Audio-visual tracking of concurrent speakers

Xinyuan Qian, Alessio Brutti, Oswald Lanz, Maurizio Omologo, Andrea Cavallaro

***Abstract*—Audio-visual tracking of an unknown number of concurrent speakers in 3D is a challenging task, especially when sound and video are collected with a compact sensing platform. In this paper, we propose a tracker that builds on generative and discriminative audio-visual likelihood models formulated in a particle filtering framework. We localize multiple concurrent speakers with a de-emphasized acoustic map assisted by the image detection-derived 3D video observations. The 3D multi-modal observations are either assigned to existing tracks for discriminative likelihood computation or used to initialize new tracks. The generative likelihoods rely on color distribution of the target and the de-emphasized acoustic map. Experiments on AV16.3 and CAV3D datasets show that the proposed tracker outperforms the uni-modal trackers and the state-of-the-art approaches both in 3D and on the image plane.**

***Index Terms*—Audio-visual fusion, 3D multiple target tracking, particle filter, concurrent speakers.**

I. Introduction

Tracking multiple targets that may produce sound is important to support sound source separation [1–3], speaker diarization [4], and speech enhancement [5]. The location of a target can be inferred by acoustic cues, estimated from Sound Source Localization (SSL) methods, which often rely on Direction of Arrival (DoA) or Time Difference of Arrival (TDoA) estimates when the geometry of the microphone array is known [6, 7]. However, acoustic cues may be inaccurate because of background noise and reverberation. Moreover, even a short overlap of sounds may disrupt source localization [8, 9]. Similarly, visual cues are less discriminative because of clutter and illumination changes. Finally, the presence of multiple targets whose number may vary makes both audio and video tracking tasks difficult.

Audio-visual fusion may help overcome the limitations of uni-modal tracking [10–24]. In audio-visual tracking, audio usually supports a visual tracker when targets are occluded or outside the Field-of-View (FoV) of the camera [10, 11, 17, 22, 25]. However, these methods are generally sensitive to audio outliers, especially in case with multiple concurrent speakers. Moreover, effectively dealing with the inter-dependencies of multi-modal observations is an unsolved problem in Multi-Target Tracking (MTT). AV3T is an audio-visual multi-speaker tracker based on a Particle Filter (PF) framework [24] that uses signals from a single RGB camera co-located with a small circular microphone array for MTT in 3D. AV3T derives 3D video observations from face detections to support tracking without using a 3D sensor [26] or stereo vision [27]. However, AV3T needs to know the number of targets in the scene.

In this paper, we present Audio-Visual 3D Tracking with Video-assisted De-emphasized Maps (AV3T-VDM), an extension of AV3T to track an unknown number of speakers. We propose interactive audio and video generative and discriminative likelihoods to compensate for the respective uni-modal weakness. AV3T-VDM enhances the data association process with spatio-temporal consistency and audio-visual features, to limit identity switches when targets overlap on the image plane. We use 3D multi-modal observations to initialize, de-activate and re-activate tracks. To the best of our knowledge, AV3T-VDM is the first multi-target tracking framework that can deal with 3D tracking of an unknown number of concurrent speakers with a compact audio-visual sensing platform.

II. Background

Audio-visual trackers can be characterized in terms of tracking space, number of targets, sensor configurations, audio-visual features, and likelihood models, data association and fusion methods. Tab. I summarizes the most representative works.

Tracking can take place on the image plane [10–15], a ground plane [16, 17], or in 3D [18–24]. Multi-modal estimates should be processed in a common space with known camera-microphone calibration information. Compared to Single Target Tracking (STT), MTT requires observation-to-track assignments, which are challenging with multi-modal signals. Audio-visual works cover both STT [12, 17–23] and MTT [10, 11, 13–16, 24].

The likelihood models used for tracking can be categorized as *generative* or *discriminative* [28, 29]. The *generative* likelihoods build adaptive models for target description in specified feature space to find the best matching target candidate [30]. For example, *generative visual likelihoods* [10–14, 21–24] learn the target appearance representation (*e.g.* color histogram) to search for the most similar image sub-region. The corresponding target representations are then updated online while suppressing the background. However, the performance degrades in visually cluttered scene [30]. For example, MeanShift models a target with non-parametric distributions of the color features and locates the target with mode shifts [31]. MeanShift can be wrongly attracted by background regions with similar color distributions. In audio processing, *generative audio likelihoods* rely on the derived spatial acoustic map, where the target position estimate eventually converges to the spatial location with the highest probability of including a sound source [19, 21–24]. However, an acoustic generative tracker may converge to a nearby local

X. Qian and A. Cavallaro are with Centre of Intelligent Sensing (CIS) at Queen Mary University of London (QMUL), London, E1 4NS, UK. (e-mail: x.qian@qmul.ac.uk; a.cavallaro@qmul.ac.uk).

A. Brutti and O. Lanz are with Fondazione Bruno Kessler (FBK), Trento, Italy (e-mail: brutti@fbk.eu; lanz@fbk.eu)

M. Omologo fully contributed to this work while he was with FBK (email: omologo@fbk.eu). He is now with Alexa Machine Learning, Amazon, Italy and USA (email: omologo@amazon.com).

TABLE I
Summary of the state-of-the-art (SoA) audio-visual tracking methods. Key – MS: Multiple Speakers; UNS: Unknown Number of Speakers; LM: Likelihood Model (D: Discriminative; G: Generative); SSM: Sam Sparse Mean; GCC-PHAT: Generalized Cross Correlation with Phase Transfrom; ILD: Interaural Level Difference; IPD: Interaural Phase Difference; DP-RTF: Direct-Path Relative Transfer Function; RTF: Relative Transfer Function; $\mathcal{H}$: histogram; det.: detection; loc.: location; CAMShift: Continuously Adaptive MeanShift; WPDA: Weighted Probabilistic Data Association; EM: Expectation–Maximization; NN: Nearest Neighbor; TDoA: Time Difference of Arrival; PF: Particle Filter; PHD: Probability Hypothesis Density Filter; GM: Graphical Model; KF: Kalman Filter; PSO: Particle Swarm Optimization; NA: information Not Available; -: information not applicable.

Ref	Space	MS	UNS	Audio		Video		Data Association		Fusion
				Feature	LM	Feature	LM	Method	Feature	
[10]	Img	✓		SSM	D	$\mathcal{H}_{HSV}$	G	NA	NA	PF
[11]	Img	✓	✓	SSM	D	$\mathcal{H}_{HSV}$	G	NA	NA	PHD
[12]	Img			GCC-PHAT	D	$\mathcal{H}_{RGB}$, change det.	D*G	WPDA	$\mathcal{H}_{RGB}$, change det, TDoA	PF
[13]	Img	✓	✓	ILD, IPD	D	$\mathcal{H}_{RGB}$, person det.	D*G	EM	target image loc., $\mathcal{H}_{RGB}$, IPD	GM+KF
[14]	Img	✓	✓	DP-RTF	D	face det. and embedding	D*G	EM	target image loc., DP-RTF	EM
[15]	Img	✓	✓	RTF	D	upper-body tracker	D*G	NN	target image loc.	GM
[16]	Grd	✓		GCC-PHAT	D	$\mathcal{H}_{NA}$ (MeanShift)	D	NN	target image loc.	KF
[17]	Grd			GCC	D	person det.	D	-	-	KF
[18]	3D			GCC-PHAT	D	$\mathcal{H}_{HSV}$ (CAMShift)	D	-	-	PSO
[19]	3D			GCC-PHAT	G	motion, face, upper-body det.	D	-	-	PF
[20]	3D			GCC	D	face det.	D	-	-	PF
[21]	3D			GCC-PHAT	G	$\mathcal{H}_{RGB}$	G	-	-	PF
[22]	3D			GCC-PHAT	G	$\mathcal{H}_{RGB}$, face det.	D+G	-	-	PF
[23]	3D			GCC-PHAT	G	$\mathcal{H}_{RGB}$	G	-	-	PF
[24]	3D	✓		GCC-PHAT	G	$\mathcal{H}_{HSV}$, face det.	D/G	NN	target image loc.	PF
Prop.	3D	✓	✓	GCC-PHAT	D*G	$\mathcal{H}_{HSV}$, face det.	D*G	WPDA	target 3D loc., time, $\mathcal{H}_{HSV}$	PF

maximum instead of the global maximum in the acoustic map. Furthermore, a coherent noise source can wrongly attract the tracker.

Discriminative likelihoods rely on a decision (*e.g.* an object localization result) that separates the target from the background [32]. For example, trackers with *discriminative visual likelihoods* consider tracking as a binary classification problem which locates the targets by maximizing the foreground and background difference [33–35]. Among those trackers, tracking-by-detection is the most widespread strategy, where outputs from any object detectors/descriptors (*e.g.* either based on hand-crafted features [36] or deep neural networks [26, 37, 38]) are fed as inputs to the tracking algorithm to retrieve trajectories [15–19]. Works in [39, 40] have exploited Convolutional Neural Network (CNN) to produce image object detections. Discriminative trackers can be designed to be robust to target occlusions and pose variations [41]. However, the tracking performance depends on the quality of the object detector results. For audio, the *discriminative likelihoods* rely on a sound source localizer which plays the role of an object detector, where spatial acoustic observations are used as the discriminative information to be fed into the tracking pipeline [10–18, 20]. Discriminative acoustic trackers are less exposed than generative ones to local maxima that may appear in the acoustic map, but are affected by strong noise sources.

In MTT, *data association* is a significant stage to make observation-to-track assignments. The association strategies can be classified into three groups [42, 43]: (i) one-to-one observation-to-track assignment at each time *e.g.* with a Nearest Neighbor (NN) strategy; (ii) multiple-to-one observation-to-track assignment at each time *e.g.* Weighted Probabilistic Data Association (WPDA), and (iii) retrospect for the best decision at a later time *e.g.* Multiple Hypothesis Tracking (MHT). Among the three, NN, which assigns the most closest observation to the track, is the most computational efficient strategy [44]. However, it may produce over-confident estimates due to the integration of False Positive (FP) observations. The WPDA, which is well suited for multi-modal signals, assigns multiple observations to a single track, weighted by their probabilities [17]. The MHT considers multiple association hypotheses over several past frames, and has the highest complexity as the number of hypotheses grows exponentially with time and cubically with the target number [45].

The number of tracks varies according to the target number. Thus, track initialization (birth) and de-activation (death) are required. Strategies usually rely on the temporal persistence of new observations and tracks, such as [13, 46, 47]. Within a given time interval, a new track is initialized if consistent un-associated observations appear in a nearby region. Existing tracks that are not assigned with observations for a while will be de-activated. However, modality absence or False Negative (FN) observation may lead to incorrect de-activations.

For audio-visual *fusion*, Bayesian inference frameworks may be used to estimate target states from noisy multi-modal observations. Kalman Filter (KF) [13, 16, 17] is the optimal solution under the assumptions of Gaussian noise and linear state functions. The PF [10, 12, 19–24] and Particle Swarm Optimization (PSO) [18] approximate the target state for a non-linear and non-Gaussian problem with particles (target hypotheses). KF, PF and PSO require the prior knowledge of the number of targets. PF is the most popular one in fusing multi-modal data, due to its flexibility to support different feature types [10] and the ability to recover lost tracks [24, 48]. The Probability Hypothesis Density (PHD) [11] estimates the number and states of targets by creating the multi-target state and observation set and it is followed by clustering as post-processing to provide identity labels. In [15], a Graphical Model (GM) is built to incorporate speech signals from localized sound sources to visually tracked targets where

the audio-visual spatial-temporal coincidence further facilitates speaker diarization. To avoid post-processed identity labelling, in [14], a variational Expectation-Maximization (EM) algorithm is derived where the observation-to-person association is integrated inside the model.

In this paper, we present a novel strategy to better exploit the complementary characteristics and the roles of audio and video modalities. This allows us to estimate trajectories of multiple concurrent speakers. The details of the proposed system are described in the next sections.

III. Audio-visual Observation

A. Overview

Let $\mathbf{s}_{1:t}$ be the audio sample sequences captured by the microphone array. Let $\mathbf{J}_{1:t}$ be the RGB image sequence captured by the camera and synchronized with the audio. The camera calibration parameters are denoted as $\boldsymbol{\Upsilon}$. We aim to estimate the 3D location, of each target $i \in I_t$ (where I_t is the set of target ground truth identity labels) at time $t = 1, ..., T$ (where T is the total number of frames): $\mathbf{p}_{t,i} = (x_{t,i}, y_{t,i}, z_{t,i})^\intercal$ (where x, y, z are the individual 3D coordinates, and $\intercal$ is the transpose of a vector).

The block diagram of our proposed algorithm AV3T-VDM is shown in Fig. 1. We use PF as the underlying temporal process, which approximates the posterior probability distribution given noisy or partial observations with a group of weighted particles. Let $\boldsymbol{\chi}_{t,i} = \{(\mathbf{p}_{t,i}^{(n)}, \omega_{t,i}^{(n)}), n = 1, ..., N\}$ be the particle set of track i at time t, where $\{\cdot\}$ denotes a set, $\mathbf{p}_{t,i}^{(n)}$ and $\omega_{t,i}^{(n)}$ are the state vector and weight associated with the n^{th} particle, N is the number of particles, and $t = t_i^s, ..., t_i^e$, where t_i^s and t_i^e are the start and end time frame of track i. The AV3T-VDM takes the multi-modal observations as the inputs. It uses a data association strategy to label the observations with track identities for likelihoods computation. To handle the unknown number of targets, the re-activation, birth, and de-activation processes are introduced. Track states are then updated as the expectation over weighted particles. Finally, we re-sample the particles (particles with higher weights are duplicated while those with lower weights are eliminated) to avoid sample depletion and predict the states for the next iteration.

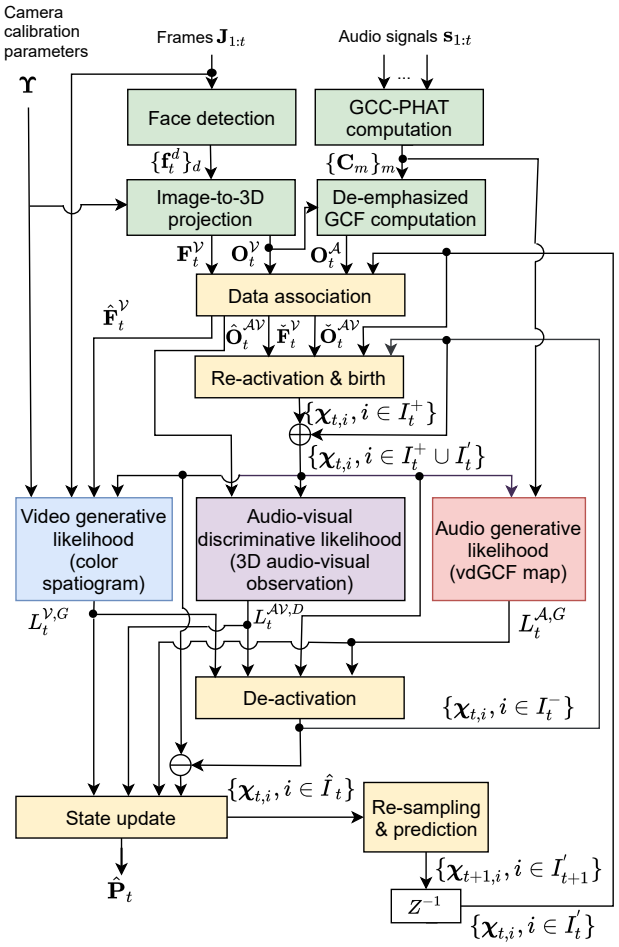


Fig. 1. Block diagram of the proposed AV3T-VDM algorithm. For simplicity, we only keep the identity label i for tracks while eliminating i in observations and likelihoods. (green: audio-visual observation; yellow: PF framework; blue: video generative likelihood; pink: audio generative likelihood; purple: multi-modal likelihood; vSSL: video-assisted SSL; vdGCF: video-assisted de-emphasized GCF. *Notations*: $\mathbf{J}_{1:t}$: images until time t; $\mathbf{s}_{1:t}$: audio signals; $\boldsymbol{\Upsilon}$: camera calibration parameters; $\{\mathbf{f}_t^d\}_d$: the set of raw face detections; $\{\mathbf{C}_m\}_m$: set of GCC-PHATs; $\mathbf{F}_t^{\mathcal{V}}$: the set of face detections; $\mathbf{O}_t^{\mathcal{V}}, \mathbf{O}_t^{\mathcal{A}}$: the set of 3D video, audio observations; $\hat{\mathbf{O}}_t^{\mathcal{AV}}$, $\hat{\mathbf{F}}_t^{\mathcal{V}}$: the set of associated observations and face detections; $\check{\mathbf{O}}_t^{\mathcal{AV}}$, $\check{\mathbf{F}}_t^{\mathcal{V}}$: the set of un-associated observations and face detections; $L_t^{\mathcal{V},G}, L_t^{\mathcal{A},G}$: video and audio generative likelihoods; $L_t^{\mathcal{AV},D}$: audio-visual discriminative likelihood; $\boldsymbol{\chi}_{t,i}$: the particle set of track i; I_t', I_t^+, I_t^- and $\hat{I}_t$: the predicted, new-born, de-activated and estimated target identity label set ($I_t' = \hat{I}_{t-1}$); $\hat{\mathbf{P}}_t$: the set of 3D target location estimates; $\oplus$: set union; $\ominus$: set exclusion.)

B. Video observation

We apply a face detector [38] on the image plane[1] to geometrically estimate the target 3D mouth location, given the camera calibration parameters and the assumption of the face diagonal size [24]. Let $\mathbf{F}_t^{\mathcal{V}} = \{\mathbf{f}_t^d, d \in D_t^{\mathcal{V}}\}$ be the face detection set at time t, where $D_t^{\mathcal{V}}$ is the detection index set and $\mathbf{f}_t^d = (u^d, v^d, w^d, h^d)_t^\intercal$ is the d^{th} face detection bounding box. Correspondingly, $\mathbf{O}_t^{\mathcal{V}} = \{\mathbf{o}_t^{\mathcal{V},d}, d \in D_t^{\mathcal{V}}\}$ is the 3D video observation set where $\mathbf{o}_t^{\mathcal{V},d}$ is the observed 3D mouth position projected from $\mathbf{f}_t^d$. The image-to-3D projection (for the visual discriminative model) is formulated as:

$$\mathbf{o}_t^{\mathcal{V},d} = \Psi(\mathbf{f}_t^d; W, H, \boldsymbol{\Upsilon}), \tag{1}$$

where Ψ is the projection operator [49]. This projection suffers from the object-camera distance ambiguity, unless prior knowledge about the object size and camera calibration information is available. Thus, we make an assumption on the human face size in 3D, and denote W and H for the width and height, respectively. $\mathbf{o}_t^{\mathcal{V},d}$ will be eliminated from set $\mathbf{O}_t^{\mathcal{V}}$ if it is outside the room range (assumed to be known), as well as the corresponding $\mathbf{f}_t^d$.

[1]A face detector is preferred to upper-body/person detection [13] because we observed that the face detection results are more robust and have fewer FPs.

We also adopt the 3D-to-image projection to enhance the visual processing part and to compensate for mis-detections by designing the visual generative model. The projection is formulated as:

$$\bar{\mathbf{f}}_t^{(n)} = \Phi(\mathbf{p}_t^{(n)}; W, H, \boldsymbol{\Upsilon}), \tag{2}$$

where $\bar{\mathbf{f}}_t^{(n)}$ is the projected face bounding box on the image, computed from a 3D hyper-rectangle derived from a particle $\mathbf{p}_t^{(n)}$, and Φ is the 3D-to-image projection operator. Detailed descriptions of Eq. 1 and Eq. 2 can be found in [24].

C. Audio observation - single speaker

The TDoA estimates are spreadly used in acoustic localization and tracking [50], which rely on the similarities of audio signals captured by a microphone pair. A general way to derive TDoA estimates is via finding the peak position of the Generalized Cross Correlation (GCC) function [51]. Since the signal amplitude depends on the speaker-array distance, distortion, and attenuation, instead of GCC, GCC-PHAT [52] is considered as a better formulation which only keeps the phase information and thus is more robust in noisy and reverberant environments. Therefore, our acoustic likelihood is built on the GCC-PHAT estimates. Given the audio signals $\mathbf{s}_t$, GCC-PHAT at the m^{th} microphone pair is computed as:

$$C_m(\tau, t) = \int_{-\infty}^{+\infty} \frac{S_{m_1}(t,f) S_{m_2}^*(t,f)}{|S_{m_1}(t,f)| \left|S_{m_2}^*(t,f)\right|} e^{j2\pi f\tau} df, \tag{3}$$

where f indicates the frequency bin, S_{m_1} and S_{m_2} are the Short-Time-Fourier-Transform (STFT) computed at the m^{th} pair with microphone indexes m_1 and m_2 ($m = 1, ..., M$ where M is the total number of microphone pairs). τ is the inter-microphone time delay and $*$ is the complex conjugate operator. Ideally, $C_m(\tau, t)$ exhibits a peak when τ equals the actual TDoA.

The GCF [53] combines the GCC-PHAT from each microphone pair in a unique representation, *e.g.* in the 3D space. As a result, the TDoA inaccuracy resulting from any single microphone pair will be averaged. The GCF is computed as:

$$g(\mathbf{p}, t) = \frac{1}{M} \sum_{m=1}^{M} C_m\left(\tau_m(\mathbf{p}), t\right), \tag{4}$$

where $\tau_m(\mathbf{p})$ is the TDoA of microphone pair m, for a source located at a generic 3D point $\mathbf{p}$. The sound location is estimated by picking the peak of the GCF.

The estimation of the distance of a sound source with a small-size circular microphone array is typically inaccurate [24, 54]. Moreover, the symmetry of a planar array causes ambiguities in the TDoA estimation [55, 56]. To assist the 3D SSL, we use the 3D video observation to suggest the most likely speaker height plane to compute a 2D GCF, which is denoted as vGCF [24]:

$$g^{\mathcal{V}}(\mathbf{p}, t) = \frac{1}{M} \sum_{m=1}^{M} C_m\left(\tau_m(\mathbf{p}_t^{\mathcal{V}_z}), t\right), \tag{5}$$

where $\mathbf{p}_t^{\mathcal{V}_z}$ indicates a 3D point with the same z value of $\mathbf{o}_t^{\mathcal{V}}$.

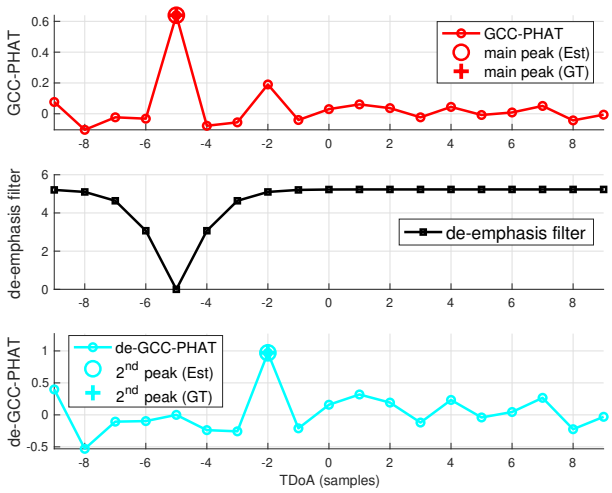


Fig. 2. Example of the de-emphasis progress [57] at a microphone pair. The red, black and cyan curves in the top, middle and bottom rows show the original GCC-PHAT, the applied de-emphasis filter and the resulting de-emphasized GCC-PHAT, respectively. The red (cyan) cross and circle indicate the ground truth and the estimated GCC-PHAT peak of the dominant (non-dominant) speaker.

D. Audio observation - multiple concurrent speakers

Finding the maxima in 3D GCF, as well as in a video-suggested 2D GCF, represents an effective way to locate a single sound source. However, in the case of concurrent speakers, only the dominant one would eventually be found. In order to locate secondary speakers, in this work we adopt the iterative method described in [57], which de-emphasizes the GCC-PHAT peak associated to the source found in the previous iteration. Fig. 2 illustrates the de-emphasis progress at a microphone pair. We can see that with de-emphasis, the non-dominant speaker can be successfully localized.

Let us denote $\bar{g}^{\mathcal{V},e}(\mathbf{p}, t)$ as the video-assisted de-emphasized GCF map *i.e.* vdGCF, where $e = 1, ..., I^E$ is the de-emphasis index with I^E the maximum concurrent speaker number for each vdGCF. When $e = 1$, $\bar{g}^{\mathcal{V},e}(\mathbf{p}, t) = g^{\mathcal{V}}(\mathbf{p}, t)$ which localizes the dominant speaker. The iterative steps at each video-suggested speaker height plane are listed as below:

Step 1: Find the 3D location of the e^{th} dominant speaker from vdGCF:

$$\mathbf{p}_t^e = \underset{\mathbf{p}}{\operatorname{argmax}}\, \bar{g}^{\mathcal{V},e}(\mathbf{p}, t). \tag{6}$$

Step 2: Compute the corresponding TDoA at each microphone pair (Fig. 2, first row):

$$\tau_m^e(\mathbf{p}_t^e) = \left\lceil \frac{||\mathbf{p}_t^e - \mathbf{p}_{m_1}|| - ||\mathbf{p}_t^e - \mathbf{p}_{m_2}||}{c} f_a \right\rfloor, \tag{7}$$

where c is the sound speed, f_a is the audio sampling frequency, $\mathbf{p}_{m_1}$ and $\mathbf{p}_{m_2}$ are the positions of the m^{th} microphone pair and $\lceil \cdot \rfloor$ is the operation that rounds a number to the nearest integer.

Step 3: Generate the de-emphasized GCC-PHAT (Fig. 2, last row):

$$C_m^{e+1}(\tau, t) = \delta\left(\tau, \tau_m^e(\mathbf{p}_t^e)\right) C_m^e(\tau, t), \tag{8}$$

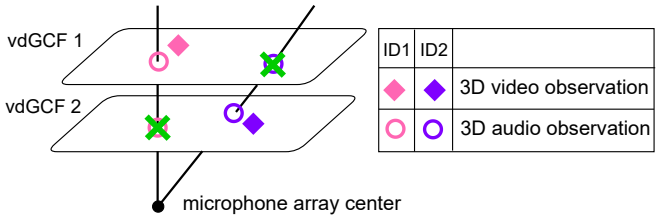


Fig. 3. Elimination (green crosses) of the 3D audio observations detected at a height plane of another video observation.

where $\delta(\tau, \tau_m^e(\mathbf{p}_t^e))$ is the de-emphasis filter (Fig. 2, middle row), derived as [57]:

$$\delta(\tau, \tau_m^e(\mathbf{p}_t^e)) = \alpha_\delta \left(1 - e^{-|\tau - \tau_m^e(\mathbf{p}_t^e)|^\varrho}\right), \tag{9}$$

where ϱ controls the notch sharpness, and α_δ is the normalization factor that guarantees the same GCC-PHAT sum at the m^{th} microphone pair after de-emphasis: $\sum_\tau C_m^{e+1}(\tau, t) = \sum_\tau C_m^e(\tau, t)$.

Step 4: Compute vdGCF as a combination of all the de-emphasized GCC-PHATs:

$$\bar{g}^{\mathcal{V},e+1}(\mathbf{p}, t) = \frac{1}{M} \sum_{m=1}^{M} C_m^{e+1}\left(\tau_m(\mathbf{p}_t^{\mathcal{V}_z}), t\right). \tag{10}$$

Finally, the corresponding vdGCF peak is thus computed as:

$$\bar{G}_t^{\mathcal{V},e+1} = \max\ \bar{g}^{\mathcal{V},e+1}(\mathbf{p}, t). \tag{11}$$

At a time frame, when multiple faces are detected, we apply vdGCF at different video observation height planes. To eliminate the 3D audio observation detected at a height plane of the other video observation, after the vdGCF progress, we compute the 3D distance between each video observation and the audio observations. We eliminate the audio observations with a larger distance to any video observations. Two audio observations are considered from the same speaker if their DoA difference is smaller than 2°. Fig. 3 illustrates an example. The 3D video observations from two targets (magenta and purple diamonds) result in four audio observations (circles) on the two planes. The audio observations (circles) with green crosses are eliminated due to the large distance with both 3D video observations.

The performance of the de-emphasis localization method drops considerably when more than two concurrent speakers are tracked [57]. To eliminate influence from the noise, constraints are applied on the vdGCF results. Let $\mathbf{O}_t^{\mathcal{A}} = \{\mathbf{o}_t^{\mathcal{A},e}, e \in D_t^{\mathcal{A}}\}$ be the 3D audio observation set and $\mathbf{O}_t^{\mathcal{AV}} = \mathbf{O}_t^{\mathcal{A}} \bigcup \mathbf{O}_t^{\mathcal{V}} = \{\mathbf{o}_t^{\mathcal{AV},d}, d \in D_t^{\mathcal{AV}}\}$ be the 3D audio-visual observation set. $D_t^{\mathcal{A}}$ and $D_t^{\mathcal{AV}}$ are the label sets of audio and audio-visual observations, respectively. Let $\varepsilon(\cdot,\cdot)$ be the euclidean distance between any two variables indicated by "$\cdot$" in the bracket. And let $\varepsilon(\cdot,\cdot|\vartheta)$ represents the result of the euclidean distance comparison whose value will be *true* if $\varepsilon(\cdot,\cdot) \leq \vartheta$ and *false* otherwise (ϑ is an arbitrary variable). The following validation stages are introduced:

Stage 1: Peak validation: The GCF peak value can be used to detect voice activities and to estimate the number of active speakers [57]. Since the applied de-emphasis filter may enhance the noise, a larger, less pronounced de-emphasized GCF peak (than the original one) should be avoided. Thus, the peak validation boolean flag for an observation with the de-emphasis index e is computed as:

$$\eta_t^p = \left\{\bar{G}_t^{\mathcal{V},e-1} \geq \bar{G}_t^{\mathcal{V},e} \geq \max\left(\frac{\bar{G}_t^{\mathcal{V}}}{2}, \vartheta_g^{\mathcal{A}}\right)\right\}, \tag{12}$$

where $\vartheta_g^{\mathcal{A}}$ is the constant peak threshold. $\frac{\bar{G}_t^{\mathcal{V}}}{2}$ is a temporal-varying threshold to guarantee the large de-emphasized GCF peak is resulting from a real sound source instead of the residual of the original peak. We use $\{\cdot\}$ to indicate a logical operator if "$\cdot$" includes comparisons or boolean results, and is set to true if all the conditions hold.

Stage 2: Temporal-spatial validation: A real sound source appears consecutively in a constrained spatial region. We denote $\check{\mathbf{o}}$ as a generic un-associated observation. The temporal-spatial validation boolean flag for an un-associated 3D audio observation $\check{\mathbf{o}}_t^{\mathcal{A},e}$ is set to *true* if in the past $t|_{\Delta t}$ frames, for at least $\vartheta_{\Delta t}^{\mathcal{AV}}$ frames, it is close to any other un-associated audio-visual observations. Let us denote this assignment through the following notation:

$$\eta_t^s = \left\{\varepsilon\left(\check{\mathbf{o}}_t^{\mathcal{A},e}, \check{\mathbf{o}}_{t'}^{\mathcal{AV},\check{d}} | \vartheta_{3D}^{\mathcal{AV}}\right), \forall\ t' \in t|_{\vartheta_{\Delta t}^{\mathcal{AV}}}, \exists\ \check{d} \in D_{t'}^{\mathcal{AV}}\right\} \tag{13}$$

where $\vartheta_{3D}^{\mathcal{AV}}$ and $\vartheta_{\Delta t}^{\mathcal{AV}}$ are the 3D error and time thresholds. $t|_{\vartheta_{\Delta t}^{\mathcal{AV}}}$ indicates any $\vartheta_{\Delta t}^{\mathcal{AV}}$ frames of $t|_{\Delta t}$ with $t|_{\Delta t} = [t - \Delta t, t)$.

With the proposed validation stages, the 3D audio observation (for the audio discriminative model) is finally estimated as:

$$\mathbf{o}_t^{\mathcal{A},e} = \underset{\mathbf{p} \in \mathbf{P}_t^{\mathcal{V}_z}}{\operatorname{argmax}}\ \bar{g}^{\mathcal{V},e}(\mathbf{p}, t), \quad \textit{if} \quad \eta_t^p \wedge\ \eta_t^s, \tag{14}$$

where $\wedge$ is the *and* operator which indicates the 3D audio observation passes the peak and temporal-spatial validations. $\mathbf{P}_t^{\mathcal{V}_z}$ are the 3D points with the same z value of $\mathbf{o}_t^{\mathcal{V}}$ within the room range.

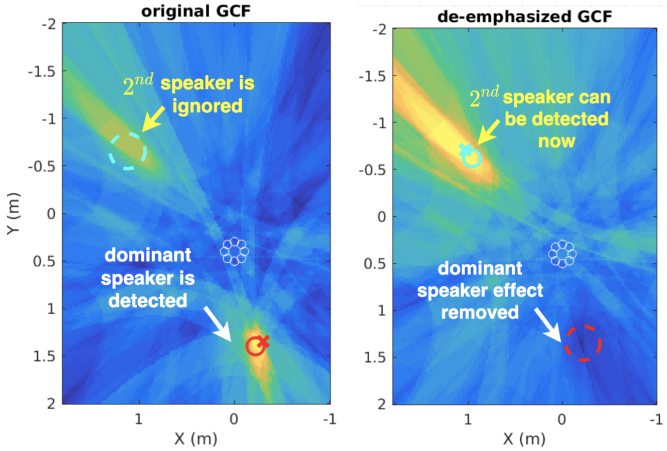
Fig. 4 illustrates the 3D audio observations from vdGCF. Influence of the dominant speaker (marked as the red circle in Fig. 4(a)) has been eliminated while the effect from the second speaker is enhanced, that results in a correct location estimate (the cyan cycle in Fig. 4(b)) at the second speaker height plane suggested by video. The pseudo-code of the multi-speaker localization method with the de-emphasized GCF is provided in *Algorithm* 1.

IV. Multi-modal likelihood

Our proposed multi-modal likelihood consists of the audio and video generative likelihoods, $L_{t,i}^{\mathcal{A},G}$ and $L_{t,i}^{\mathcal{V},G}$; and the audio-visual discriminative likelihood, $L_{t,i}^{\mathcal{AV},D}$:

$$L_{t,i}^{\mathcal{AV}}(\mathbf{J}_t, \mathbf{s}_t|\mathbf{p}) = L_{t,i}^{\mathcal{A},G}(\mathbf{s}_t|\mathbf{p}) L_{t,i}^{\mathcal{V},G}(\mathbf{J}_t|\mathbf{p}) L_{t,i}^{\mathcal{AV},D}(\mathbf{J}_t, \mathbf{s}_t|\mathbf{p}). \tag{15}$$

Assuming that they are independent, likelihood components are multiplied. On one hand, this promotes an equal contribution of each component. On the other hand, this formulation



(a) (b)

Fig. 4. Normalized video-assisted acoustic map at different speaker height planes: (a) the original map, vGCF, at the dominant speaker height plane; (b) the de-emphasized map, vdGCF, at the secondary speaker height plane: it de-emphasizes the dominant speaker (the red cross in Fig. 4(a)) to localize the secondary speaker (cyan cross in Fig. 4(b)). Circles and crosses indicate the location estimates and ground truth, respectively. Blue (yellow) represents the region with a lower (higher) probability of including a sound source.

does not fully exploit the inter-dependencies of the multi-modal likelihoods.

A. Generative Likelihoods

The *video generative likelihood* consists of the color spatiogram [58]. It is used to find the image sub-region with the highest similarity to the reference image (*i.e.* last face detection) of the tracked target and is defined as:

$$h^b_{\mathbf{f}} = (r^b_{\mathbf{f}}, \mu^b_{\mathbf{f}}, \Sigma^b_{\mathbf{f}}), \quad b = 1, ..., B \tag{16}$$

where B is the number of bins per channel, $r^b_{\mathbf{f}}$ is the number of pixels at image region $\mathbf{f}$ whose value is within the b^{th} bin, $\mu^b_{\mathbf{f}}$ and $\Sigma^b_{\mathbf{f}}$ are the spatial mean and covariance matrix of the pixel coordinates, respectively. For each 3D hypothesis $\mathbf{p}$, we create a 3D hyper-rectangle and project it onto the image plane to derive a face bounding box $\bar{\mathbf{f}}_{t,i}$ (Eq. 2). For simplicity, we eliminate the time and identity index (t and i) in computing the similarity score $\beta_{t,i}$:

$$\beta(\mathbf{f}, \bar{\mathbf{f}}) = \sum_{b=1}^{B} \sqrt{r^b_{\mathbf{f}} r^b_{\bar{\mathbf{f}}}} \left(8\pi |\Sigma^b_{\mathbf{f}} \Sigma^b_{\bar{\mathbf{f}}}|^{\frac{1}{4}} \mathcal{N}(\mu^b_{\mathbf{f}} | \mu^b_{\bar{\mathbf{f}}}, 2(\Sigma^b_{\mathbf{f}} + \Sigma^b_{\bar{\mathbf{f}}})) \right), \tag{17}$$

where $\mathcal{N}$ is the normal distribution.

The likelihood is thus derived as:

$$L^{\mathcal{V},G}_{t,i}(\mathbf{J}_t|\mathbf{p}) = \begin{cases} \beta_{t,i} & \text{if } \beta_{t,i} \geq \vartheta^{\mathcal{V}}_G, \\ \mathcal{U}(\mathbf{p}) & \text{otherwise}, \end{cases} \tag{18}$$

where $\mathcal{U}$ is the uniform distribution and $\vartheta^{\mathcal{V}}_G$ is the spatiogram similarity threshold.

When a 3D audio observation is associated to track i, we use vdGCF as the likelihood. The *audio generative likelihood* is formulated as:

$$L^{\mathcal{A},G}_{t,i}(\mathbf{s}_t|\mathbf{p}) = \begin{cases} \bar{g}^{\mathcal{V},e}(\mathbf{p}, t) & \text{if } \bar{G}^{\mathcal{V},e}_{t,i} \geq \vartheta^{\mathcal{A}}_G, \\ \mathcal{U}(\mathbf{p}) & \text{otherwise}, \end{cases} \tag{19}$$

Algorithm 1: De-emphasized GCF (for localization of multiple concurrent speakers)

Input:
$\mathbf{O}^{\mathcal{V}}_t$: 3D video observations
$\{\mathbf{C}_m\}_m$: GCC-PHATs
Output:
$\mathbf{O}^{\mathcal{A}}_t$: 3D audio observations

for $d \in D^{\mathcal{V}}_t$ **do**
  $\mathbf{P}^{\mathcal{V}_z}_t = \{\mathbf{p}^{\mathcal{V}_z}_t\}$ *% create a set of 3D grid points*
  $e = 1$ *% de-emphasis index*
  while $e \leq I^E$ **do**
    compute $\bar{g}^{\mathcal{V},e}(\mathbf{P}^{\mathcal{V}_z}_t, t)$ *% vdGCF*
    compute η^p_t, Eq. 12 *% peak validation*
    compute η^s_t, Eq. 13 *% temporal-spatial validation*
    if $\eta^p_t \wedge \eta^s_t$ **then**
      $\mathbf{o}^{\mathcal{A},e}_t = \underset{\mathbf{p} \in \mathbf{P}^{\mathcal{V}_z}_t}{\operatorname{argmax}} \bar{g}^{\mathcal{V},e}(\mathbf{p}, t)$
    end
    $\bar{G}^{\mathcal{V},e}(t) = \underset{\mathbf{p} \in \mathbf{P}^{\mathcal{V}_z}_t}{\max} \bar{g}^{\mathcal{V},e}(\mathbf{p}, t)$ *% peak value*
    $e = e + 1$
  end
end

% eliminate audio observation detected at the height plane of the other video observation, Fig.3.
if $|D^{\mathcal{V}}_t| \geq 2$ **then**
  $e_1 \in D^{\mathcal{A}}_t$
  $e_2 \in D^{\mathcal{A}}_t \setminus e_1$ *% "\" means exclusion*
  if $\varepsilon^{\theta}(\mathbf{o}^{\mathcal{A},e_1}_t, \mathbf{o}^{\mathcal{A},e_2}_t) < 2^\circ$ **then**
    $e = \underset{e \in \{e_1, e_2\}}{\operatorname{argmax}} \left(\underset{d \in D^{\mathcal{V}}_t}{\min} \varepsilon(\mathbf{o}^{\mathcal{A},e}_t, \mathbf{o}^{\mathcal{V},d}_t) \right)$
    $\mathbf{O}^{\mathcal{A}}_t = \mathbf{O}^{\mathcal{A}}_t \setminus \mathbf{o}^{\mathcal{A},e}_t$ *% eliminate observation*
  end
end

where $\bar{G}^{\mathcal{V},e}_{t,i}$ is the vdGCF peak computed among all the particles belonging to track i, and $\vartheta^{\mathcal{A}}_G$ is the Voice Activity Detection (VAD) threshold.

B. Discriminative Likelihoods

The main inaccuracy of the 3D video observations is due to the scaling factor ambiguity in the image-to-3D projection in Eq. 1. This occurs because our implementation assumes a constant diagonal size of target faces. However, the face detection bounding box size varies with different target identities and their face orientations. In addition, a face detector is more reliable on a frontal view and becomes weak when a face is not oriented towards the camera [15]. Thus, in the camera's spherical coordinates, compared to the azimuth and elevation estimates, the depth estimate is less accurate. Similar error characteristics can be observed on the 3D audio observation (Eq. 14). Due to the use of a small-size circular microphone array, the depth estimate is less accurate than in the other dimensions. Therefore, considering the error characteristics of

the 3D observations, a multi-variant Gaussian model is built in the sensor's spherical coordinates.

The *audio-visual discriminative likelihood* is finally derived as:

$$L_{t,i}^{\mathcal{AV},D}(\mathbf{J}_t, \mathbf{s}_t|\mathbf{p}) = \begin{cases} \sum\limits_{d\in D_{t,i}^{\mathcal{AV}}} \alpha_{t,i}^{d}\mathcal{N}(\tilde{\mathbf{p}}|\tilde{\mathbf{o}}_{t,i}^{\mathcal{AV},d}, \Sigma^{\mathcal{AV}}) & \textit{if } D_{t,i}^{\mathcal{AV}} \neq \varnothing, \\ \mathcal{U}(\mathbf{p}) & \textit{otherwise}, \end{cases} \tag{20}$$

where $\tilde{(\cdot)}$ indicates the location vector $(\cdot)$ in the sensor's spherical coordinates; $D_{t,i}^{\mathcal{AV}}$ is observation label set of track i; $\Sigma^{\mathcal{AV}}$ is the covariance matrix which accounts for the observation reliability across different dimensions and it is set to diagonal assuming independent error distribution as discussed in [24]; $\alpha_{t,i}^{d}$ is the reliability measure derived from the observation-to-track association which is defined in the next section (in Eq.23).

In summary, the audio-visual discriminative likelihood helps because (1) the video-assisted de-emphasized GCF provides 3D audio observations which compensate the error from the video-only observations; (2) the 3D observation error is modeled in individual spherical coordinate that separates the inaccurate depth estimates from the other coordinates and (3) it assists the tracker to manage a varying number of targets, as described in the next section.

V. Track management

For MTT, apart from data association between new observations and the existing tracks, additional stages such as track birth, re-activation, and de-activation are introduced to manage a varying number of targets. We use I_t', I_t^+, I_t^-, and $\hat{I}_t$ to denote the predicted, new-born, de-activated, and estimated target identity label sets at time t, respectively. We set $I_t' = \hat{I}_{t-1}$ for the prediction to coincide with the previous identity label set.

Data association assigns observations to tracks. The task is challenging when targets are closer in the observation space than the observation errors [59]. Observation uncertainties make the association even more complicated. For example, when the multi-modal system re-detects or misses a target, tracks may have ID swaps or incorrect de-activations may occur. Furthermore, while observations from different modalities may lead to more robust operation, they also further complicate the association task. Since audio and video modalities may generate different 3D observations from the same target (as described in Sec. III), we use the WPDA strategy for multiple-to-one observation-to-track assignment.

Gating is used to validate possible matching observations and a binary gating flag between track i and observation d is computed as:

$$G_t(i,d) = \left\{\varepsilon^{\star}\left(\tilde{\mathbf{p}}_{t-1,i}, \tilde{\mathbf{o}}_t^{\mathcal{AV},d}|\vartheta_{\star}^{\mathcal{AV}}\right), \forall \star \in \{\theta,\phi,R\}\right\}, \tag{21}$$

where θ, ϕ, R denote the azimuth, elevation and radius of the spherical coordinates, respectively. $\varepsilon^{\star}(\cdot,\cdot|\vartheta_{\star}^{\mathcal{AV}})$ indicates the distance comparison result in coordinate *. $\vartheta_{\star}^{\mathcal{AV}}$ is the gating limit which specifies the elliptic cone-shaped gate region. Remind that $\{\cdot\}$ is a logical operation outputs true if all the comparisons in "$\cdot$" satisfy.

The *association* of a gate-validated observation to a track is based on minimizing the association distance, which is computed using (i) the de-activation time; (ii) the visual dis-similarity measure and (iii) the spatial error in sensor's spherical coordinates:

$$A_t(i,d) = \frac{t - t_i^e}{T_n} + (1 - \beta_{t,i}) + \sum_{\star\in\{\theta,\phi,R\}} \frac{\varepsilon^{\star}(\tilde{\mathbf{p}}_{t-1,i}, \tilde{\mathbf{o}}_t^{\mathcal{AV},d})}{\vartheta_{\star}^{\mathcal{AV}}}, \tag{22}$$

where t_i^e is the end frame of track i, $\beta_{t,i}$ is computed in Eq.17 and T_n is the maximum de-activation time.

The reliability measure in the multi-modal likelihood (Eq. 20) is computed as:

$$\alpha_{t,i}^{d} = \frac{A_t(i,d)}{\sum\limits_{d\in D_{t,i}^{\mathcal{AV}}} A_t(i,d)}. \tag{23}$$

Re-activation and *birth* are applied to observations that are outside the gating region. The un-associated observations, denoted as $\check{\mathbf{O}}_t^{\mathcal{AV}} = \{\check{\mathbf{o}}_t^{\mathcal{AV},d},\ d \in \check{D}_t^{\mathcal{AV}}\}$, may result either from a de-activated track or a new track. We re-activate a de-activated track i at a detection with label d if the association distance $\alpha_{t,i}^{d}$ is below a threshold A^R, and is smaller than the association distance to any of the active tracks. Then, the remaining un-associated observations are used to generate new tracks: a new track will be assigned to $\check{\mathbf{o}}_t^{\mathcal{AV},d}$ if in the past $t|_{\Delta t}$ frames, for at least $\vartheta_{\Delta t}^{\mathcal{AV}}$ frames, $\check{\mathbf{o}}_t^{\mathcal{AV},d}$ is close to any other un-associated observations. The *temporal-spatial validation* is defined as:

$$\eta_d^+ = \left\{\varepsilon\left(\check{\mathbf{o}}_t^{\mathcal{AV},d}, \check{\mathbf{o}}_{t'}^{\mathcal{AV},\check{d}}|\vartheta_{3D}^{\mathcal{AV}}\right), \forall t \in t|_{\vartheta_{\Delta t}^{\mathcal{AV}}}, \exists\ \check{d} \in \check{D}_{t'}^{\mathcal{AV}}\right\} \tag{24}$$

The pseudo-code of data association, track re-activation and birth are given in *Algorithm* 2 and *Algorithm* 3, respectively. The variable $\mathbf{q}$ used for track re-activation and birth in *Algorithm* 3 is Gaussian distributed noise whose covariance matrix is the same as in the audio-visual discriminative likelihood ($\Sigma^{\mathcal{AV}}$ in Eq. 20), since $\Sigma^{\mathcal{AV}}$ represents the observation error allowance.

A track i is *de-activated*, if during the past time interval $t|_{\Delta t}$, it is not active or it overlaps with another track.

A track is considered as *non-active* if in $t|_{\Delta t}$ the likelihood follows a uniform distribution:

$$\eta_i^- = \left\{L_{t,i}^{\mathcal{AV}}(\mathbf{J}_t, \mathbf{s}_t|\mathbf{p}) \overset{i.i.d.}{\sim} \mathcal{U}(\mathbf{p}), \forall\ t \in t|_{\Delta t}\right\}, \tag{25}$$

where *i.i.d.* means independent and identically distributed.

A track overlaps with another track if their 3D distance is small, formulated as:

$$\eta_i^{\times} = \left\{\varepsilon\left(\bar{\mathbf{p}}_{t|_{\Delta t},i}, \bar{\mathbf{p}}_{t|_{\Delta t},\tilde{i}} \mid \vartheta_{3D}^{\mathcal{AV}}\right), \exists\ \tilde{i}\right\}, \tag{26}$$

where $\tilde{i}$ indicates the label of any other tracks and $\bar{\mathbf{p}}_{t|_{\Delta t},i}$ is the average target state estimate in $t|_{\Delta t}$.

The pseudo-code of the de-activation process is given in *Algorithm* 4.

Algorithm 2: Data association

Input:
$\mathbf{F}_t^{\mathcal{V}}$: face detections
$\mathbf{O}_t^{\mathcal{AV}}$: 3D audio-visual observations
$\{\boldsymbol{\chi}_{t,i}, i \in I_t'\}$: predicted tracks
Output:
$\hat{\mathbf{O}}_t^{\mathcal{AV}}$, $\check{\mathbf{O}}_t^{\mathcal{AV}}$: associated, un-associated observations
$\hat{\mathbf{F}}_t^{\mathcal{V}}$, $\check{\mathbf{F}}_t^{\mathcal{V}}$: associated, un-associated face detections

for $i \in I_t'$ *and* $d \in D_t^{\mathcal{AV}}$ **do**
  compute $G_t(i,d)$, Eq. 21 *% gating score*
  compute $A_t(i,d)$, Eq. 22 *% association distance*
end

for $d \in D^{\mathcal{AV}}$ **do**
  $i^\star = \underset{i^\star \in I_t'}{\operatorname{argmin}} A_t(i,d)$
  if $G_t(i^\star, d)$ **then**
    $\hat{\mathbf{O}}_{t,i^\star}^{\mathcal{AV}} \leftarrow \mathbf{o}_t^{\mathcal{AV},d}$ *% observation-track association*
    $\hat{\mathbf{F}}_{t,i^\star}^{\mathcal{V}} \leftarrow \mathbf{f}_t^{\mathcal{AV},d}$
  else
    $\check{\mathbf{O}}_t^{\mathcal{AV}} \leftarrow \mathbf{o}_t^{\mathcal{AV},d}$ *% un-associated observation*
    $\check{\mathbf{F}}_t^{\mathcal{V}} \leftarrow \mathbf{f}_t^{\mathcal{AV},d}$
    go to *Re-activation* and *birth* (*Algorithm* 3)
  end
end

Algorithm 3: Re-activation and birth

Input:
$\{\boldsymbol{\chi}_{t,i}, i \in I_t'\}$: predicted tracks
$\{\boldsymbol{\chi}_{t,i}, i \in I_t^-\}$: de-activated tracks
$\check{\mathbf{O}}_t^{\mathcal{AV}}$: un-associated observations
$\check{\mathbf{F}}_t^{\mathcal{V}}$: un-associated face detections
Output:
$\{\boldsymbol{\chi}_{t,i}, i \in I_t^+\}$: new-born tracks

for $d \in \check{D}_t^{\mathcal{AV}}$ *and* $i \in I_t^-$ **do**
  compute $A_t(i,d)$, Eq. 22 % association distance
end

if $\left\{\alpha_{t,i}^d \leq A^R, \exists\ i \in I_t^-\right\}$ **then**
  $i^+ = \operatorname{argmin}_{i \in I_t^-} A_t(i,d)$ *% re-activation label*
  $\hat{\mathbf{p}}_{t,i^+} = \check{\mathbf{o}}_t^{\mathcal{AV},d}$ *% re-activate at observation*
  $(\mathbf{p}_{t,i^+}^{(n)}, \omega_{t,i^+}^{(n)}) = (\hat{\mathbf{p}}_{t,i^+} + \mathbf{q}, \frac{1}{N})$
else
  compute η_t^+, Eq. 24
  if η_t^+ **then**
    $i^+ = |I_t'| + 1$
    $I_t^+ \leftarrow i^+$ *% new-born label*
    $t_{i^+}^0 = t$ *% birth time*
    $\hat{\mathbf{p}}_{t,i^+} = \check{\mathbf{o}}_t^{\mathcal{AV},d}$ *% birth at observation*
    $(\mathbf{p}_{t,i^+}^{(n)}, \omega_{t,i^+}^{(n)}) = (\hat{\mathbf{p}}_{t,i^+} + \mathbf{q}, \frac{1}{N})$
  end
end

Algorithm 4: De-activation

Input:
$\{\boldsymbol{\chi}_{t,i}, i \in I_t^+ \cap I_t'\}$: active tracks
$L_t^{\mathcal{A},G}, L_t^{\mathcal{V},G}, L_t^{\mathcal{AV},D}$: likelihoods
Output:
$\{\boldsymbol{\chi}_{t,i}, i \in I_t^-\}$: de-activated tracks

for $i \in I_t^+ \cap I_t'$ **do**
  if $(t - t_i^0) \leq \Delta t$ **then**
    return *% don't de-activate new tracks*
  end
  compute η_i^-*, Eq. 25* *% non-active*
  compute $\eta_i^\times$*, Eq. 26* *% overlapped*
  if η_i^- *or* $\eta_i^\times$ **then**
    $I_t^- \leftarrow i$ *% de-activation label*
    $t_i^e = t$ *% de-activation time*
  end
end

VI. Experimental Results

We compare AV3T-VDM with the SoA methods [10, 11, 60] on the AV16.3 [61] and the CAV3D dataset [24]. We first describe the implementation details and evaluation measures and then the benefits of the multi-modal likelihood modeling and contributions of each component. Finally, we discuss the results of the comparison and discuss the limitations of the proposed tracker.

A. Implementation details

We validate AV3T-VDM on the multi-speaker sequences of the AV16.3 and of the CAV3D datasets. With respect to other audio-visual corpora, these two datasets are recorded using small-size circular microphone arrays and standard RGB cameras, and include sensor calibration information and 3D target location as the annotations.

For the AV16.3 dataset, we use the same sequences specified in [11] and [60]. The first circular microphone array and each (of the three) corner cameras are used individually for the experiments. For the CAV3D dataset, we use the multi-speaker sequences (*i.e.* the CAV3D-MOT subset). Specifications of the datasets and experiments are provided in Tab.II. The audio sampling frequency f_a is 16 kHz in AV16.3 and 96 kHz in CAV3D while the video frame rate f_v is 25 fps (360×288 pixels) and 15 fps (1024×768 pixels), respectively. Besides, we list the number of sequences, summarize the total frame number, the % of frames with target crosses and the % of frames where a target is outside the camera's FoV.

We apply a CNN based face detector [38] on each frame and geometrically extract mouth positions through the image-to-3D projection (Eq. 2). The STFT window length for audio processing equals to 2^{12} ($256\ ms$) and 2^{15} ($341\ ms$) for AV16.3 and CAV3D respectively, with the overlapping factor sets for one-to-one audio-visual frame correspondence; the VAD threshold $\vartheta_G^{\mathcal{A}}$ for the acoustic generative model, is set

TABLE II
Specifications of the datasets and experiments.

	AV16.3	CAV3D
Image frame rate	25 fps	15 fps
Image resolution	360×288	1024×768
Audio sampling rate	16 kHz	96 kHz
Training sequences	seq08,11,12,18,19	CAV3D-SOT subset
Testing sequences	seq24, 25, 30, 45	CAV3D-MOT subset
# fr: total (2, 3 speakers)	8512 (6276, 2236)	2455 (1571, 884)
% fr: target crosses	21.0%	27.9%
% fr: target outside FoV	27.7%	37.1%

to 0.1 and 0.03 correspondingly[2], with the multi-speaker localization threshold $\vartheta_g^{\mathcal{A}} = 1.5\vartheta_G^{\mathcal{A}}$; the number of bins per color channel is $B = 8$ and the visual spatiogram threshold $\vartheta_G^{\mathcal{V}}$ equals 0.4; $(W, H) = (14, 18)\ cm$ is an approximate average size of the central region of a face, which is fixed for every target; the thresholds in data association is set as $T_n = 100$ frames, $\vartheta_\star^{\mathcal{AV}} = (15°, 15°, 1.5m)$. The activation time Δt in data association, birth, and re-activation, and de-activation stages is set as Δt =5 frames, with the time and 3D validation $\vartheta_{\Delta t}^{\mathcal{AV}} = \Delta t - 1$ and $\vartheta_{3D}^{\mathcal{AV}} = 0.3m$ (Eq. 13); all the microphone pairs are used for GCF computation, thus $M = 28$; the particle number (per target) is $N = 50$; the covariance matrix in the audio-visual discriminative likelihood (Eq. 20) is $\Sigma^{\mathcal{AV}}$ =diag$(2°, 2°, 0.4m)$; the maximum concurrent speaker number I^E for each vdGCF is set to 2, as such in [57], considering the degrading performance with the presence of more speakers; ϱ is set to 1.3 empirically; the re-activation threshold A^R is set to 5, which is the item number in A_t (Eq. 22).

Parameters are optimized on the training sequences (seq08, 11, 12, 18 and 19 in AV16.3 and the CAV3D-SOT subset), which are not used for testing. Excluding parameters related to the audio sampling rate (that differs considerably), we use the same parameters for both datasets. Given the rather different room geometries and target behaviours, the proposed AV3T-VDM proves to be effective and extendable in different conditions.

For the tracking results, we repeat the experiments for 10 iterations and report the mean and the standard deviations (the number after "±") as the final results.

B. Performance measures

We use Multiple Object Tracking Accuracy (MOTA) and Optimal SubPattern Assignment for Tracks (OSPAT) to quantitatively evaluate the performance of different tracking algorithms. The symbols '↑' and '↓' mean 'the higher the better results' and 'the lower the better results', respectively.

Different from Mean Absolute Error (MAE), which only measures the average error over the whole sequence without considering the instant error made by a tracker, MOTA [62] is a more expressive metric that combines three sources of instant errors *i.e.* FP, FN and IDentity switches (IDs):

$$\text{MOTA} = 100 \times \left(1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDs}_t)}{\sum_t GT_t}\right), \quad (27)$$

[2]Note that different audio parameters are due to different audio sampling frequencies in the two datasets.

TABLE III
Comparison between our proposed AV3T-VDM likelihood and the one used in [24]. Results are given as the mean and standard deviations (the number after '±') over 10 iterations. fps: frame per second.

	Likelihoods	OSPAT↓	MOTA$_{img}$ ↑	MOTA$_{3D}$ ↑	fps
AV16.3	[24]	11.5±0.9	77.9±4.9	27.7±9.0	5.3
	prop.	**9.5**±0.9	**85.6**±3.6	**65.0**±1.2	3.2
CAV3D	[24]	15.4 ±2.2	71.2±5.0	35.5±5.3	1.6
	prop.	**15.2**±1.3	**75.4**±4.1	**45.3**±2.4	1.3

where GT_t is the number of ground truth targets at time t. Note that MOTA $\in (-\infty, 100]$ and it may be negative in cases when the number of errors made by the tracker exceeds the target number in the scene. MOTA is evaluated both in 3D (MOTA$_{3D}$) as well as on the image plane (MOTA$_{img}$).

The OSPAT [63] considers the cardinality differences between the ground truth and the estimated target set [64], which is suitable for evaluating tracking with a varying number of targets. It also allows us to make a direct comparison with the SoA method [11]. The equation is given as:

$$\text{OSPAT}_t = \min_{\gamma \in \Gamma_{\hat{I}_t, I_t}} \sqrt[a]{\frac{1}{\hat{I}_t}\left(\sum_{i=1}^{|\hat{I}_t|} \varepsilon^{(c)}(\hat{\mathbf{p}}_{i,t}, \mathbf{p}_{i,t})^{\mathcal{A}} + c^{\mathcal{A}}(|I_t| - |\hat{I}_t|)\right)}, \quad (28)$$

where $\Gamma_{\hat{I}_t, I_t}$ indicates the set of maps: $\gamma : 1, ..., |\hat{I}_t| \rightarrow 1, ..., |I_t|$. $\varepsilon^{(c)}(\cdot)$ is a function defined as $\min(c, \varepsilon(\cdot))$ where c is the cut-off value controls the penality assigned to the identity and localization errors. a is a metric order which relates to the sensitivity to the outliers.

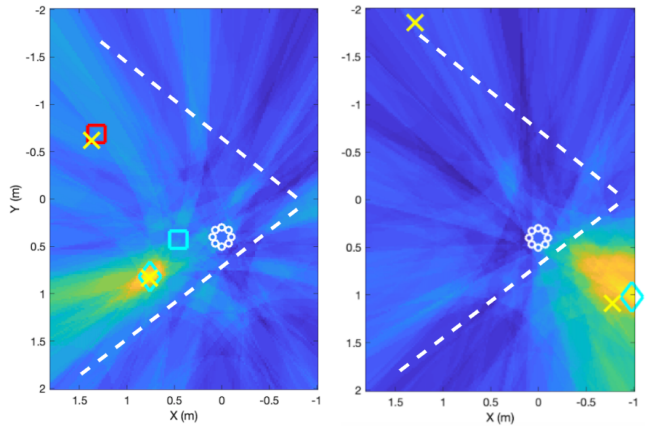
Apart from MOTA, Multiple Object Tracking Precision (MOTP) is another popular MTT evaluation metric which eliminates the instant error and presents the average position error among all the correctly tracked targets [62]. We don't use this metric here since this information is already included in OSPAT.

In this paper, MOTA$_{img}$ considers FP with the state whose pixel error is smaller than $\frac{1}{15}$ of the image diagonal size, where MOTA$_{3D}$ considers FP with $\vartheta_{3D}^{\mathcal{AV}}$. OSPAT in Eq. 28 is computed with c and a set to 35 and 2 respectively, as such in [10, 11, 60].

C. Multi-modal likelihood evaluation

We evaluate the proposed multi-modal likelihood (Eq. 15) and the data association method (*Algorithm* 2), by replacing the counterparts in the tracking framework of [24]. For the experiment, we use the same sequences as in [24] and assume a known number of targets to eliminate the influence of track initialization and de-activation. Tracks are initialized at the ground truth.

The comparison results are given in Tab. III. The execution time given in the last column was measured running the algorithms on a 3.4 GHz Intel(R) Core(TM) i7-4770 CUP. From the results, we can see that AV3T-VDM provides a better tracking accuracy both on the image plane and in 3D which compensates the slower processing speed. Large improvements can be observed in terms of MOTA$_{3D}$. This is because in [24], MOTA$_{3D}$ becomes very small due to the inaccurate 3D



(a) (b)

Fig. 5. Localization of multiple concurrent speakers: when (a) two targets are inside the camera's FoV (the triangular region surrounded by the two white dashed lines) and (b) an active speaker is outside the FoV. (yellow crosses: ground truth; red and cyan indicate the two identities; squares: video observations; diamonds: audio observations.)

TABLE IV
Mean Absolute Error (m) of the target 3D location estimates on single speaker sequences of the AV16.3 dataset: v-SSL (video-based SSL): image-to-3D projection from face detections [38]; va-SSL (video-assisted SSL): GCF on video-assisted speaker height plane.

	seq01	seq02	seq03	seq08	seq11	seq12	seq15
v-SSL	0.177	0.183	0.117	0.155	0.331	0.197	0.192
va-SSL	0.167	0.174	0.105	0.124	0.236	0.196	0.115

video observations, or the ID switches when targets overlap on the image. Compared to [24], improvements from the AV3T-VDM likelihood are mainly because of the 3D multiple concurrent speakers localization for acoustic discriminative likelihood modelling. In this case, acoustic information from the non-dominant speakers can compensate for the 3D video observation inaccuracies. A graphical example is depicted in Fig. 5(a) where the audio observation (cyan diamond) is closer to the ground truth (yellow cross) than the video observation (cyan square). Besides, the new acoustic discriminative likelihood may contribute a reliable 3D audio observation also when an active speaker is outside the camera's FoV, such as in Fig. 5(b). In this case, the detected audio observation may trigger a new track for an appearing target or re-activate a de-activated track which has stayed outside the camera's FoV for a long period[3].

D. Ablation study on the AV3T-VDM components

Compared to the audio-only SSL (*i.e.* GCF [53]), the *video-assisted localization* method brings several benefits [24], such as (i) reducing the searching space from 3D to a 2D plane parallel to the ground for an improved localization accuracy and (ii) eliminating the TDoA ambiguity resulting from the symmetric configuration of the circular microphone array. The improvements provided by the video-assisted audio localization are evident in Tab. IV, where the proposed method

[3]video sample: https://tinyurl.com/y4q8r8of

TABLE V
Influence of each likelihood component in the proposed AV3T-VDM tracking framework. '✓' indicates the likelihood is used. The audio-visual discriminative likelihood $L^{\mathcal{AV},D}$ is decomposed into audio ($d \in D_t^{\mathcal{A}}$) and video parts ($d \in D_t^{\mathcal{V}}$), respectively. Results are given as the mean and standard deviations (the number after '±') over 10 iterations.

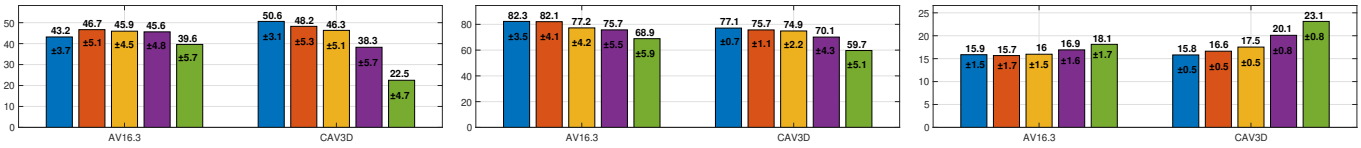
		$L^{\mathcal{AV},D},\ d \in D_t^{\mathcal{A}}$	$L^{\mathcal{AV},D},\ d \in D_t^{\mathcal{V}}$	$L^{\mathcal{A},G}$	$L^{\mathcal{V},G}$	MOTA$_{3D}$ ↑	MOTA$_{img}$ ↑	OSPAT↓
AV16.3	1	✓	✓	✓	✓	**43.2**±3.7	**82.3**±3.5	**15.9**±1.5
	2	✓	✓			30.3±3.7	72.4±3.2	19.8±1.2
	3		✓	✓		29.7±6.9	70.5±6.9	18.2±1.9
	4	✓			✓	39.4±5.9	74.7±4.1	17.8±2.0
	5			✓	✓	29.6±7.9	71.7±7.9	18.5±2.3
CAV3D	1	✓	✓	✓	✓	**50.6**±3.1	**77.1**±0.7	**15.8**±0.5
	2	✓	✓			49.8±2.2	76.7±1.6	15.9±0.5
	3		✓	✓		43.7±3.5	76.5±2.0	15.6±0.7
	4	✓			✓	23.6±6.9	76.8±2.3	22.0±1.0
	5			✓	✓	-4.2±12.6	69.5±6.8	23.4±1.5

is compared against the video-based localization (image-to-3D projection from detected faces, Eq.1) in terms of MAE on single-speaker sequences in the AV16.3 dataset. Note that we use single-speaker sequences to disentangle other factors which affect the overall performance and are dealt with by the tracking framework.

We evaluate the influence of each likelihood component on AV3T-VDM by removing the others from the tracking framework. Tab. V shows the comparison results where '✓' means the likelihood is used. For example, in the 1^{st} row of both datasets, '✓' is placed under $L^{\mathcal{A},D}, L^{\mathcal{A},G}, L^{\mathcal{AV},D}(d \in D_t^{\mathcal{A}}, D_t^{\mathcal{V}})$, this indicates all the likelihood models are used for tracking. From the comparison results, we can see that the discriminative model (Eq. 20) is significant for a good 3D tracking accuracy: when removing it from the framework (4^{th} row, which only uses the generative models), MOTA$_{3D}$ gets the lowest value among all the others. This is because when occlusion happens, the visual generative likelihood either fails to recover the trajectory of the occluded target or swaps the targets' identity labels. For the audio generative likelihood, when targets are close in 3D, the de-emphasized acoustic map may produce two nearby peaks, which either results in track overlap or identity swap. Even though the generative models are not as crucial as the discriminative ones, they still bring improvements, especially during short-term absence of the discriminative models (no 3D observations). A video sample is provided[4]. In summary, by letting generative and discriminative multi-modal likelihoods interact, uni-modal weakness is compensated for the best tracking performance both in 3D and on the image plane.

Since the face detection plays an important role in AV3T-VDM to generate the 3D video observations and to assist the acoustic map computation, we evaluate the influence of the face detector. Fig. 6 shows the variations of the tracking results when we randomly remove 0%, 20%, 40%, 60% and 80% detections from the tracking framework. The mean values over 10 repetitions are listed on top of each bar while the standard deviations are indicated after the symbol "±". The performance on the image plane (*i.e.* MOTA$_{img}$ and OSPAT)

[4]video sample: https://tinyurl.com/y2nn6azu



(a) MOTA$_{3D}$ (b) MOTA$_{img}$ (c) OSPAT

Fig. 6. Influence of removing % of face detections on (a) MOTA$_{3D}$; (b) MOTA$_{img}$ and (c) OSPAT, in AV16.3 and CAV3D, respectively (The mean value over 10 repetitions are listed on top of each bar while the standard deviations are indicated after the symbol "±". Blue, red, orange, purple and green indicate removing 0%, 20%, 40%, 60%, 80% of the face detections).

TABLE VI
Tracking results on the AV16.3 dataset as the mean and standard deviations (std) over 10 iterations.

	OSPAT ↓				MOTA$_{img}$ ↑			MOTA$_{3D}$ ↑
seq-C	[60]	[10]	[11]	prop.	[10]	[11]	prop.	prop.
24-1	22.3	15.7	18.7	**11.2**	60.3	37.4	**89.4**	62.4
24-2	17.6	15.9	17.3	**15.3**	71.9	68.0	**91.0**	48.3
24-3	28.2	**14.5**	15.6	17.3	61.0	53.3	**91.0**	39.8
25-1	21.5	16.4	18.1	**11.7**	43.0	39.6	**94.8**	58.6
25-2	19.2	14.8	15.8	**13.9**	65.6	54.5	**85.9**	49.4
25-3	29.4	15.6	18.1	**12.6**	66.6	57.7	**91.2**	50.6
30-1	36.0	**18.8**	22.5	19.5	**46.5**	41.4	40.8	7.2
30-2	28.4	**16.2**	18.2	19.3	60.3	54.1	**82.7**	37.3
30-3	34.6	20.7	20.9	**20.0**	49.3	47.1	**64.7**	27.5
45-1	NA	23.4	23.9	**15.7**	29.5	32.5	**87.5**	53.4
45-2	NA	21.8	23.8	**16.6**	49.4	41.2	**88.3**	55.7
45-3	NA	23.7	22.1	**17.4**	44.4	48.6	**80.1**	28.0
mean	26.3	18.1	19.6	**15.9**	54.0	48.0	**82.3**	43.2
±std	NA	±0.6	±0.4	±1.5	±3.0	±2.1	±3.5	±3.7

TABLE VII
Tracking results on the CAV3D dataset as the mean and standard deviations (std) over 10 iterations..

	OSPAT ↓		MOTA$_{img}$ ↑		MOTA$_{3D}$ ↑
seq	[11]	prop.	[11]	prop.	prop.
22	28.5	**13.2**	52.4	**97.4**	69.0
23	19.7	**14.7**	**96.1**	70.0	53.7
24	17.9	**12.8**	45.8	**64.0**	21.2
25	28.4	**20.7**	47.7	**82.6**	39.6
26	27.4	**17.6**	38.4	**71.5**	69.5
mean	24.4	**15.8**	56.1	**77.1**	50.6
±std	±1.1	±0.5	±7.9	±0.7	±3.1

drops gradually with more face detection removal. MOTA$_{3D}$ has different variation tendencies on the AV16.3 and the CAV3D dataset. For the CAV3D dataset, MOTA$_{3D}$ has a significant decrease with a larger face removal percentage. However, for the AV16.3 dataset, MOTA$_{3D}$ is stable unless 80% of detections are removed. This difference results from the more challenging acoustic environment in CAV3D, which includes: stronger room reverberation, more frequent non-direct head orientations and larger distances of the speakers from the sensing platform. In this case, when we remove face detections, the 3D audio observations are less accurate than the ones in AV16.3, leading to different MOTA$_{3D}$ performance in the two datasets.

E. Comparisons with state-of-the-art methods

Tab. VI and Tab. VII display the tracking results of different methods [10, 11, 60] on the AV16.3 and the CAV3D dataset. [60] is a visual SoA multi-speaker tracking method without audio usage, we report their results here to emphasize the benefits of audio-visual fusion. Results of [10] and [11] are generated by running their provided source code[5] where we optimize the parameters on the training sequences.

Compared to the SoA trackers performed on the 2D image plane [10, 11, 60], the proposed AV3T-VDM algorithm considers a higher tracking dimension *i.e.* 3D, which introduces more uncertainties in multi-modal processing. Even though, the comparison results still demonstrate clear improvements in both datasets. This is mainly due to the fact that in the three SoA methods, the audio modality only plays an assistant role to guide the visual tracker. Thus, the performance drops significantly during video failure, such as unfavourable targets' orientations or targets outside the camera's FoV, or when the audio observation is unreliable. Differently, in our AV3T-VDM pipeline, audio is also guided by the video and it is still processed independently during the absence of video information. For example, when a target talks outside the camera's FoV, AV3T-VDM updates the likelihood using the acoustic generative model, as well as the discriminative model if a 3D acoustic observation is validated (Fig. 5(b)). Besides, since AV3T-VDM adopts 3D target tracking, apart from the image plane information, the target-sensor distance information can also be used to improve the tracking performance.

F. Limitations

Limitations of AV3T-VDM include (i) dealing with reflected sounds that generate consistent FP observations, which may result in an incorrect new-born track when a speaker faces a wall and (ii) target re-identification after a long absence of a speaker from the FOV of the camera, especially when the target re-enters with a non-frontal view[6]. In this case, the spatiogram component of the association distance may fail because the reference image, updated with the latest face detection, is less discriminative as the target is captured from very different viewpoints.

VII. Conclusion

We presented AV3T-VDM, a particle filtering framework using audio-visual signals captured from a small-size sensing platform to perform multiple target tracking in 3D. AV3T-VDM relies on a video-assisted de-emphasized acoustic

[5]http://personal.ee.surrey.ac.uk/Personal/W.Wang/codes/AV_A_PF.html
[6]video sample: https://tinyurl.com/yyvau6zg

map to localize multiple concurrent speakers in 3D and on a joint multi-modal discriminative and generative likelihood modeling. A distinctive feature of AV3T-VDM is its track management with cross-modal cues to handle an unknown number of targets. AV3T-VDM can track multiple concurrent speakers with good tracking performance both in 3D and on the image plane.

Directions for future work include the adaptation of the 3D face diagonal size to reduce inaccuracies in the image-to-3D face detection projection and the inclusion of biometric acoustic features of each individual to enrich the audio likelihood modeling.

REFERENCES

[1] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *American Society for Engineering Education Zone Conf. Proc.*, Pittsburgh, PA, USA, Jun 2008, pp. 1–7.

[2] S. M. Naqvi, W. Wang, M. S. Khan, M. Barnard, and J. Chambers, "Multimodal (audiovisual) source separation exploiting multi-speaker tracking, robust beamforming and time-frequency masking," *IET Signal Processing*, vol. 6, no. 5, pp. 466–477, Sept 2012.

[3] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 895–910, Jun 2010.

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

[5] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[6] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1327–1339, Apr 2007.

[7] S. Lin, "Robust pitch estimation and tracking for speakers based on subband encoding and the generalized labeled multi-bernoulli filter," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 27, no. 4, pp. 827–841, Apr 2019.

[8] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 21, no. 10, pp. 2193–2206, Jul 2013.

[9] S. Lin, "Reverberation-robust localization of speakers using distinct speech onsets and multichannel cross correlations," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 26, no. 11, pp. 2098–2111, Jul 2018.

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

[11] 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. on Multimedia*, vol. 18, no. 12, pp. 2417–2431, Aug 2016.

[12] H. Zhou, M. Taj, and A. Cavallaro, "Target detection and tracking with heterogeneous sensors," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 4, pp. 503–513, Sept 2008.

[13] Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *Int. Conf. on Computer Vision Workshop on Computer Vision for Audio-Visual Media*, Venice, Italy, Oct 2017, pp. 446–454.

[14] Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, "Variational Bayesian inference for audio-visual tracking of multiple speakers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Nov 2019.

[15] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1086–1099, Jan 2017.

[16] E. D'Arca, N. M. Robertson, and J. Hopgood, "Person tracking via audio and video fusion," in *Data Fusion and Target Tracking Conf.: Algorithms and Applications*, London,UK, May 2012, pp. 1–6.

[17] M. Taj and A. Cavallaro, "Audio-assisted trajectory estimation in non-overlapping multi-camera networks," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, Taipei, Taiwan, Apr 2009, pp. 3517–3520.

[18] U. Kirchmaier, S. Hawe, and K. Diepold, "Dynamical information fusion of heterogeneous sensors for 3D tracking using particle swarm optimization," *Information Fusion*, vol. 12, no. 4, pp. 275–283, Oct 2011.

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

[20] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 1, pp. 1154–1164, Dec 2002.

[21] A. Brutti and O. Lanz, "A joint particle filter to track the position and head orientation of people using audio visual cues," in *Proc. of European Signal Processing Conf.*, Aalborg, Denmark, Aug 2010, pp. 974–978.

[22] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 38, no. 3, pp. 799–807, Dec 2008.

[23] R. Brunelli, A. Brutti, P. Chippendale, O. Lanz, M. Omologo, P. Svaizer, and F. Tobia, "A generative approach to audio-visual person tracking," in *Int. Evaluation Workshop on Classification of Events, Activities and Relationships*, Southampton, UK, Apr 2006, pp. 55–68.

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

[25] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey," *Proc. of the IEEE*, vol. 98, no. 10, pp. 1692–1715, Aug 2010.

[26] W. Luo, B. Yang, and R. Urtasun, "Fast and furious: Real time end-to-end 3D detection, tracking and motion forecasting with a single convolutional net," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, Jun 2018, pp. 3569–3577.

[27] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals using stereo imagery for accurate object class detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1259–1272, May 2017.

[28] Q. Yu, T. B. Dinh, and G. Medioni, "Online tracking and reacquisition using co-trained generative and discriminative trackers," in *Proc. of European Conf. on Computer Vision*, Marseille, France, Oct 2008, pp. 678–691.

[29] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neurocomput.*, vol. 74, no. 18, pp. 3823–3831, Nov 2011.

[30] Y. Lei, X. Ding, and S. Wang, "Visual tracker using sequential bayesian learning: Discriminative, generative, and hybrid," *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 6, pp. 1578–1591, Oct 2008.

[31] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, vol. 2, Hilton Head Island, SC, USA, Jun 2000, pp. 142–149.
[32] Y. Xie, W. Zhang, C. Li, S. Lin, Y. Qu, and Y. Zhang, "Discriminative object tracking via sparse representation and online dictionary learning," *IEEE Trans. on cybernetics*, vol. 44, no. 4, pp. 539–553, Jun 2013.
[33] C. Sun, F. Li, H. Lu, and G. Hua, "Visual tracking via joint discriminative appearance learning," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2567–2577, Dec 2017.
[34] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *Proc. of Int. Conf. on Computer Vision*, Kyoto, Japan, Sept 2009, pp. 1515–1522.
[35] L. Ting, W. Gang, and Y. Qingxiong, "Real-time part-based visual tracking via adaptive correlation filters," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, Jun 2015, pp. 4902–4912.
[36] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. of British Machine Vision Conf.*, Nottingham, UK, Sept 2014, pp. 1–11.
[37] X. Wu, R. He, and Z. Sun, "A lightened CNN for deep face representation with noisy labels," *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, Nov 2018.
[38] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, Jul 2017, pp. 1522–1530.
[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. of Int. Conf. on Neural Information Proc. Systems*, Montreal, Canada, Dec 2015, pp. 91–99.
[40] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, Jun 2014, pp. 580–587.
[41] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Proc. of Int. Conf. on Neural Information Proc. Systems*, Lake Tahoe, Nevada, USA, Dec 2013, pp. 809–817.
[42] Y. Bar-Shalom, P. K. Willett, and X. Tian, *Tracking and data fusion*. YBS publishing Storrs, CT, USA, 2011.
[43] M. Liggins II, D. Hall, and J. Llinas, *Handbook of multisensor data fusion: theory and practice*. CRC press, 2017.
[44] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society of Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, Mar 1957.
[45] K. Panta, B.-N. Vo, S. Singh, and A. Doucet, "Probability hypothesis density filter versus multiple hypothesis tracking," in *Defense and Security*, Orlando, FL, USA, Apr 2004, pp. 284–295.
[46] X. Alameda-Pineda, S. Arias, Y. Ban, G. Delorme, L. Girin, R. Horaud, X. Li, B. Morgue, and G. Sarrazin, "Audio-visual variational fusion for multi-person tracking with robots," in *Proc. of ACM Int. Conf. on Multimedia*, Nice, France, Oct 2019, p. 1059–1061.
[47] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 88–103, Mar 2019.
[48] E. Maggio and A. Cavallaro, "Hybrid particle filter and mean shift tracker with adaptive transition model," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, Philadelphia, PA, USA, Mar 2005, pp. 221–224.
[49] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.
[50] Q. Liu, W. Wang, d. C. Teófilo, P. Jackson, and A. Hilton, "Multiple speaker tracking in spatial audio via PHD filtering and depth-audio fusion," *IEEE Trans. on Multimedia*, vol. 20, no. 7, pp. 1767–1780, Nov 2017.
[51] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug 1976.
[52] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, May 1997.
[53] M. Omologo, P. Svaizer, and R. De Mori, "Acoustic transduction," in *Spoken Dialogue with Computer*. Academic Press, 1998, ch. 2, pp. 1–46.
[54] X. Qian, A. Brutti, M. Omologo, and A. Cavallaro, "3D audio-visual speaker tracking with an adaptive particle filter," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, New Orleans, LA, USA, Mar 2017, pp. 2896–2900.
[55] I. Marković and I. Petrović, "Speaker localization and tracking with a microphone array on a mobile robot using von mises distribution and particle filtering," *Robotics and Autonomous Systems*, vol. 58, no. 11, pp. 1185–1196, Nov 2010.
[56] P. Smaragdis and P. Boufounos, "Position and trajectory learning for microphone arrays," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 358–368, Dec 2006.
[57] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, pp. 1–17, Jan 2010.
[58] S. T. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region-based tracking," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, Jun 2005, pp. 1158–1163.
[59] J. Vermaak, S. J. Godsill, and P. Perez, "Monte carlo filtering for multi target tracking and data association," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 41, no. 1, pp. 309–332, Jan 2005.
[60] 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. on Multimedia*, vol. 16, no. 3, pp. 864–880, Apr 2014.
[61] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16. 3: an audio-visual corpus for speaker localization and tracking," in *Machine Learning for Multimodal Interaction*. Martigny, Switzerland: Springer, Jun 2004, pp. 182–195.
[62] R. Stiefelhagen and J. Garofolo, *Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities and Relationships (CLEAR)*. Berlin, Heidelberg: Springer-Verlag, 2007.
[63] B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Trans. on Signal Processing*, vol. 59, no. 7, pp. 3452–3457, Jul 2011.
[64] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, Jul 2008.