted in the ellipses in Fig. 3. To estimate the target state in the presence of these outliers, filtering methods can be used. The classic Bayesian filtering framework and its relaxed models such as distributed Bayesian formulation [43] and linear programming [44] propagate the multitarget posterior, and this leads to increased computational complexity for an increasing number of targets. PHD filtering addresses the above limitation by propagating the first-order statistical moment of the multitarget posterior instead. The moment is also referred to as the probability hypothesis density or intensity, "whose integral is the expected number of targets" [32], i.e. the PHD (or intensity) models how densely the targets are distributed. The SMC-PHD framework [34] is used here.



(a) 3D view

(b) 2D side view

Fig. 3: Two views of depth-based head tracking results containing two participants in the second sequence of our recorded dataset [42]. The listener stands still in the centre (ID=5), highlighted by the cyan diamond. The other person (ID=3) walks behind along a semicircle as shown by the dotted trajectory. When this person gets occluded by the static person in front, head tracking becomes invalid and outliers are observed, as highlighted within these ellipses. The depth sensor is at position $[0, 0, 0]^\top$, the z axis is depth and x is roughly aligned with the horizon. The x-z plane is not necessarily parallel to the floor, in the acquisition of our dataset the sensor was approximately 23° tipped downwards.

Assume m_k persons are detected at the k-th frame, denoted as the observation set $\mathbf{Z}_k = \{\mathbf{z}_1, \cdots, \mathbf{z}_{m_k}\}$, where each element $\mathbf{z}$ is a 3D position vector $\mathbf{z} = [x, y, z]^\top$ within the Kinect coordinate system shown in Fig. 2. From the noisy observation sequence $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k$ that contains clutters and mis-detections, we aim to estimate the real target status $\mathbf{X}_k = \{\mathbf{x}_1, \cdots, \mathbf{x}_{n_k}\}$ at the k-th frame, where $\mathbf{x}$ represents the position as well as the velocity of a target $\mathbf{x} = [x, y, z, \dot{x}, \dot{y}, \dot{z}]^\top$, and n_k is the number of targets.

For a target at the $(k-1)$-th frame, i.e. $\mathbf{x}_{k-1}$, it can either survive with probability P_S or disappear (die) with probability $1-P_S$ at the next frame k. For a surviving/persistent target, its new state $\mathbf{x}_k$ evolves from $\mathbf{x}_{k-1}$ by following the single target transition density $\pi(\mathbf{x}_k|\mathbf{x}_{k-1})$, which approximates the motion model. Moreover, a new target might be born with intensity of ν. The detection uncertainty is also considered when a target might be mis-detected with probability P_D. A successfully detected target $\mathbf{x}_k$ is related to the associated head tracking result $\mathbf{z}_k$, with a single target likelihood function $g(\mathbf{z}_k|\mathbf{x}_k)$.

In SMC-PHD, the target intensity at time k is represented by N_k particles $\mathbf{x}_k^{(n)}$ with weights $w_k^{(n)}, n = 1, 2, \cdots, N_k$.

Algorithm 1: FRAMEWORK OF SMC-PHD.

```
Input: Measurement set $\mathbf{Z}_k$, particles from the previous
       frame $\{(\mathbf{x}_{k-1}^{(n)}, w_{k-1}^{(n)})\}_{n=1}^{N_{k-1}}$
Output: Particles $\{(\mathbf{x}_k^{(n)}, w_k^{(n)})\}_{n=1}^{N_k}$, state $\mathbf{X}_k$
% Prediction step
foreach $n = \{1, 2, \cdots, N_{k-1}\}$ do
    Draw $\mathbf{x}_{k|k-1}^{(n)} \sim \pi(\cdot|\mathbf{x}_{k-1})$
    $w_{k|k-1}^{(n)} = P_S w_{k-1}^{(n)}$
foreach $\mathbf{z} \in \mathbf{Z}_k$ do
    Draw $M_b$ new particles $\sim \mathcal{N}(\cdot|\mathbf{z}, \Sigma)$, with equal
    weight of $\frac{\nu}{m_k M_b}$
% Update step
foreach $n = \{1, 2, \cdots, N_{k-1} + m_k M_b\}$ do
    if $n <= N_{k-1}$ then
        $w_{k|k}^{(n)} =$
        $(1 - P_D) w_{k|k-1}^{(n)} + \sum_{\mathbf{z} \in \mathbf{Z}_k} \frac{P_D g(\mathbf{z}|\mathbf{x}_{k|k-1}^{(n)}) w_{k|k-1}^{(n)}}{\mathcal{L}(\mathbf{z})}$
    else
        $w_{k|k}^{(n)} = \sum_{\mathbf{z} \in \mathbf{Z}_k} \frac{w_{k|k-1}^{(n)}}{\mathcal{L}(\mathbf{z})}$
    where
    $\mathcal{L}(\mathbf{z}) = \kappa(\mathbf{z}) + \nu + \sum_{j=1}^{N_{k-1}} P_D g(\mathbf{z}|\mathbf{x}_{k|k-1}^{(j)}) w_{k|k-1}^{(j)}$
% Resampling
Resample $\{(\mathbf{x}_{k|k}^{(n)}, w_{k|k}^{(n)})\}$ to obtain $\{(\mathbf{x}_k^{(n)}, w_k^{(n)})\}$.
% State estimation
Cluster particles for the final state $\mathbf{X}_k$ estimation.
```

The new-born target intensity ν is represented by m_k groups of Gaussian-distributed particles, with each group centred at one detected person containing M_b new-born particles. $\kappa(\mathbf{z})$ is the clutter intensity [32], [45], whose integral over $\mathbf{z}$ is the expected number of clutters in the current frame. SMC-PHD propagates over time with recursive prediction and update steps, whose principles are summarised in Algorithm 1.

To estimate the final positions, a simple clustering method can be applied as follows. Particles in the proximity of a measurement $\mathbf{z}$, i.e. whose distances to $\mathbf{z}$ are smaller than a pre-defined threshold ζ_0, are grouped as a cluster. If the accumulated weight is greater than 0.5, then we consider that there exists a target at $\mathbf{z}$.

B. ID Association

Although the outliers can be mitigated with the PHD filtering, a problem persists in the head tracking results, that the IDs of detected targets sometimes get swapped or assigned with new values when the head tracker re-detects a person, introducing inconsistent IDs as illustrated in Fig. 4.

We proposed in our early work in [46] an ID association scheme with short- and long-term analysis. The principle for the short-term analysis is to keep the consistency and continuity of a target's movements within a small time interval. The long-term analysis exploits the common scenario when a person gets occluded by the persons in front and is therefore mis-detected, whose ID is inaccurately assigned when this person is re-detected. Both the PHD filter and the ID association scheme are performed on a frame-wise basis, thus they can be combined together at each frame.

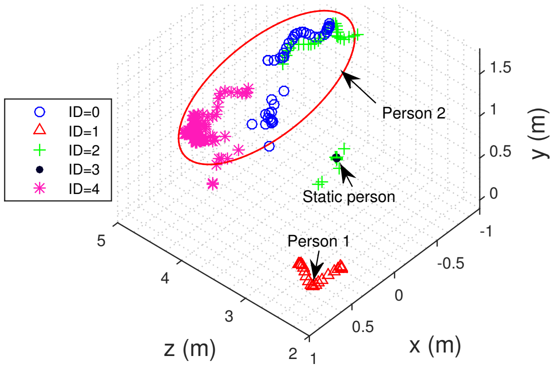


Fig. 4: Applying Kinect head tracking to the third sequence involving three people in our recorded dataset [42]. The static person and Person 1 are close to the sensor, which often occlude Person 2. This results in inconsistent IDs being assigned to Person 2. It can be observed that the detected trajectory for Person 2 is represented with three different patterns (IDs), which means the ID for Person 2 changed at least twice, of which one is shared with the static person in the centre. For illustration purposes, the original detected points are down-sampled.

C. Audio Tracker

Time delay of arrival (TDOA) cues have been widely used in audio tracking [4], [9]–[11], which are calculated by comparing the difference between a pair of microphones. TDOA cues are often obtained via finding the peak positions from the generalised cross correlation (GCC) [47] function based on the maximum likelihood (ML) principle. The phase-transform GCC (PHAT-GCC) function provides more robustness against noise. Suppose $L_k(\omega)$ and $R_k(\omega)$ are the short time Fourier transform (STFT) of the two audio segments at time frame k. The PHAT-GCC function can be calculated as:

$$C(\tau) = \int_{-\infty}^{\infty} \frac{L_k(\omega) R_k^*(\omega)}{|L_k(\omega) R_k^*(\omega)|} e^{j\omega\tau} \, d\omega, \tag{1}$$

where the superscript $*$ is the conjugate operator. For practical implementation, its discrete version, i.e. summation over frequency bins, is employed.

If the microphone pair is a binaural microphone, e.g. a HATS, the TDOA cues become interaural time difference (ITD) introduced by binaural room impulse responses (BRIRs). TDOA cues provide information about the bilateral azimuth or input angle of the speaker relative to the microphone pair. The relationship between the azimuth α versus the TDOA τ varies between different binaural microphone models. From three sets of BRIRs recorded with different HATSs in different rooms, we found a similar pattern in the relationship, as illustrated in Fig. 5, that the relationship between α and τ can be approximated with a third-order polynomial:

$$\tau = \mathcal{T}(\alpha) = p_1 \alpha + p_3 \alpha^3, \tag{2}$$

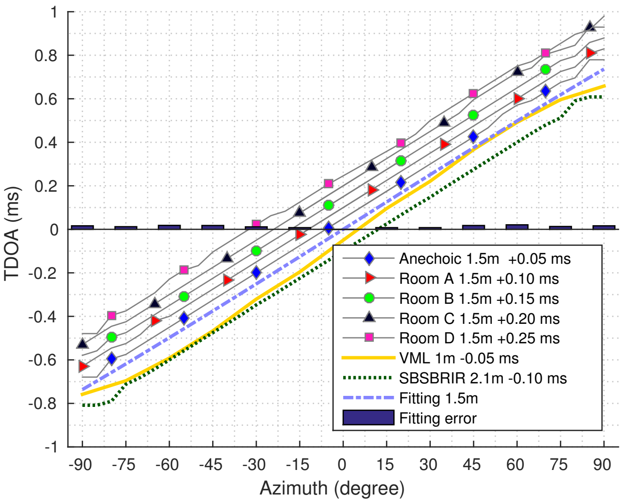


Fig. 5: Illustration of the relationship between different azimuths and the yielding TDOA cues, obtained from three BRIR datasets. The first dataset is recorded by Cortex MK2 HATS in five different rooms with RT60 of $[0, 320, 470, 680, 890]$ ms, denoted as Anechoic, Room A, Room B, Room C and Room D respectively. The distance between the sound source and the HATS is fixed at 1.5 m [48]. The second dataset is recorded with a Brüel & Kjær HATS denoted as SBSBRIR, with the fixed distance of 2.1 m [49]. The third dataset is recorded by Cortex MK2 HATS denoted as VML with the fixed distance of 1 m [42]. The relationship is consistent with different binaural head models, distances and reverberation levels. The dash-dotted line is fitted on the first dataset with a third-order polynomial. The TDOA features are shifted for illustration purposes and the average fitting error is calculated over the above three datasets.

with α being the azimuth in front of the binaural microphone in degree ($[-90, 90]$) and τ in millisecond. For an azimuth that is from the back of the HATS ($[-180, -90)$ and $(90, 180]$), it should be first mirror-symmetrically mapped to the front to calculate the resultant delay, due to the front-back ambiguity of the binaural microphone.

Note that, there are some essential limitations to the audio cues, especially in complex environments with high reverberation and strong background noise. However, the biggest challenge is the dynamic nature of an acoustic environment and nonstationarity of speech signals such as the varying number of speakers and energy fluctuations during a conversation.

IV. Our Contributions

A. PHD with A Novel Clutter Intensity Model

In SMC-PHD filtering in Algorithm 1, the clutter intensity $\kappa(\mathbf{z})$ plays a critical role in the convergence. When the overall clutter intensity is bigger than 1, the weights of all particles (new-born or not) decrease. When it is small, e.g. 0.1, the weights of new-born targets dramatically increase. Instead of using the uniform distribution as in [34], [35], here we present a novel measurement-driven clutter intensity model that takes into account the depth sensor's FOV as well as occlusion detection:

$$\kappa(\mathbf{z}|\mathbf{Z}) = \kappa + \kappa_1(\mathbf{z}) + \kappa_2(x, z) + \kappa_3(\mathbf{z}|\mathbf{Z}). \tag{3}$$

In the above equation κ is the clutter intensity for a measurement in the FOV of the depth sensor. κ_1 is the clutter intensity increment for a measurement which is out of the sensor's person tracking range, i.e. the near range $r_n = 0.5$ m and far range $r_f = 4.5$ m [17].

$$\kappa_1(\mathbf{z}) = \begin{cases} \dfrac{r_n - \|\mathbf{z}\|}{r_n} c_n, & \text{if } \|\mathbf{z}\| < r_n \\ \dfrac{\|\mathbf{z}\| - r_f}{r_f} c_f, & \text{if } \|\mathbf{z}\| > r_f \end{cases} \tag{4}$$

where $\|\cdot\|$ is the Euclidean norm. c_n and c_f determine how quickly the intensity increases when a target is out of range. $\kappa_2(x, z)$ is the clutter intensity increment based on the depth sensor's FOV. Kinect2 has a wide view angle of $70.6°$ in the horizontal (x-z) plane, and we define

$$\kappa_2(x, z) = 1, \text{ if } |x| > \tan(35.3°)|z|, \tag{5}$$

where $|\cdot|$ is a modulus operator. An increment of 1 in κ_2 attenuates all particles out of the FOV. $\kappa_3(\mathbf{z}|\mathbf{Z})$ considers the case that a detected target $\mathbf{z}_i$ is likely to occlude $\mathbf{z}$, if it is closer to the sensor (i.e. to the origin) than $\mathbf{z}$, i.e. $\|\mathbf{z}_i\| < \|\mathbf{z}\|$. The motivation is to increase the clutter density for the observation $\mathbf{z}$ if it is under the occluded area of other detected targets. The occlusion-based clutter function was then defined with a Gaussian mixture model:

$$\kappa_3(\mathbf{z}|\mathbf{Z}) = \sum_{\mathbf{z}_i \in \mathbf{Z}, \|\mathbf{z}_i\| < \|\mathbf{z}\|} w \exp\left(\frac{-d_i^2 \|\mathbf{z}_i\|^2}{2\delta^2 \|\mathbf{z}\|^2}\right), \tag{6}$$

where w determines the increment amount. δ is the half-width of $\mathbf{z}_i$. Scaled by the Euclidean relative distance, the half occluded width by $\mathbf{z}_i$ is $\frac{\|\mathbf{z}\|}{\|\mathbf{z}_i\|}\delta$ at the distance $\|\mathbf{z}\|$, centred by the line connecting the origin and $\mathbf{z}_i$, and d_i is the distance from $\mathbf{z}$ to this line.

B. Trajectory Plane Fitting

The audio azimuth with respect to the HATS extends over a conical surface, known as the cone of confusion. To locate the 3D position, a simple and effective solution is to use the fact that a person's head tends to move on trajectories that are roughly parallel to the ground plane. With this prior information, the search space is greatly simplified, as the cone of confusion becomes a pair of lines on a 2D plane. With a further distance constraint learned from the depth trajectory, azimuths enable the estimation of 3D trajectories during periods of occlusion.

A person's head trajectory can be approximated by a plane, if the person does not engage complex movements such as jump and bending over. This is usually the case in most tracking applications, such as surveillance in public spaces. The plane of possible head positions for the i-th person can be represented by:

$$ax + by + cz + d_i = 0, \tag{7}$$

where $[a, b, c]^\top$ is the normal vector of the plane, which is the same for all the persons, and d_i is related to the height. If the sensor's principal axis is orthogonal to the x-axis, we can enforce $a = 0$, which is standard for depth sensors developed for games. Ignoring the velocity in the state vector from the PHD-filtered results, and assuming $\mathbf{x}_k$ to be the position for the i-th person at the k-th frame, the plane parameters can be found as follows.

Denoting $\bar{\mathbf{x}}_k$ as the 3D position associated with one target after subtracting the mean $\bar{\mathbf{x}}$ from $\mathbf{x}_k$, we apply eigenvalue decomposition (EVD) to the expectation of $\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$, where the columns of $\bar{\mathbf{X}}$ are $\bar{\mathbf{x}}_k$ at different frames. The normal vector equals the eigen vector corresponding to the minimum eigen value, and $d_i = [a, b, c]\bar{\mathbf{x}}$. This plane is parallel to the floor, and thus, floor detection from the depth map can also be exploited to estimate the normal vector, though this is not trivial in cluttered environments. Although most of the outliers have been removed from the PHD filtering, there are still some residue outliers that might affect EVD. To address this limitation, we employ a bootstrap plane-fitting method, by iteratively choosing 95% of the data that best match the plane and then applying EVD.

C. Depth Azimuth Mapping

With the prior information of the 3D microphone position and orientation in the Cartesian coordinate of the depth sensor, we can calculate the 1D azimuth of each person relative to the HATS. To avoid the issue caused by the height difference between the person and the HATS, we map the HATS to the plane associated with the person, denoted as $\mathbf{x}_0 = [x_0, y_0, z_0]^\top$. Ignoring the velocity, the person detected at $\mathbf{x} = [x, y, z]^\top$ yields the depth azimuth

$$\alpha^{(d)}=\begin{cases} \arctan\dfrac{x_0-x}{\sqrt{(y_0-y)^2+(z_0-z)^2}}, & \text{if } z_0\geq z \\ \left[180-\arctan\dfrac{x_0-x}{\sqrt{(y_0-y)^2+(z_0-z)^2}}\right]_{-180}^{180}, & \text{otherwise,} \end{cases} \tag{8}$$

where $[\cdot]_{-180}^{180}$ maps the angle into the range of $[-180, 180]$. The above depth azimuth mapping process depends on the setup of the Kinect and the HATS, which face each other in this paper. Different geometric mapping from 3D positions to 1D azimuths can be built flexibly for other setups.

If the depth azimuth $\alpha^{(d)}$ equals the real audio azimuth α, it should be related to the TDOA feature τ with Eq. (2). However, the HATS might be slightly rotated, which results in an azimuth shift Δ:

$$\tau = \mathcal{T}(\alpha^{(d)} + \Delta). \tag{9}$$

This azimuth shift Δ can be calculated by minimising the difference between the associated TDOA feature τ_k from the audio stream with the analytical TDOA from the depth stream:

$$\hat{\Delta} = \operatorname*{argmin}_{\Delta} \sum_k \|\tau_k - \mathcal{T}(\alpha_k^{(d)} + \Delta)\|^2. \tag{10}$$

Note that, at each time frame k, there might be several TDOA features and several depth detections, and we need to choose only these *matched* pairs for optimising Δ. The direct association between the audio and the depth streams becomes a speaker diarisation problem [50]. Even if the audio features are not continuous in all the frames during the segment periods, e.g. a speaker might stop talking, Δ can still be statistically derived given enough concurrent depth+audio pairs, for aligning the audio and depth information. When the depth stream is missing but audio is available, Δ will be used in the gap filling stage, introduced as follows.

D. Gap Filling

During the segment periods when the depth tracker successfully tracks the involved targets, our proposed method will use the 3D tracking results from the PHD-filtered head tracker. However, gaps are observed when the depth tracker fails to detect the target. Take a gap spanning the time period between t_1 and t_2 for example. We aim to estimate the person's azimuth at each frame k s.t. $[t_1 f_d] < k < [t_2 f_d]$ from the concurrent audio and then map it to 3D space, where $[\cdot]$ rounds a number to its nearest integer and f_d is the depth stream sampling rate. Suppose the two points enclosing this gap (beginning and end points) yield the depth azimuths of $\alpha_1^{(d)}$ and $\alpha_2^{(d)}$ respectively, which are equivalent to the audio azimuths $\alpha_1 = \alpha_1^{(d)} + \Delta$ and $\alpha_2 = \alpha_2^{(d)} + \Delta$. We further denote L_1 and L_2 as the distances of the enclosing points to the HATS. Note that, for the gap in the beginning or end of each sequence, we have only one point enclosing this gap.

During the gap period $[t_1, t_2]$, we implemented a single-target particle filtering method to the associated audio TDOA features, to robustly track the audio azimuths. The linear Gaussian state-space model, as used in Kalman filtering [51] is exploited here. The TDOA measurements reflect velocity changes and the angular velocity for each particle is evolving with time using the employed motion model. The weights of the particles that follow the velocity changes will be increased. For particles that are not adaptive to the velocity changes, their weights will be attenuated. The two enclosing audio azimuths α_1 and α_2 are used for initialisations and thresholding the particles. The audio tracker can detect only active speakers and if there is no valid audio, due to the lack of both depth and audio data during the gap period, the proposed method cannot re-track the mis-detected target and hence will fail in this case. However, in spatial audio systems, metadata (e.g. the positions of the speakers) are only required for the active speakers.

To map this 1D azimuth to 3D position, we calculate the distance at the j-th frame as $L_j = L_1 + \frac{(L_2 - L_1)}{J}(j - 0.5)$ with a linear assumption. The 3D position for the j-th frame is

$$\begin{bmatrix} x_0 + L_j \cos(\alpha_j + 90^\circ) \\ y_0 - L_j \sin(\alpha_j + 90^\circ)\frac{c}{\sqrt{b^2+c^2}} \\ z_0 - L_j \sin(\alpha_j + 90^\circ)\frac{b}{\sqrt{b^2+c^2}} \end{bmatrix}. \tag{11}$$

As for the special occasion at the beginning or end of each sequence, where the beginning point or the end point is missing, we analyse the trajectory as follows. Line fitting is applied to the valid segment enclosing the gap on the associated plane, and the fitted line is extended during the gap period, which we refer to as the depth line. As a contrast, the audio line starts from the HATS, and exhibits the current audio azimuth α_j, i.e. crosses the HATS as well as a point defined in Eq. (11) with any length L_j (e.g. $L_j = 1$). The intersection between the depth and audio lines is the estimated position.

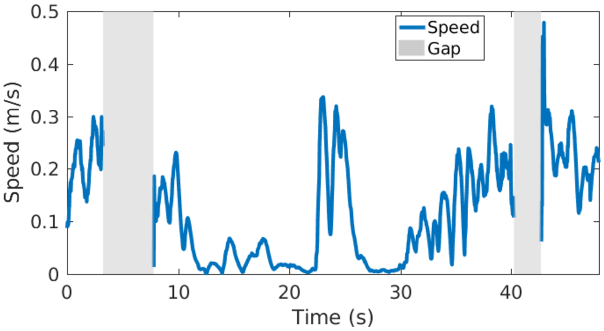


Fig. 6: Norm of the speed calculated from the depth sensor for Person 2 in Fig. 3. Outliers have been removed for speed calculation. Gap periods where there is no valid depth detection are indicated by gray areas. Rapid speed changes can be observed.

An alternative is to use the velocity calculated from the depth stream for gap filling. However, in a confined living room with multiple moving persons, both the speed and direction might change quickly. For instance, using the depth trajectory of Person 2 in Fig. 3, the speed is plotted in Fig. 6. The standard deviation is obtained as 0.1 m/s, which is very high, as compared to the mean speed of 0.12 m/s. As a result, the velocity is un-predictable during gap periods, thus gap filling or trajectory recovery with velocity information obtained from the depth sensor is not reliable. However, this problem is mitigated in our system due to the use of audio information, e.g. via the associated TDOA measurements. We will now evaluate our proposed algorithm on real-room recordings, and analyse the experimental results.

V. Experiments

A. Data Recording Setup

We recorded a dataset[1] for spatial audio production in living room conditions, where each speech signal is an audio object. The data were recorded on a set constructed in a TV/film studio built following professional media production standards. The room had furnitures and a size of $244 \times 396 \times 242$ cm^3, which is very similar to that of a typical living room. As with typical TV/film production sets, its ceiling and one of the walls were missing, though this set was assembled inside a larger room. The reverberation time of this room is about 430 ms. The binaural microphone, i.e. Cortex MK2 as the HATS, was located in the centre of the room with ear height of 165 cm. The depth sensor, i.e. Kinect2, faced Cortex MK2 at the distance of 329 cm just outside the openside wall of the recording room to get a full view, at the height of 170 cm. We could have set the depth sensor overhead or with a much more tilted angle to reduce the amount of occlusions. However, we used a skeletal tracking method that relies on the supervised learning method of [18], which was trained for horizontal view angles (as in the view from a set top box). This Kinect setup was obtained from a pilot test of skeletal tracking before recording the dataset, which is a trade-off between the optimal range that the Kinect SDK can successfully apply skeletal tracking and occlusion issues. The overall setup is illustrated in Fig. 7.

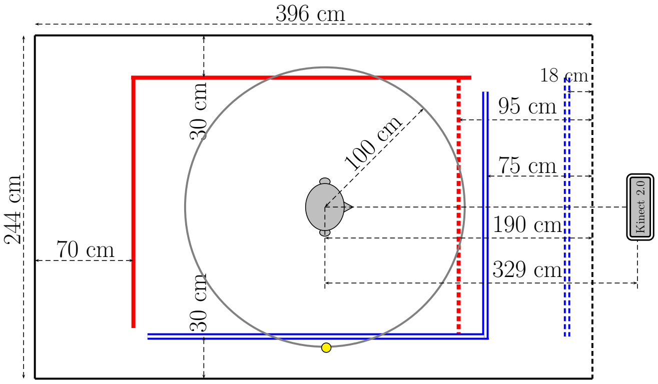


Fig. 7: Setup for data recordings.

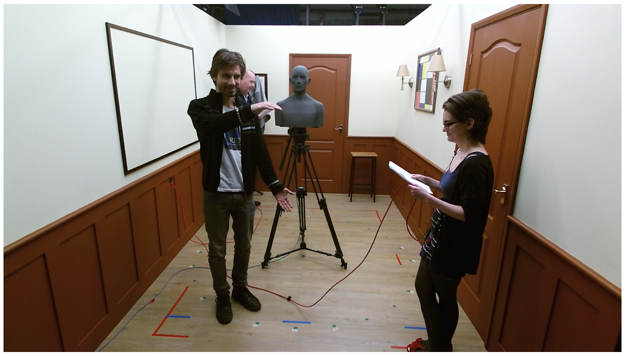
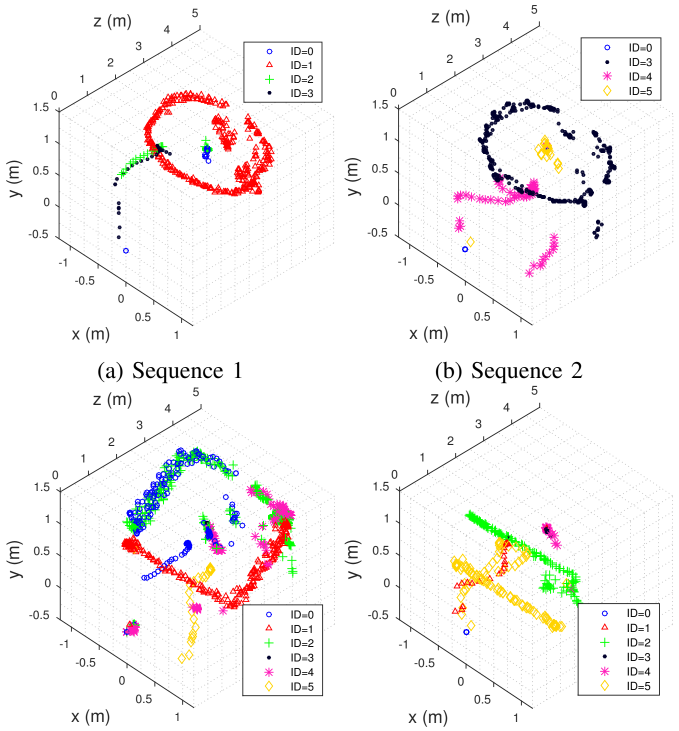


Fig. 8: A sample image in Sequence 3 from Kinect's viewpoint. Actor (only the head can be partially seen) stands on the back corner while Actress stands in front. An additional subject is clapping his hands for synchronisation of the depth and audio streams.

Four sequences were recorded which in total last about 10 minutes, involving two actors denoted as Actor and Actress, with height of about 190 cm and 160 cm respectively. The audio materials contain 120 phonetically-balanced transcripts from the TIMIT corpus, which is a large scale speech database widely used in speech processing research [52]. In Sequence 1, Actor started at the position highlighted by the small circle in Fig. 7, facing the centre (i.e. the HATS), walking slowly, sideways, along the circular trajectory anti-clockwise while reading the audio materials. After completing one circle, he returned clockwise back to the starting point. In Sequence 2, Actress repeated this process with a faster pace. In Sequence 3, Actor walked back and forth along the L-shaped path highlighted with the single solid line. At the same time, Actress walked along the L-shaped curve in the double solid line. They walked independently from each other, both at their preferred pace and facing forward. Fig. 8 shows a sample image at the beginning of Sequence 3 from the viewpoint of the depth sensor. In Sequence 4, Actor walked along the single dashed line while Actress along the double dashed line at their preferred pace, both facing the centre of the room (where the HATS was located) while reading concurrently. The two engaged subjects in our recordings were not walking naturally, as they were restricted by the hardware such as the tripod and

[1] Data underlying the findings are fully available without restriction at http://cvssp.org/data/s3a.



(a) Sequence 1

(b) Sequence 2

(c) Sequence 3

(d) Sequence 4

Fig. 9: Head tracking results from the depth stream for the 4 sequences, down-sampled for better visualisation. Note that there are short trajectories in each sequence, e.g. these crossed and dotted “tails” in Sequence 1, which are caused by a person leaving and entering the scene in the beginning and end of the sequence to synchronise the audio and depth streams with hand claps.

lifted cables while they read and walk concurrently, and they also had to follow fixed trajectories which were designed to facilitate ground-truth labelling.

The depth and audio streams were recorded with different hardware and software, with sampling rates of 30 Hz and 44.1 kHz respectively. We used hand clapping at the beginning and end of each recording session to synchronise these two streams. This was done by an additional subject who entered the scene, standing in front of the depth sensor and clapping his hands. The hand clap is used for synchronisation, which can be detected from both the audio recordings and the depth-based head tracking results.

B. Parameter Setup

1) PHD filter: parameters for the SMC-PHD filter are set empirically as follows. The survival rate was set to $P_S = 0.98$. We also employed an adaptive detection probability P_D, which was set to 0.9 when the depth stream detected any person, and 0.2 when there was no person detected. This avoids the situation when the depth sensor fails to process a frame during consecutive frames. Measurement-centred new-born targets were drawn with a Gaussian distribution whose covariance matrix Σ equals to the identity matrix times 0.02. The birth intensity $\nu = 0.1$ and the particle number per persistent or new-born target was $M_p = M_b = 400$. The measurement likelihood followed $g(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\|\mathbf{z}-\mathbf{x}\|_2|0, 0.02)$. The clutter intensity in Eq. (3) $\kappa = 0.5$ and $c_n = 4$ and $c_f = 2$ in Eq. (4). In Eq. (6), we set $w = 0.3$ and $\delta = 0.2$ m, which is approximately the half shoulder width. The clustering threshold was set $\zeta_0 = 0.5$ m. The above parameters were determined based on a validation subset from Sequence 3.

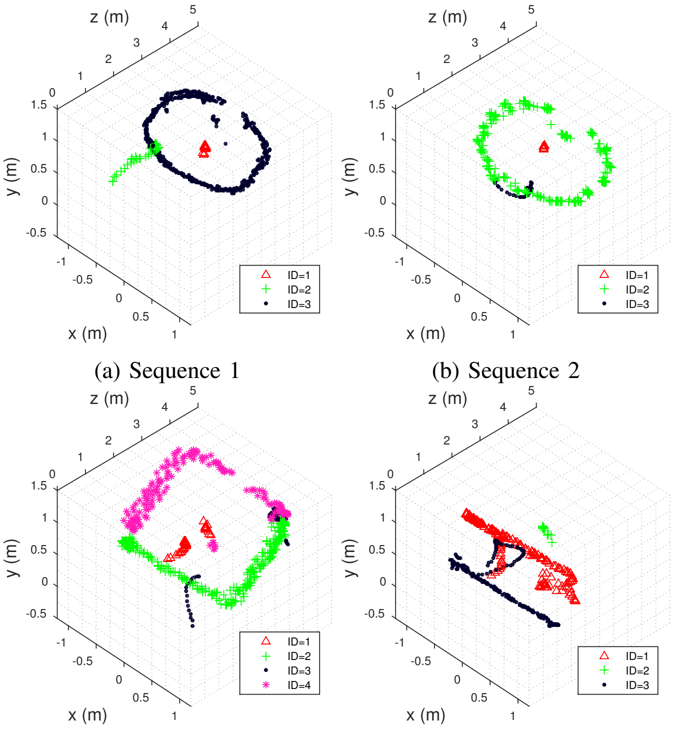
2) Depth-audio fusion: to extract the audio TDOA features, we first applied STFT to the recordings with the window size of 4096 samples overlapped at 2626 samples between neighbouring windows. Thus the audio features were extracted with the same temporal resolution as the depth stream, i.e. 30 Hz. The PHAT-GCC function in Eq. (1) was then employed to each frame and the candidate delays τ were set between $[-1,\ 1]$ ms with a resolution of 0.05 ms. The above candidate delays cover the maximum interaural time difference for the HATS, and their fine resolution supports accurate azimuth localisation. At each frame, we extracted at most two TDOAs as the audio feature. To avoid the influence of high-frequency noise, we modified Eq. (1) by summing over only the voiced frequency band of $[300,\ 3400]$ Hz, instead of the whole frequency band. The extracted TDOA features correspond to bilateral azimuths with Eq. (2). Trained on the BRIR dataset [48], we obtained $p_1 = 9.72 \times 10^{-3}$ and $p_3 = -2.19 \times 10^{-7}$.

C. Results

*1) **PHD filtering**:* in the original depth head tracking results as shown in Fig. 9, a large number of outliers were observed. As a contrast, after applying the proposed PHD filtering method, the majority of these outliers have been filtered out, as shown in Fig. 10. Note that the HATS stands at the centre to mimic a real listener, and it sometimes occludes the other subjects. To mimic a virtual listener and mitigate the introduced occlusions, more compact binaural microphones such as a dummy head with built-in microphones can be used.

Although the outlier problem was greatly mitigated, there might still be some remaining outliers, as can be observed in Sequence 2. These points were grouped to Actress during the period when she was occluded by the HATS. Since these points had a large population and they remained in consistent positions for several seconds, the PHD filtering method did not classify them as outliers, but associated them with the target. This also happened to several points in Sequence 3.

To quantitatively evaluate the PHD filtering method, we established the ground truth by manually labelling each head position as either a valid detection or an outlier as follows. We mapped each 3D head position associated with a target onto the associated depth image. If it fell in the centre of the associated person's head, it was labelled as a valid detection, i.e. true positive (tp); otherwise, it was labelled as an outlier, i.e. false positive (fp). If a frame didn't output a valid detection associated with the engaged person, this frame was labelled as false negative (fn), which included both the conditions when the detection was inaccurate and when the target was mis-detected. We used the precision & recall as evaluation metrics, where precision=$\frac{\#\text{tp}}{\#\text{tp}+\#\text{fp}}$, and recall=$\frac{\#\text{tp}}{\#\text{tp}+\#\text{fn}}$. Note that, since Actor and Actress were engaged throughout these sequences, $\#\text{tp} + \#\text{fn}$ equals the



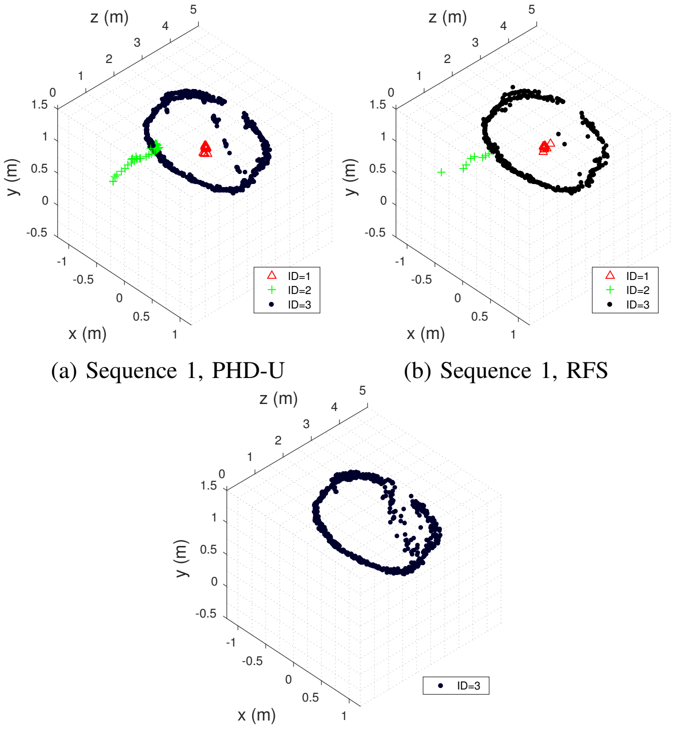
(a) Sequence 1 (b) Sequence 2

(c) Sequence 3 (d) Sequence 4

Fig. 10: Applying the proposed PHD filtering method with the ID association scheme to the head tracking results in Fig. 9. The PHD-filtered results are down-sampled here.

frame number of each sequence. Localisation accuracy such as RMSE in mm was not used, since it is difficult to manually label the 3D position on the depth image without errors up to a few centimetres, while errors $1-7$ mm have already been reported in [31] in the range of 1.2 m to 3.5 m for Kinect skeletal tracking. The performance based on precision & recall is summarised in TABLE I.

Several competing methods were implemented. The first one is the SMC-PHD filter with a uniform clutter intensity $\kappa(\mathbf{z}|\mathbf{Z}) = 0.5$, where the adaptive detection probability is also employed, denoted as “PHD-U”. The second one is a Bayesian random finite set (RFS) filtering method, denoted as “RFS”, with the same principle as used in [4]. However, due to the complex and explicit association involved, its computational complexity becomes extremely high with the increasing number of targets. Since the HATS was also detected as a person by Kinect, we employed “RFS” only for Sequences 1 and 2 where only two targets were detected most of the time. For Sequences 3 and 4 where three targets were detected in most frames, “RFS” lacks the mathematical formulation for the explicit association and thus fails to work. The third one is the Kalman filtering method [51], denoted as “Kalman”. Note that, “Kalman” cannot deal with the complex data association problem in multi-person tracking. As a result, we extracted all the depth measurements related to Actor and Actress in advance, which requires prior information. Moreover, to deal with frames without any associated depth data, we freeze the filtering process and set the output as empty until new depth data coming. The tracking results on Sequence 1 for “PHD-U”, “RFS”, “Kalman” are presented in Fig. 11.



(a) Sequence 1, PHD-U (b) Sequence 1, RFS

(c) Sequence 1, Kalman

Fig. 11: Applying “PHD-U” (a), “RFS” (b) and “Kalman” (c) to Sequence 1. IDs of “RFS” and “Kalman” are manually corrected to be consistent with PHD-filtered results. Results related to each sequence are down-sampled with the same sampling rate as in Fig. 10 (a), compared to which “PHD-U” and “Kalman” have more outliers; “RFS” has fewer valid detections, which can be observed from its sparse pattern. This is further validated by the lower precision for “PHD-U” and “Kalman”, as well as significantly lower recall for “RFS” in TABLE I.

From TABLE I, we first analyse the evaluation results in terms of recall. We notice that the two PHD filtering methods, “proposed” and “PHD-U” greatly increased the recall for all sequences, which means the valid detections after PHD filtering out-numbered that of the original data. This is because we employed an adaptive detection probability P_D, which mitigated the problem when the head tracker failed to process a frame of the depth image and output nothing. “RFS” however suffered from poor performance in terms of recall. In other words, some valid detections in the original depth tracking results were filtered out. This is because the particle weights are dramatically reduced for “RFS” during the frames for which the head tracker fails to process, and it takes several frames for these particle weights to recover to the required level for convergence. “Kalman” obtained a similar recall as compared to the original data, since it cannot estimate mis-detected positions when no depth data is available, neither can it effectively correct outliers due to the lack of the mechanism to detect outliers.

Following we analyse the results in terms of precision. For the two PHD-based algorithms, “PHD-U” and “proposed”, very high precision of almost 1 was obtained for Sequences 1 and 3, which means most of the false positives, i.e. outliers,

TABLE I: Quantitative evaluations of the proposed PHD filtering method in terms of precision & recall.

	Performance	Actor		Actress	
		precision	recall	precision	recall
Seq 1	original	0.9612	0.8597	N/A	
	proposed	0.9976	0.9130		
	PHD-U	0.9945	0.9134		
	RFS	0.9950	0.6746		
	Kalman	0.9591	0.8573		
Seq 2	original	N/A		0.9166	0.7891
	proposed			0.9360	0.8407
	PHD-U			0.9269	0.8410
	RFS			0.9341	0.5621
	Kalman			0.9273	0.7979
Seq 3	original	0.9848	0.8243	0.9998	0.9123
	proposed	0.9969	0.8657	1.0000	0.9690
	PHD-U	0.9967	0.8660	1.0000	0.9688
	Kalman	0.9824	0.8219	1.0000	0.9118
Seq 4	original	1.0000	0.7147	0.9304	0.8111
	proposed	0.9660	0.7737	0.9318	0.8599
	PHD-U	0.9653	0.7737	0.9307	0.8605
	Kalman	1.0000	0.7141	0.9826	0.8003

TABLE II: Outlier evaluations as a percentage for each sequence.

Evaluation in percentage		Seq 1	Seq 2	Seq 3	Seq 4	Avg
$\frac{\#\text{outliers}}{\#\text{detections}}$	original	2.17	4.77	1.55	0.28	2.19
	proposed	0.12	3.15	0.30	0.17	0.94
	PHD-U	0.28	4.13	0.43	0.17	1.25
$\frac{\#\text{outliers}}{\#\text{frames}}$	original	3.98	8.96	4.25	0.62	4.45
	proposed	0.22	5.81	0.85	0.40	1.82
	PHD-U	0.51	7.69	1.20	0.40	2.45

were removed from these sequences. In Sequence 2, the above two methods converged to a group of inaccurate detections lasting several seconds as mentioned earlier, which resulted in a relatively low precision. However in Sequence 4, the precision of Actor is worse after the PHD filtering. This doesn't mean the PHD filtering methods have converged to some irrelevant data points and thus introduced new clutters. This was caused by another subject (the rectangle "tail" in Fig. 10) being mis-identified as Actor. This also happened towards the end when this subject was mis-identified as Actress (the dotted "tail" in Fig. 10), which resulted in the relatively low precision. Ignoring mis-identifications, the precision for Actor and Actress is almost 1 for both PHD filtering methods in Sequence 4. "RFS" gained slightly higher precision than PHD filtering, however at the cost of very low recall. "Kalman" still does not show an advantage over the original head tracker for most conditions. Since "Kalman" employs a linear Gaussian time state model, where the linear motion model does not consider the chance for a data point being a clutter, thus "Kalman" will gradually converge to outlying positions where outliers emerge during a continuous time slot.

We notice that "proposed" and "PHD-U" outperform the other filtering methods for both recall and precision. They are both PHD-based approaches, with the only difference lies on the clutter intensity, where measurement-driven and uniform clutter intensities were used respectively. The "proposed" PHD modification gives a similar recall to the one by "PHD-U", i.e. same numbers of valid detections were retained. However, "proposed" performs better than "PHD-U" in terms of the measure of precision, which increases from 96.90% to 97.14% on average. The improvement is relatively small for the following two reasons. Firstly, the number of outliers related to the two targets is not significant, which is reflected by the very high precision in the original data. Secondly, outliers related to the other engaged subjects were not included yet. Considering each sequence as a whole and ignoring mis-identifications, we calculated the total number of outliers and compared it over (1) the total number of detections and (2) the total frame number in TABLE II. Particularly, the total number of outliers over the total frame number, i.e. the outlier rate, is a direct quantification of the chance of each frame being affected by occlusions.

From TABLE II, it can be seen that "proposed" outperforms "PHD-U" in terms of outlier rate, which reduces from 2.45% to 1.82% on average. To analyse whether the improvement is statistically significant, we performed a one-way ANOVA test [53]. If the resultant p-value is less than a threshold (i.e. the significance level, which is often set as 0.05), it rejects the null hypothesis that different groups are drawn from the same distribution. For our specific case, the null hypothesis that the two algorithms have no performance difference is rejected if $p < 0.05$, i.e. statistically significant results are therefore justified (either positive or negative). We repeated "proposed" and "PHD-U" 50 times respectively, with different initialisations and random particle re-sampling and creation processes. Comparing "proposed" and "PHD-U" using the average outlier rate between sequences, a p-value of 0.00 was obtained, suggesting that our proposed method outperforms "PHD-U" with statistical significance.

In summary, our proposed filtering method shows advantages over the competing methods for the following reasons. Firstly, our filtering method employed the "PHD" framework, which propagates the first-order statistical moment of the multiple target posterior, and therefore avoids the complex data association problem in multiple target scenarios. As a contrast, "RFS" cannot cope with more than two target situations; while prior information is required to apply "Kalman" to associate the observed data to each target. In addition our proposed method used an adaptive detection probability to deal with the situation when the depth sensor fails to process a frame during consecutive frames. "RFS" does not have a mechanism to maintain the weights of particles associated with targets in such a condition and needs several frames to recover, thus suffers from a low recall. Secondly, the novel clutter intensity that takes into account the Kinect's limited FOV as well as occlusions were used. This provides the opportunity to classify observations constrained by sensor limitations as outliers. Also the filtering method is less likely to converge to a person being occluded due to the higher clutter intensity exhibited if the person falls in the occluded area. "PHD-U" employed a uniform clutter intensity, which cannot effectively reduce the weights of the particles related to outliers, especially in the occluded period where groups of outliers exist and still show temporal-spatial continuity. As a result, "PHD-U" is more likely to converge to persistent outliers as compared to the proposed method.

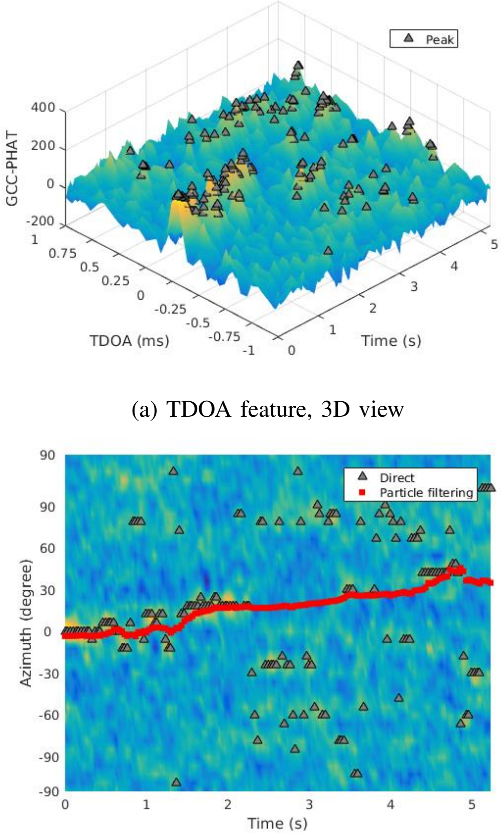
*2) **Depth-audio fusion**:* the depth stream sometimes fails to detect the target, and thus the depth detection related to each target is composed of several segments (see, e.g. Fig. 10(b)). To fill in a gap between two adjacent segments where a target was mis-detected, audio features extracted from the simultaneous audio stream were integrated.

The 1D audio azimuth of the active speaker w.r.t. the HATS yields the audio TDOA feature, as described in Eq. (2). The TDOA features are affected by interference from both competing speakers and background noise, as well as the nonstationarity of speech signals, thus they consist of outliers. If we directly employ the inverse process of Eq. (2) to TDOA features, 1D audio azimuths are obtained with noise. Tested on Sequence 1 that contains one active speaker, approximately 51% frames have the TDOA features yielded by azimuths within $\pm 5°$ deviation from the groundtruth, and 85% frames fall in the range of $\pm 10°$ deviation. Heavier noise can be observed in more challenging multiple speaker scenarios, as shown in Fig. 12. To mitigate the noise in audio cues, particle filtering as mentioned in Section. IV-D is applied to the TDOA features.

The 1D audio azimuths can be further mapped to the 3D positions as introduced in Section IV-D. Extra constraints learned from the depth stream that trajectories of each person lie on a plane, as discussed in Section IV-B, are enforced in the mapping process. Each plane can be represented by a normal vector plus a scalar in Eq. (7), which can be iteratively calculated by data clustering and parameter estimation. Take Sequence 1 after the PHD filtering as an example, the top 98% of points that best fit the converged plane associated with Actor have an average fitting error of 1.25 cm, which also confirms that the plane assumption is valid. When multiple targets are involved, the planes related to the engaged targets are assumed to be parallel to each other. We need to stress that the audio stream and the depth stream might not be spatially aligned, as shown in Fig. 13. In such conditions, we need to tune one modality before applying the fusion to mitigate this difference.

To rectify erroneous IDs, we have applied an ID association scheme (top right corner in Fig. 1) [46], which can be applied to the PHD-filtered results directly. However, there were still some remaining ID errors, as shown in Sequences 3 and 4 in Fig. 10, where the entering subject (tail) was mis-identified as one of the other subjects. Before applying the gap filling, we manually corrected the remaining ID errors, such that the segments enclosing each gap are associated with the same speaker. The final tracking results after integrating the audio cues are shown in Fig. 14. It can be observed that the estimated positions successfully filled the gap between adjacent depth segments. Moreover, outliers that did not fit onto the target plane were removed.

To better illustrate the gap filling in the temporal domain, we also plotted audio azimuths of the engaged targets relative to the HATS over time in Fig. 15, where targets have consistent colours as in Fig. 14. Each azimuth trajectory is composed by different segments and gaps can be observed between them. During the above gap periods, audio azimuths were calculated via the proposed gap filling method, and these audio azimuths matched very well with the designed trajectory.



(a) TDOA feature, 3D view

(b) Audio azimuth, 2D view

Fig. 12: TDOA features related to the peaks in the GCC-PHAT function are extracted from the binaural audio recordings during a gap period in Sequence 3. The first peak in each frame is highlighted in triangles at the top plot. These TDOA features can be directly converted to audio azimuths (highlighted in triangles in the bottom plot), however exhibiting heavy noise. The converted audio azimuths are wrapped between $-90°$ and $90°$. By applying particle filtering to the TDOA features, the noise is mitigated, and more accurate audio azimuths are obtained, as shown by the red dots.

We then quantitatively evaluated the tracking results with precision & recall, as shown in TABLE III. The ground truth during gap periods was manually established in the same way as the one used for PHD evaluations in Section V-C1, by mapping 3D positions from the associated depth images. The precision was almost 100% for most sequences. This is because the outliers that did not fit onto the plane were further rectified. Moreover, the new detections from the audio stream accurately tracked the mis-detected target, resulting in a very high recall, as compared to the PHD-filtered results in TABLE I. Take Actor in Sequence 3 as an example. Previously, of the 4491 frames, there were 3888 valid detections and 12 outliers associated with Actor as well as 591 frames where Actor was mis-detected. After gap filling, the 12 outliers were further rectified, and extra 494 new detections successfully tracked the target from the above mis-detected frames. Thus both the precision and recall were greatly improved. The recall improvement is more significant for Actress in Sequence 4, where it can be observed in Fig. 15 that a long gap exists in

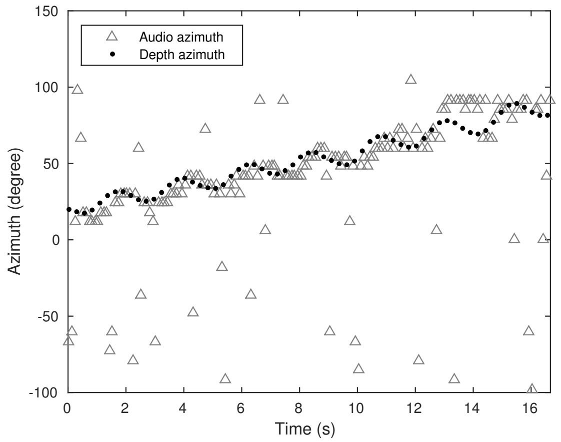
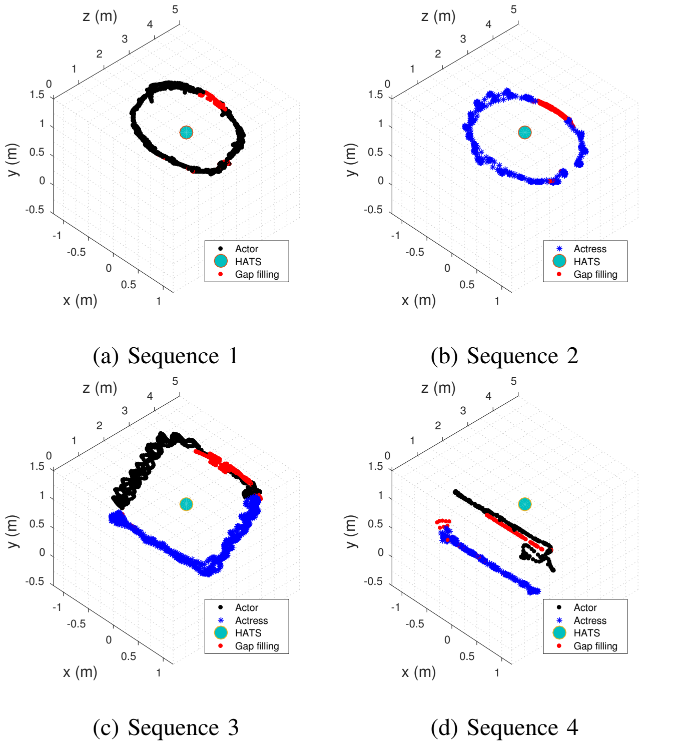


Fig. 13: The HATS might be tilted and thus not synchronised with the Kinect coordinate system. The audio azimuth is lower than the depth azimuth, which means the HATS is tilted to the right ear direction. As a reference, we calculated the HATS rotation from Kinect SDK, and obtained the head orientation of 1°, which is consistent with our estimated angle shift 0.45°. The audio and depth azimuths are down-sampled here for better visualisation.

TABLE III: Precision & recall after integrating the audio cues.

	Seq 1	Seq 2	Seq 3		Seq 4	
	Actor	Actress	Actor	Actress	Actor	Actress
Precision	1.0	1.0	1.0	0.9936	0.9892	0.9865
Recall	0.9781	0.9391	0.9784	0.9722	0.9348	0.9138



(a) Sequence 1

(b) Sequence 2

(c) Sequence 3

(d) Sequence 4

Fig. 14: Depth-audio fusion applied to the PHD-filtered results in Fig. 10, to fill the gaps between valid depth detections using the audio stream and the planar constraint. The data points are down-sampled for illustration purposes. Trajectories associated with Actor and Actress are plotted, and filled detections using the audio stream are highlighted in red dots. The HATS is highlighted in the centre with a cyan circle. Actor and Actress are shown with consistent patterns between different sequences.

the beginning for Actor. During the gap period particle filtering tracked 300 (lasting in total 10 s) frames, of which 292 frames have valid detections.

VI. Conclusions

We have presented a multi-person tracking system for object-based spatial audio production, combining the binaural audio recordings and the concurrent depth stream. The depth-based head tracker gives positions in 3D. Yet, it suffers from outliers and mis-detections, which are often introduced by occlusions in multi-person scenarios. To remove outliers, we introduced a modified PHD filtering method with adaptive clutter intensity. To mitigate mis-detections when the depth stream fails to track a person, the binaural recordings originally collected for spatial audio evaluations are utilised. Applied to real room recordings, we have compared our proposed PHD filtering method with several other baseline filtering methods. Quantitative evaluations in terms of precision and recall show our proposed method can effectively remove outliers, and the integration of extra audio information successfully compensates mis-detections; showing advantages over depth-only tracking, particularly when there are multiple people with a significant amount of occlusions.

As future work, improvements may be obtained by incorporating more advanced motion models, e.g., taking periodicity in people's trajectories. Insights from Yan et al. [54] may be helpful if batch processing is allowed. Tracklets can be combined using a shortest path search approach, though Yan et al.'s method will have to be adapted for multiple target tracking. In addition, the proposed method may be applied to more challenging scenarios, for example, involving more subjects and occlusions, and compared with other state-of-the-art methods.

Acknowledgements

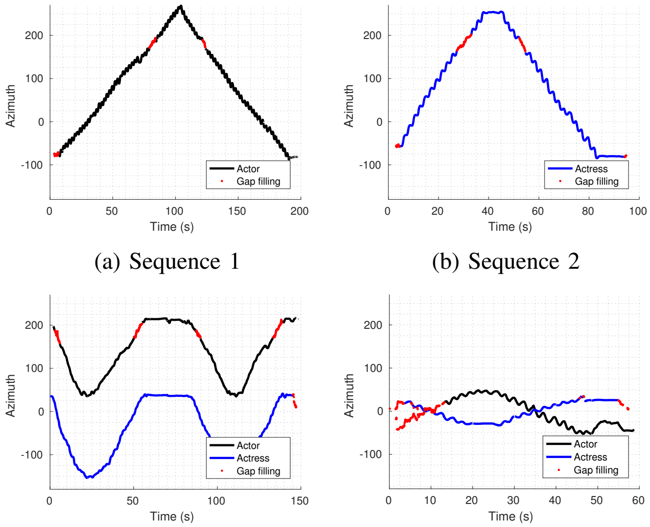
Thanks to Mark Barnard, Luca Remaggi and Philip Coleman for the help with the acquisition of the dataset. We would like to thank the Associate Editor and the anonymous reviewers for their contributions to improving the quality of the paper. We would like to acknowledge the support of the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

References

[1] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proc. IEEE*, vol. 101, no. 9, pp. 1920–1938, Sept. 2013.

[2] J. Francombe, T. Brookes, R. Mason, R. Flindt, P. Coleman, Q. Liu, and P. Jackson, "Production and reproduction of program material for a variety of spatial audio formats," in *Proc. 138th Audio Eng. Soc. Conv.*, May 2015.

[3] P. Coleman, A. Franck, J. Francombe, Q. Liu, T. de Campos, R. J. Hughes, D. Menzies, M. F. Simón Gálvez, Y. Tang, J. Woodcock, P. J. B. Jackson, F. Melchior, C. Pike, F. M. Fazi, T. J. Cox, and A. Hilton, "An audio-visual system for object-based audio: From recording to listening," *IEEE Trans. Multimedia*, submitted for publication.



(a) Sequence 1

(b) Sequence 2

(c) Sequence 3

(d) Sequence 4

Fig. 15: The azimuth of Actor and Actress relative to the HATS. The filled gaps integrating information from the audio stream are highlighted in red. To show the spatial consistency of the tracking results, some azimuths in (-90, -180] are mapped to the range of [180, 270).

[4] 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, 2006.

[5] "PhaseSpace: Impulse motion capture," Online, Retrieved in Dec. 2015, http://www.phasespace.com/.

[6] Y. Raja, S. J. McKenna, and S. Gong, "Tracking and segmenting people in varying lighting conditions using colour," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recog.*, 1998, pp. 228–233.

[7] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Comput. Vis. Image Underst.*, vol. 80, no. 1, pp. 42–56, 2000.

[8] S. Zhang, X. Yu, Y. Sui, S. Zhao, and L. Zhang, "Object tracking with multi-view support vector machines," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 265–278, Mar. 2015.

[9] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 5, 2001, pp. 3021–3024.

[10] E. A. Lehmann and R. C. Williamson, "Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments," *EURASIP J. Adv. Signal Process.*, no. 1, pp. 1–9, 2006.

[11] 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.

[12] U. Varshney, "Pervasive healthcare and wireless health monitoring," *Mobile Netw. Appl.*, vol. 12, no. 2-3, pp. 113–127, Mar. 2007.

[13] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Comput. Vis. Image Underst.*, vol. 106, no. 23, pp. 288–299, 2007.

[14] S. J. Krotosky and M. M. Trivedi, "Mutual information based registration of multimodal stereo videos for person tracking," *Comput. Vis. Image Underst.*, vol. 106, no. 23, pp. 270–287, 2007.

[15] K. O. Arras, S. Grzonka, M. Luber, and W. Burgard, "Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities," in *Proc. IEEE Int. Conf. Robot Autom.*, May 2008, pp. 1710–1715.

[16] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[17] Microsoft, "Kinect for Windows," Online, Retrieved in Sept. 2015, https://dev.windows.com/en-us/kinect/.

[18] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2011.

[19] M. Ye, X. Wang, R. Yang, L. Ren, and M. Pollefeys, "Accurate 3D pose estimation from a single depth image," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 731–738.

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

[21] 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, no. 2, pp. 601–616, Feb. 2007.

[22] A. O'Donovan, R. Duraiswami, and J. Neumann, "Microphone arrays as generalized cameras for integrated audio visual processing," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, June 2007, pp. 1–8.

[23] M. Barnard, P. Koniusz, W. Wang, J. Kittler, S. M. Naqvi, and J. Chambers, "Robust multi-speaker tracking via dictionary learning and identity modeling," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 864–880, Apr. 2014.

[24] 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.

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

[26] M. P. Michalowski and R. Simmons, "Multimodal person tracking and attention classification," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction*, Mar. 2006, pp. 347–348.

[27] M. Luber, L. Spinello, and K. O. Arras, "People tracking in RGB-D data with on-line boosted target models," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sept. 2011, pp. 3844–3849.

[28] W. Choi, C. Pantofaru, and S. Savarese, "Detecting and tracking people using an RGB-D camera via multiple detector fusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2001.

[29] L. S. K. O. Arras, "People detection in RGB-D data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sept. 2011, pp. 3838–3843.

[30] A. Mogelmose, C. Bahnsen, T. B. Moeslund, A. Clapes, and S. Escalera, "Tri-modal person re-identification with RGB, depth and thermal features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern. Recog. Workshops*, June 2013, pp. 301–307.

[31] M. A. Livingston, J. Sebastian, Z. Ai, and J. W. Decker, "Performance measurements for the Microsoft Kinect skeleton," in *Proc. IEEE Virtual Reality Short Papers and Posters*, Mar. 2012, pp. 119–120.

[32] R. P. S. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, Oct. 2003.

[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] B. Ristic, D. Clark, B.-N. Vo, and B.-T. Vo, "Adaptive target birth intensity for PHD and CPHD filters," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 48, no. 2, pp. 1656–1668, 2012.

[35] 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.

[36] "More about kinect sensor placement," https://support.xbox.com/en-US/xbox-360/accessories/sensor-placement, accessed: 2016-11-25.

[37] R. Levorato and E. Pagello, "Probabilistic 2d acoustic source localization using direction of arrivals in robot sensor networks," in *Proc. Int. Conf. Simulation, Modeling, Programming Autonomous Robots*, 2014, pp. 474–485.

[38] C. Morato, K. N. Kaipa, B. Zhao, and S. K. Gupta, "Toward safe human robot collaboration by using multiple kinects based real-time human tracking," *J. Computing Inf. Sci. Eng.*, vol. 14, Jan. 2014.

[39] L. Sevrin, N. Noury, N. Abouchi, F. Jumel, B. Massot, and J. Saraydaryan, "Preliminary results on algorithms for multi-kinect trajectory fusion in a living lab," *IRBM*, vol. 36, no. 6, pp. 361–366, 2015.

[40] "Skeletal tracking with multiple kinect sensors," https://social.msdn.microsoft.com/Forums/en-US/0ea87ba7-f958-4970-9fec-8718a2d87768/skeletal-tracking-with-multiple-kinect-sensors?forum=kinectv2sdk, accessed: 2016-11-25.

[41] "Skeleton tracking with multiple kinect sensors," https://msdn.microsoft.com/en-us/library/dn188677.aspx, accessed: 2016-11-25.

[42] Q. Liu, T. de Campos, W. Wang, P. Jackson, and A. Hilton, "Person tracking using audio and depth cues," in *Proc. ICCV Workshop 3D Reconstruction Underst. Video Sound*, Dec. 2015.

[43] W. Qu, D. Schonfeld, and M. Mohamed, "Real-time distributed multi-object tracking using multiple interactive trackers and a magnetic-inertia potential model," *IEEE Trans. Multimedia*, vol. 9, no. 3, pp. 511–519, Apr. 2007.

[44] H. Jiang, S. Fels, and J. J. Little, "Optimizing multiple object tracking and best view video synthesis," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 997–1012, Oct. 2008.

[45] F. Lian, C. Han, and W. Liu, "Estimating unknown clutter intensity for PHD filter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 4, pp. 2066–2078, Oct. 2010.

[46] Q. Liu, T. de Campos, W. Wang, and A. Hilton, "Identity association using PHD filters in multiple head tracking with depth sensors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016.

[47] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[48] C. Hummersone, "Binaural room impulse response measurements," Online, Aug. 2011, http://iosr.surrey.ac.uk/software/#BRIRs.

[49] D. Satongar, Y. W. Lam, and C. Pike, "Measurement and analysis of a spatially sampled binaural room impulse response dataset," in *Proc. Int. Congr. Sound Vibration*, July 2014, http://www.bbc.co.uk/rd/publications/sbsbrir.

[50] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 99, 2017.

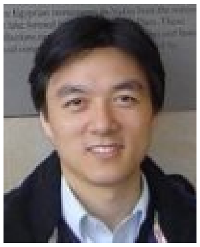
[51] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME–J. Basic Eng.*, vol. 82, pp. 35–45, 1960.

[52] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Tech. Rep. N*, vol. 93, Feb. 1993, https://catalog.ldc.upenn.edu/ldc93s1.

[53] P. G. Hoel, *Elementary Statistics*. New York: Wiley, 1976.

[54] F. Yan, W. Christmas, and J. Kittler, "Layered data association using graph-theoretic formulation with application to tennis ball tracking in monocular sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1814–1830, Oct. 2008.

Qingju Liu received the B.Sc. degree in electronic information engineering from Shandong University, Jinan, China in 2008, and the Ph.D. degree in signal processing in 2013 from the Centre for Vision, Speech and Signal Processing (CVSSP) in University of Surrey, Guildford, U.K. Since October 2013, she has been working as a research fellow in CVSSP. Her current research interests include audio-visual signal processing, spatial audio, neural networks and machine learning.



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

He then worked in King's College London, Cardiff University, Tao Group Ltd. (now Antix Labs Ltd.), and Creative Labs, before joining University of Surrey in 2007, where he is currently a Reader in Signal Processing, and a Co-Director of the Machine Audition Lab.

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. He has (co)-authored over 200 publications in these areas. He is currently an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Teófilo de Campos completed his doctorate at the University of Oxford in 2006, on 3D hand tracking. In 2001 he completed his masters on Feature Selection at the U. of São Paulo. He has worked in the research laboratories of Sharp, Microsoft and Xerox. From 2009 to 2016 he worked at the University of Surrey on action recognition, transfer learning, anomaly detection, head tracking and semantic segmentation. He was also a researcher in the Machine Learning group at Sheffield from 2013-2014. Currently, he is a Professor Adjunto at the University of Brasilia.

Philip J. B. Jackson is Senior Lecturer in machine audition at the Centre for Vision, Speech & Signal Processing (University of Surrey, UK) with MA in Engineering (Cambridge University, UK) and PhD in Electronic Engineering (University of Southampton, UK). His broad interests in acoustical signal processing provide research contributions in speech production, auditory processing and recognition, audio-visual machine learning, blind source separation, articulatory modelling, visual speech synthesis, spatial audio recording, reproduction and quality evaluation, and sound field control [http://bit.ly/2oTRw1C]. He leads research on object-based production in the S3A project on spatial audio, and enjoys life in sound.

Adrian Hilton received the B.S. (Hons.) and D.Phil. degrees from University of Sussex, Sussex, U.K., in 1988 and 1992, respectively. He is a currently a Professor of Computer Vision and Graphics and Director of the Centre for Vision, Speech, and Signal Processing at the University of Surrey, Surrey, U.K. His research interests include robust computer vision to model and understand real world scenes. Contributions include technologies for the first handheld 3-D scanner, modelling of people from images and 3-D video for games, broadcast, and film production. He currently leads research investigating the use of computer vision for applications in entertainment content production, visual interaction, and clinical analysis.