Audio-Visual SLAM towards Human Tracking and Human-Robot Interaction in Indoor Environments

Aaron Chau, Kouhei Sekiguchi, Aditya Arie Nugraha, Kazuyoshi Yoshii, Kotaro Funakoshi

***Abstract*— We propose a novel audio-visual simultaneous and localization (SLAM) framework that exploits human pose and acoustic speech of human sound sources to allow a robot equipped with a microphone array and a monocular camera to track, map, and interact with human partners in an indoor environment. Since human interaction is characterized by features perceived in not only the visual modality, but the acoustic modality as well, SLAM systems must utilize information from both modalities. Using a state-of-the-art beamforming technique, we obtain sound components correspondent to speech and noise; and estimate the Direction-of-Arrival (DoA) estimates of active sound sources as useful representations of observed features in the acoustic modality. Through estimated human pose by a monocular camera, we obtain the relative positions of humans as representation of observed features in the visual modality. Using these techniques, we attempt to eliminate restrictions imposed by intermittent speech, noisy periods, reverberant periods, triangulation of sound-source range, and limited visual field-of-views; and subsequently perform early fusion on these representations. We develop a system that allows for complimentary action between audio-visual sensor modalities in the simultaneous mapping of multiple human sound sources and the localization of observer position.**

I. Introduction

Human interaction is inherently characterized by features dominant in the acoustic and visual modalities. The synergistic use of audio and visual information allows systems in human-robot interaction to exploit spatial information determined from speech and movement for human tracking.

Simultaneous Localization and Mapping (SLAM) has been actively investigated to enable mobile robots to localize themselves in an environment while they concurrently map landmarks of interest from the surrounding area. SLAM provides benefit in the creation of live maps populated by landmarks of interest, and also allows robots to gain representations of self-location in the environment they move within. In the classical context of human localization algorithms, the visual modality is often used as the most dominant modality for estimating target positions. However, humans speak, listen, and move when they communicate and interact with objects. As such, the acoustic modality also plays an important role in localizing humans from acoustic and visual data provided during interactions.

Visual and acoustic features have a complementary relationship in the SLAM problem. Classical SLAM algorithms reliant on the visual modality provide robust information of the physical environment and can be decomposed into parametric, feature-based representations that directly lead into estimation of Cartesian target positions. Such Visual SLAM however, is limited by narrow sensor field-of-views (FoV), feature occlusions, and optical degradations, especially in the case of monocular vision. On the other hand, Acoustic SLAM is traditionally implemented by sole use of acoustic signal. Acoustic SLAM inherently possesses a full sensor FoV, and provides the important ability to localize targets based off speech interaction; however, it requires inference of Cartesian target positions from acquired audio features and introduces new susceptibility to environmental reverberance and noise. Features present in the acoustic modality can be used to supplement Visual SLAM's disadvantages in narrow FoV and optical degradations, whereas visual-based features compensate for Acoustic SLAM's undetermined estimation of Cartesian target positions and acoustic degradation.

In this paper, we propose the Audio-Visual SLAM (AV-SLAM) framework that exploits features extracted from both the acoustic and the visual modalities for human-robot interaction in indoor environments. We built this approach upon the existing theoretical bases laid out by the Acoustic SLAM (aSLAM) [1] and the multitarget tracking methods of Random Finite Sets (RFSs) [2]–[4]. The RFSs allow the AV-SLAM to map and track multiple intermittent targets distinguished by sound sources or human targets captured in the acoustic and visual FoVs, respectively. We also propose a robust feature acquisition pipeline performed prior to AV-SLAM. The full proposed framework consists of the following components:

1) *Audio feature extraction* by the deep neural network (DNN) based spectral mask estimation [5] followed by the GSVD-MUSIC [6] for a noise-resilient estimation of the direction of arrival (DoA) of sound sources.
2) *Visual feature extraction* by estimating the relative Cartesian target positions given the keypoints provided by the OpenPose [7], [8], which is a state-of-the-art human pose estimator, using monocular vision.
3) *Acoustic-visual modality reliability measures* based on an approximated Signal-to-Noise Ratio (SNR) and human confidence values.
4) *Early fusion* by a multi-stream standard-product weighting approach to the acoustic and visual features.

This paper was supported by by the Cooperative Intelligence Joint Research Chair with Honda Research Institute Japan Co., Ltd. and JSPS KAKENHI No. 19H04137.

A. Chau is with the Schulich School of Engineering, University of Calgary, Calgary, T2N 1N4, Canada. aaron.chau@ucalgary.ca

K. Sekiguchi, A. A. Nugraha, and K. Yoshii are with the Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan. adityaarie.nugraha@riken.jp

K. Sekiguchi, K. Yoshii, and K. Funakoshi are additionally with the Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan. sekiguch@kuis.kyoto-u.ac.jp, {yoshii, funakoshi.k}@i.kyoto-u.ac.jp

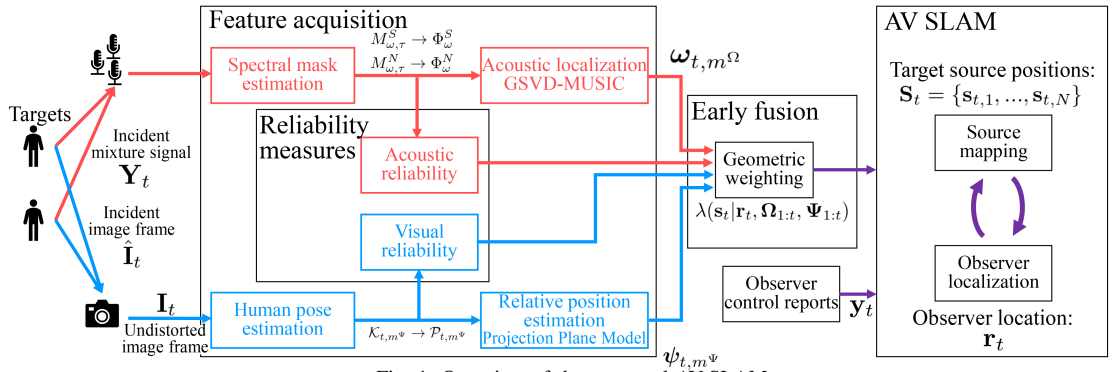


Fig. 1: Overview of the proposed AV-SLAM system.

II. Proposed Framework

The AV-SLAM framework is summarized in Fig. 1. Pertinent components of the framework are explained as follows.

In this paper, we represent all spatial quantities in respect to the coordinate frame of Fig. 2. The AV-SLAM targets are assumed to be of human nature and assumed to emit human speech. Given the observed DoAs $\boldsymbol{\omega}_t$, the observed position $\boldsymbol{\psi}_t$ relative to the observer (robot), and the observer control report $\mathbf{y}_t$, we want to locate and track the observer $\mathbf{r}_t$ and the sources $\mathbf{S}_t$. The state space model is shown in Fig. 3.

Inputs to Feature Acquisition consist of observations from the different modalities:

- Acoustic: $\mathbb{M}$-channel input audio source signal, $\mathbf{Y}_t$,
- Visual: $\mathbb{I} \times \mathbb{J} \times 3$ matrix of monocular RGB-video, $\mathbf{I}_t$.

Inputs to the AV-SLAM are characterized by three input variables correspondent to the observed states shown in Fig. 3. The Acoustic and Visual observations are derived from incident observed variables in the Feature Acquisition processing step. And in standard practice of SLAM algorithms, the control report, $\mathbf{y}(t)$, is acquired by various odometry sensors in the robot platform [9].

- Acoustic input: $\boldsymbol{\omega}_{t,m^\Omega}$, is the time-varying DoA estimation vector of, $m^\Omega = 1, \ldots, M_t^\Omega$, estimations in the spherical coordinate system, defined in Eq. (7).
- Visual input: $\boldsymbol{\psi}_{t,m^\Psi}$, is the time-varying relative target position (RTP) estimation vector of, $m^\Psi = 1, \ldots, M_t^\Psi$, estimations in the cylindrical coordinate system, defined in Eq. (11).
- Control input: $\mathbf{y}(t) \triangleq [y_{t,\nu}, y_{t,\gamma}]^T$, where, $y_{t,\nu}$, and, $y_{t,\gamma}$, are the control speed and control orientation respectively. In this formulation, the source motion is assumed to be constrained to the X-Y Plane of Fig. 2, i.e. control speed is in the direction of control orientation.

Outputs of the AV-SLAM framework are given by:

- Observer positional state: Denoted with vector: $\mathbf{r}_t \triangleq [x_t, y_t, z_t, \gamma_t]^T$, where, (x_t, y_t, z_t) , denote Cartesian position, and, γ_t, denotes observer orientation.
- Relative Target Positional States: Denoted with: $\boldsymbol{s}_{t,n} \triangleq [x_{t,n}, y_{t,n}, z_{t,n}]$, where, $n = 1, \ldots, N_t$, is the number of estimated sources.

Inputs and outputs to the AV-SLAM framework are represented at each time point, t.

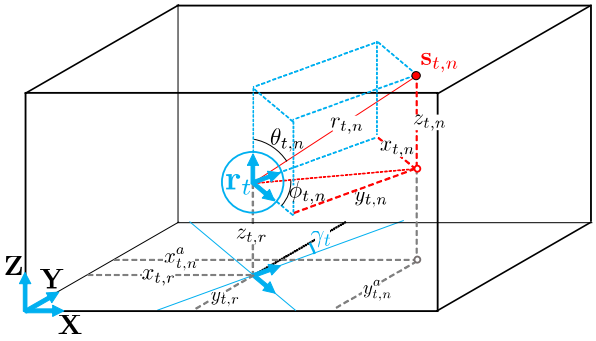


Fig. 2: Coordinate frame.

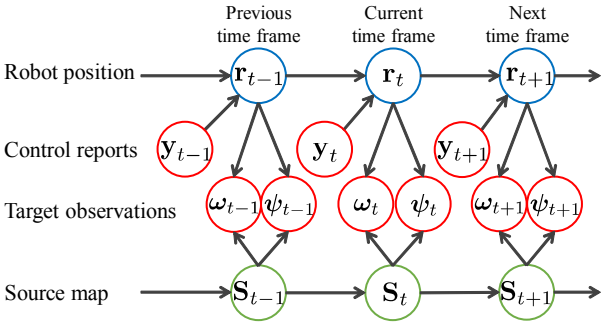


Fig. 3: State space model.

Targets in the environment are expected to be intermittent in presence due to discontinuous speech in the acoustic modality and unseen targets in the visual modality. As such, the number of targets N_t is time-varying and unknown [1], [3]. Akin to aSLAM, the number of sources and the corresponding states are modelled by a RFS:

$$\mathbf{S}_t = \left[\bigcup_{n=1}^{N_t - 1} P(\mathbf{s}_{t-1,n}) \right] \cup B_t, \tag{1a}$$

$$P(\mathbf{s}_{t-1,n}) = \begin{cases} \{\boldsymbol{s}_{t,n}\} & \text{if source } n \text{ was active,} \\ \emptyset & \text{otherwise,} \end{cases} \tag{1b}$$

where B_t is a point process that models newborn sources.

III. Proposed Method

A. Feature Acquisition

The DoAs, $\boldsymbol{\omega}_t$ and the relative target positions, $\boldsymbol{\psi}_t$ extracted from the acoustic and the visual modalities, respectively, and modeled as follows.

1) Acoustic Modality: The DoA estimates are obtained by the Generalized Singular Value Decomposition based Multiple Signal Classification (GSVD-MUSIC) [6], [10]. We assume that pre-recorded, microphone-array-specific room transfer functions (RTFs) are available for a number of DoAs (ϕ, θ). To estimate the required power spectral density (PSD) of unwanted signals, we employ a deep neural network based spectral mask estimation [5]. This approach works on the time-frequency domain. We extract the short-time Fourier transform (STFT) representation $\mathbf{Y}_{\omega,\tau} \in \mathbb{C}^{M\times 1}$, where ω is the angular frequency and τ is the time frame index, from the audio input $\mathbf{Y}_t$. Since the audio sampling frequency is much higher than the AV-SLAM time step frequency, we obtain multiple frames T of STFT coefficients for each t.

Given $\mathbf{Y}_{\omega,\tau}$, the noise PSD $\mathbf{\Phi}^{\mathrm{N}}_{\omega} \in \mathbb{C}^{M\times M}$ is computed as

$$\mathbf{\Phi}^{N}_{\omega} = \sum_{\tau=1}^{\mathrm{T}} M^{\mathrm{N}}_{\omega,\tau} \mathbf{Y}_{\omega,\tau} \mathbf{Y}^{\star}_{\omega,\tau}, \tag{2}$$

where $M^{\mathrm{N}}_{\omega,\tau}$ is the noise mask estimate and $\cdot^{\star}$ is the conjugate transposition. The MUSIC spectrogram is then computed as

$$p_{\phi,\theta} = \frac{1}{\omega_u - \omega_l + 1} \sum_{\omega=\omega_l}^{\omega_u} \frac{\mathbf{a}^{\star}_{\omega,\phi,\theta} \mathbf{a}_{\omega,\phi,\theta}}{\mathbf{a}^{\star}_{\omega,\phi,\theta} \mathbf{E}^{\mathrm{N}}_{\omega} \mathbf{E}^{\mathrm{N}\star}_{\omega} \mathbf{a}_{\omega,\phi,\theta}}. \tag{3}$$

where ω_l and ω_u denote the lower and the upper bounds, respectively, $\mathbf{a}_{\omega,\phi,\theta} \in \mathbb{C}^{M\times 1}$ is a room transfer function (RTF) and $\mathbf{E}^{\mathrm{N}}_{\omega} \in \mathbb{C}^{M\times M}$ are the left-singular vectors obtained by applying GSVD on $\mathbf{\Phi}^{\mathrm{N}}_{\omega}$. Finally, we pick a set of DoA estimates $[\phi_{t,m^{\Omega}}, \theta_{t,m^{\Omega}}]$ for $m^{\Omega} \in \{1, \ldots, M^{\Omega}_t\}$, used to populate a full DoA estimate expressed by:

$$\boldsymbol{\omega}_{t,m^{\Omega}} = [\rho_{t,m_{\Omega}}, \phi_{t,m^{\Omega}}, \theta_{t,m^{\Omega}}]^{\top}, \tag{4}$$

where, $\rho_{t,m^{\Omega}}$ is an initialized range hypothesis discussed later, and M^{Ω}_t is the pre-specified number of DoA estimates that maximize $p_{\phi,\theta}$.

Following the aSLAM [1], we model the DoA estimates using a RFS:

$$\mathbf{\Omega}_t = \left[\bigcup_{n=1}^{N_t} D^{\Omega}(\boldsymbol{s}_{t,n}) \right] \cup K^{\Omega}_t, \tag{5}$$

where K^{Ω}_t is a Poisson point process modeling the false DoA estimates [4] and $D^{\Omega}(\boldsymbol{s}_{t,n})$ represents the detection process:

$$D^{\Omega}(\boldsymbol{s}_{t,n}) = \begin{cases} \{\boldsymbol{\omega}_{t,m^{\Omega}}\} & \text{if source } n \text{ is detected,} \\ \emptyset & \text{if source } n \text{ is not detected.} \end{cases} \tag{6}$$

The relation between the DoA estimate and the target positional state $\mathbf{s}_{t,n}$ is expressed as

$$\boldsymbol{\omega}_{t,m^{\Omega}} = \hat{\vartheta}\left(g(\mathbf{s}_{t,n}) + \mathbf{e}_{t,m^{\Omega}}\right). \tag{7}$$

where $g(\cdot)$ is the Cartesian-to-spherical transformation, $\mathbf{e}_{t,m^{\Omega}} \sim \mathcal{N}(\mathbf{0}_{2\times 1}, \mathbf{R}^{\Omega}_{t,m^{\Omega}})$ represents the estimation error, and $\hat{\vartheta}(\cdot)$ wraps a DoA estimate so that $\phi_{t,m^{\Omega}} \in [0, 2\pi)$ and $\theta_{t,m^{\Omega}} \in [0, \pi]$.

2) Visual Modality: The relative target positions (RTPs) are estimated through the combined use of the OpenPose [7], [8] and the target projection plane model [11].

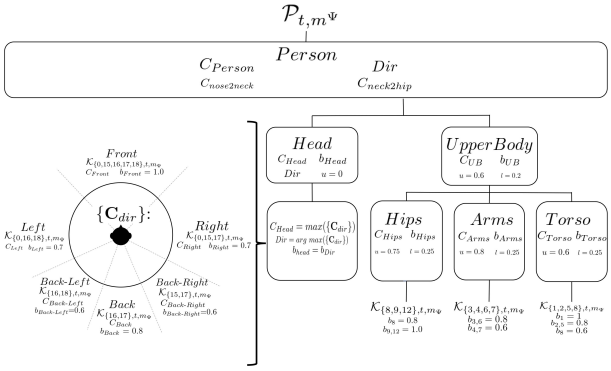


Fig. 4: Person confidence decision tree.

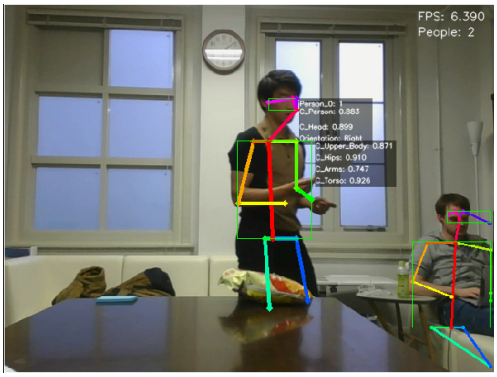


Fig. 5: Person confidence gui.

We use OpenPose for its availability and applicability towards real-time, multi-target, pose estimation in monocular image frames. From our input image, we receive 25 keypoints based on the COCO dataset, extended for BODY+FOOT estimation [7], [12], denoted by matrix $\boldsymbol{\mathcal{K}}_{t,m^{\Psi}}$. To introduce resilience towards partial body occlusions, we apply the decision tree shown in Fig. 4 to group keypoints together into part confidence and head orientation measures, denoted by matrix $\boldsymbol{\mathcal{P}}_{t,m^{\Psi}}$. Individual keypoint confidences are merged by a weighted average scheme, with heuristically defined weights unique to desired parts and merged keypoints.

From $\boldsymbol{\mathcal{P}}_{t,m^{\Psi}}$, we take two potential target planes for use in the target projection plane model: *nose2neck* or *neck2hip*, to enforce that visual estimates are obtained even in event of partial occlusion. The *neck2hip* target plane is applied in the event that the *Head* and *Upper Body* confidences are high. Otherwise, we default to the use of the *nose2neck* target plane; so that estimates are still obtained in event of poor confidence values. With known prior knowledge of physical target plane length L_0, pixel target plane length l^0_f, initialized recording distance d_0, and human height h_0, the RTPs are then estimated using the projection plane model as in [11]. The RTP estimates are expressed as

$$\boldsymbol{\psi}_{t,m^{\Psi}} = [d_{t,m^{\Psi}}, h_{t,m^{\Psi}}, \upsilon_{t,m^{\Psi}}]^{\top}, \tag{8}$$

where, $d_{t,m^{\Psi}}$, $h_{t,m^{\Psi}}$, and $\upsilon_{t,m^{\Psi}}$, are the target depth, height, and yaw, respectively, relative to the observer.

Similarly for the DoA estimates, we model the RTP esti-

mates using a RFS:

$$\mathbf{\Psi}_t = \left[\bigcup_{n=1}^{N_t} D^{\Psi}(\boldsymbol{s}_{t,n}) \right] \cup K_t^{\Psi}, \tag{9}$$

where K_t^{Ψ} is a Poisson point process modeling the false RTP estimates and $D^{\Psi}(\boldsymbol{s}_{t,n})$ represents the detection process:

$$D^{\Psi}(\boldsymbol{s}_{t,n}) = \begin{cases} \{\boldsymbol{\psi}_{t,m^{\Psi}}\} & \text{if source } n \text{ is seen in FoV,} \\ \emptyset & \text{if no source is seen in FoV.} \end{cases} \tag{10}$$

The relation between RTP estimate and the target positional state $\mathbf{s}_{t,n}$ is expressed as

$$\boldsymbol{\psi}_{t,m^{\Psi}} = h(\boldsymbol{s}_{t,n}) + \boldsymbol{e}_{t,m^{\Psi}}, \tag{11}$$

where $h(\cdot)$ is the Cartesian-to-cylindrical transformation, $\boldsymbol{e}_{t,m^{\Psi}} \sim \mathcal{N}(\mathbf{0}^{3\times 1}, \mathbf{R}^{\Psi}_{t,m^{\Psi}})$ represents the estimation error, $m^{\Psi} \in \{1, \ldots, M_t^{\Psi}\}$ with M_t^{Ψ} is the number of RTP estimates at time point t. The RTP estimate values are constrained to be in the visual FoV, defined as a function of the horizontal and the vertical camera angle-of-views.

B. Modality Reliability Measures

AV-SLAM dynamically extracts measures of reliability for each modality to account for environmental changes that degrade DoA or RTP estimations.

1) Acoustic Modality - Approximated SNR: We select a single microphone channel $y^{\text{lead}}_{\omega,f} \in \mathbb{C}$ that is most closely oriented to the estimated DoA $\boldsymbol{\omega}_{t,m^{\Omega}}$. The reliability measure based on an approximated SNR is then computed as

$$\text{SNR}_{\text{approx},t} = \sum_{\omega=\omega_l}^{\omega_u} \frac{\sum_{\tau=1}^{\text{T}} M^{\text{S}}_{\omega,\tau} |y^{\text{lead}}_{\omega,f}|^2}{\sum_{\tau=1}^{\text{T}} M^{\text{N}}_{\omega,\tau} |y^{\text{lead}}_{\omega,f}|^2}, \tag{12}$$

where $M^{\text{S}}_{\omega,\tau}$ and $M^{\text{N}}_{\omega,\tau}$ are the speech and noise masks estimated as in Section III-A.1.

2) Visual Modality - Generalized Human Confidence: Successive weighted arithmetic means are performed in Fig. 4, in which, $\mathcal{K}_i$, denotes the pixel values of the i^{th} human pose keypoint, $b_i \in [0, 1]$, provides its associated weight, and, I, is the cardinality of the keypoint set.

$$C_{\text{part}} = \frac{\sum_{i=1}^{I} \mathcal{K}_i b_i}{\sum_{i=1}^{I} b_i}, \tag{13}$$

From the derived person confidence measures, $\mathcal{P}_t, m^{\Psi}$, we obtain a measure for generalized human confidence by maximum person confidence across RTP estimations at each timepoint. The maximum, as opposed to the mean of person confidences, are taken to account for situations in which human targets are occluded or located at the edge of visual FoV. If the mean of confidences are taken, then targets of poor confidence can incorrectly degrade otherwise positive visual conditions, as made evident by another possible target of high confidence at the same timepoint. The visual feature acquisition system is more prone to spurious measurements from targets of poor person confidence; high person confidence is indicative of a properly-functioning modality.

$$HC_t = \max\left(\mathcal{P}_{t,m^{\Psi}}\right). \tag{14}$$

3) Reliability Mapping: The reliability measures of both modalities are then mapped as $\alpha \leftarrow \text{map}(\text{SNR}_{\text{approx},t})$ and $\beta \leftarrow \text{map}(HC_t)$ so that $\alpha, \beta \in [0, 1]$ where 1 represents the highest reliability, and vice versa. We heuristically define the mapping function to be a simple linear function.

C. AV-SLAM Theoretical Foundations

The AV-SLAM posterior probability density function (PDF) is expressed as

$$\begin{aligned} &p(\mathbf{r}_t, \mathbf{S}_t | \boldsymbol{y}_{1:t}, \mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t}) \\ &\quad = p(\mathbf{r}_t | \boldsymbol{y}_{1:t}, \mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t})\, p(\mathbf{S}_t | \mathbf{r}_t, \mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t}), \end{aligned} \tag{15}$$

where $p(\mathbf{r}_t | \boldsymbol{y}_{1:t}, \mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t})$ is the observer posterior PDF and $p(\mathbf{S}_t | \mathbf{r}_t, \mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t})$ is the multi-source posterior PDF.

The multi-source posterior pdf can be propagated with Bayes' theorem, under the assumption of conditional independence between audio observations, $\mathbf{\Omega}_t$, and visual observations, $\mathbf{\Psi}_t$, conditioned on observer pose, $\mathbf{r}_t$, and source states, $\mathbf{S}_t$:

$$\begin{aligned} &p(\mathbf{S}_t | \mathbf{r}_t, \mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t}) \\ &= \frac{p(\mathbf{\Omega}_t | \mathbf{r}_t, \mathbf{S}_t)\, p(\mathbf{\Psi}_t | \mathbf{r}_t, \mathbf{S}_t)\, p(\mathbf{S}_t | \mathbf{r}_t, \mathbf{\Omega}_{1:t-1}, \mathbf{\Psi}_{1:t-1})}{\int p(\mathbf{\Omega}_t | \mathbf{r}_t, \mathbf{S}_t)\, p(\mathbf{\Psi}_t | \mathbf{r}_t, \mathbf{S}_t)\, p(\mathbf{S}_t | \mathbf{r}_t, \mathbf{\Omega}_{1:t-1}, \mathbf{\Psi}_{1:t-1})\, \mathrm{d}\mathbf{S}_t}. \end{aligned} \tag{16}$$

1) Posterior SLAM PHD: For simplicity in expression, the SLAM PDF uses, $\boldsymbol{X_t} \triangleq (\mathbf{r}_t, \mathbf{S}_t)$, and, $\mathbf{Z}_t \triangleq (\boldsymbol{y}_t, \mathbf{\Omega}_t, \mathbf{\Psi}_t)$, to represent the joint states and observations, respectively. The posterior SLAM PDF can then be written through Bayes' theorem as:

$$p(\mathbf{X}_t | \mathbf{Z}_{1:t}) = \frac{p(\mathbf{Z}_t | \mathbf{X}_t) p(\mathbf{X}_t | \mathbf{Z}_{1:t-1})}{\int p(\mathbf{Z}_t | \mathbf{X}_t) p(\mathbf{X}_t | \mathbf{Z}_{1:t-1}) \mathrm{d}\mathbf{X}_t}, \tag{17}$$

It is important to note that we assume the class-conditional independence of audio and visual HMM streams inherent to AV-SLAM as per results reported in human perception studies [13], [14].

$$p(\mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t} | \mathbf{X}_t) = p(\mathbf{\Omega}_{1:t} | \mathbf{X}_t) p(\mathbf{\Psi}_{1:t} | \mathbf{X}_t), \tag{18}$$

Application of the probability chain rule, and class-conditional independence of the audio and visual streams yields the following posterior SLAM PDF,

$$\begin{aligned} &p(\mathbf{X}_t | \mathbf{Z}_{1:t}) \\ &= \frac{p(\mathbf{r}_t | \boldsymbol{y}_{1:t})\, p(\mathbf{\Omega}_t | \mathbf{r}_t)\, p(\mathbf{\Psi}_t | \mathbf{r}_t)}{\int p(\mathbf{r}_t | \boldsymbol{y}_{1:t})\, p(\mathbf{\Omega}_t | \mathbf{r}_t)\, p(\mathbf{\Psi}_t | \mathbf{r}_t)\, \mathrm{d}\mathbf{r}_t}\, p(\mathbf{S}_t | \mathbf{r}_t, \mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t}). \end{aligned} \tag{19}$$

And through PHD approximation, the posterior PDF is represented as,

$$\begin{aligned} &\lambda(\mathbf{r}_t, \mathbf{S}_t | \boldsymbol{y}_{1:t}, \mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t}) \\ &= \frac{\mathcal{L}(\mathbf{\Omega}_t | \mathbf{r}_t)\, \mathcal{L}(\mathbf{\Psi}_t | \mathbf{r}_t)\, p(\mathbf{r}_t | \boldsymbol{y}_{1:t}, \mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t})}{\int \mathcal{L}(\mathbf{\Omega}_t | \mathbf{r}_t)\, \mathcal{L}(\mathbf{\Psi}_t | \mathbf{r}_t)\, p(\mathbf{r}_t | \boldsymbol{y}_{1:t}, \mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t})\, \mathrm{d}\mathbf{r}_t} \\ &\lambda(\boldsymbol{s}_t | \mathbf{r}_t, \mathbf{\Omega}_{1:t}, \mathbf{\Psi}_{1:t}), \end{aligned} \tag{20}$$

With estimation evidence given by,

$$\mathcal{L}(\boldsymbol{\mathcal{M}}_t|\mathbf{r}_t) \triangleq e^{-N_{t,c}-N_{t|t-1}} \prod_{m_{\boldsymbol{\mathcal{M}}}=1}^{M_{t,\boldsymbol{\mathcal{M}}}} \ell(\mathcal{O}|\mathbf{r}_t). \quad (21)$$

For simplicity in notation, the acoustic and visual modalities will be denoted by set, $\boldsymbol{\mathcal{M}} \in \{\boldsymbol{\Omega}, \boldsymbol{\Psi}\}$, with their corresponding, representative modality estimations given by, $\boldsymbol{\mathcal{O}} \in \{\boldsymbol{\omega}_{t,m_{\boldsymbol{\Omega}}}, \boldsymbol{\psi}_{t,m_{\boldsymbol{\Psi}}}\}$, for the remainder of this paper.

2) Multi-Source Posterior PHD: In contrast to the previous work by Evers and Naylor 2018, [1], we now present the multi-source posterior PHD to be additionally dependent on the RFS of visual estimates up until the current time point. To do this, we express the multi-source posterior PHD, $\lambda(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t}, \boldsymbol{\Psi}_{1:t})$, as the sum of the birth PHD, and a new term denoted the full existent PHD, $\lambda_e(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t}, \boldsymbol{\Psi}_{1:t})$.

$$\begin{aligned}&\lambda(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t}, \boldsymbol{\Psi}_{1:t})\\ &\quad= p_b\lambda_b(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_t, \boldsymbol{\Psi}_t) + \lambda_e(\mathbf{s}_t|\mathbf{r}_t)\\ &\qquad+ (1-p_d)\lambda(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t-1}, \boldsymbol{\Psi}_{1:t-1}), \end{aligned} \quad (22)$$

Applying conditional independence between audio and visual modalities for newborn targets yields,

$$\begin{aligned}&\lambda(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t}, \boldsymbol{\Psi}_{1:t})\\ &\quad= p_b^{\Omega}\lambda_b(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_t) + p_b^{\Psi}\lambda_b(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Psi}_t) + \lambda_e(\mathbf{s}_t|\mathbf{r}_t)\\ &\qquad+ (1-p_d)\lambda(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t-1}, \boldsymbol{\Psi}_{1:t-1}). \end{aligned} \quad (23)$$

$$p_b = p_b^{\Omega} + p_b^{\Psi}, \quad (24a)$$

$$p_d = p_d^{\Omega} p_d^{\Psi}. \quad (24b)$$

Where p_d, p_b, are the time-independent full-detection, and full-birth probabilities. p_d^{Ω}, p_d^{Ψ}, and, p_b^{Ω}, p_b^{Ψ}, are the detection and birth probabilities of each respective modality, and $\lambda_e(\mathbf{s}_t|\mathbf{r}_t)$, is the newly-introduced full existent PHD. The full-existent PHD is representative of all presently detected components in both modalities, and is found as a function of the conditionally independent acoustic and visual detection PHDs, by the standard product algorithm for multi-stream fusion presented in [15].

$$\lambda_e(\mathbf{s}_t|\mathbf{r}_t) = \frac{\lambda_{\Omega}^{\alpha}(\mathbf{s}_t|\mathbf{r}_t)\lambda_{\Psi}^{\beta}(\mathbf{s}_t|\mathbf{r}_t)}{\int(\lambda_{\Omega}^{\alpha}(\mathbf{s}_t|\mathbf{r}_t)\lambda_{\Psi}^{\beta}(\mathbf{s}_t|\mathbf{r}_t))} \quad (25)$$

Where, α, and, β, are the acoustic and visual stream weights.

D. AV-SLAM Target Mapping

The inclusion of the visual modality and its associated observations allows for the introduction of additional source estimate components in the Gaussian Mixture Model (GMM) realization of AV-SLAM. The newborn, existent, (and by derivative) detected, and missing PHD terms which comprise the multi-source posterior PHD shown in Eq. (22), are realized as GMMs, indicative of different steps in the GM-PHD and Extended Kalman Filter equations [16], [17].

1) Realized Birth PHD - Newborn Detections: In area without overlap of acoustic and visual FoVs, source states, $\mathbf{s}_{t,n}$, remain undetermined due to a lack of source-observer range values, $\mathbf{r}_{t,m^{\Omega}}$, provided in DoA estimations. In the previous work by acoustic-SLAM, the algorithm offers a solution to this underdetermined problem by means of temporal triangulation [1]. However, RTP estimates provided in the visual modality provide full source-observer range hypotheses, but can only be born and detected in the camera FoV. The source-observer range still remains unknown in the acoustic FoV. In AV-SLAM, we use the visual modality's RTP range estimates to help the convergence of range hypothesis.

The birth PHD can be modelled as a GMM of acoustic and video modalities. The audio model deals with the undetermined system by introducing components representative of, J_b, range hypotheses, whereas the video model is used to only account for errors in RTP estimation.

The representative Gaussians of the visual RTP estimates are given as,

$$\mathbf{m}_{t,m^{\Psi}} = h^{-1}(\hat{\psi}_{t,m^{\Psi}}) \sim \mathcal{N}^c(\boldsymbol{\psi}_{t,m^{\Psi}}, \mathbf{R}^{\Psi}_{t,m^{\Psi}}). \quad (26)$$

Where ,$\mathcal{N}^c$, is a regular, and wrapped normal distribution applied to the cylindrical coordinates of the RTP estimate, $\boldsymbol{\psi}_{t,m^{\Psi}}$. Regular normal distributions are applied to the depth, $d_{t,m^{\Psi}}$, and height, $h_{t,m^{\Psi}}$, whereas a wrapped normal distribution is applied to the yaw estimate. $\upsilon_{t,m^{\Psi}}$ of the RTP estimates. $h^{-1}(\cdot)$, is the cylindrical-to-Cartesian transformation, $\hat{\boldsymbol{\psi}}_{t,m^{\Psi}}$, is the Gaussian of the RTP estimates, and $\mathbf{R}_{t,m^{\Psi}}$, is the associated covariance matrix, assumed known *a priori*.

Conversely, the acoustic modality incorporates J_b range hypotheses, $\hat{r}^{(j)}_{t,m}$, set along each DoA detection, $\omega_{t,m}$ for all $m^{\Omega} = 1, \ldots, M_t^{\Omega}$ as per acoustic-SLAM [1].

$$\hat{r}^{(j)}_{t,m^{\Omega}} \sim \mathcal{U}(r_{min}, r_{max}), \quad (27)$$

Where, r_{min}, and r_{max}, are the minimum, and maximum of source-observer range hypotheses respectively.

Errors in newborn DoA estimations are then born by J_b hypotheses of, $\hat{\boldsymbol{\omega}}^{(j)}_{t,m^{\Omega}}$ as modelled by a wrapped Gaussian distribution around a unit sphere, which allows for the initialization of source states observed by the acoustic modality, $\mathbf{m}^{(j)}_{b,t,m^{\Omega}}$.

$$\hat{\boldsymbol{\omega}}^{(j)}_{t,m^{\Omega}} \sim \mathcal{N}^w(\boldsymbol{\omega}_{t,m^{\Omega}}, \mathbf{R}^{\Omega}_{t,m^{\Omega}}), \quad (28a)$$

$$\mathbf{m}^{(j)}_{b,t,m^{\Omega}} = [g^{-1}([\hat{r}^{(j)}_{t,m^{\Omega}}, [\hat{\boldsymbol{\omega}}^{(j)}_{t,m^{\Omega}}]^T])]^T. \quad (28b)$$

Through Eqs. (26), and (28b) the birth PHD can be modelled as a GMM formed by the superposition of GMs of individual sensor modalities,

$$\lambda_b(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_t) = \sum_{m^{\Omega}=1}^{M_t^{\Omega}} \sum_{j=1}^{J_b} w^{(j)}_{b,t,m^{\Omega}} \mathcal{N}(\mathbf{s}_t|\mathbf{m}^{(j)}_{b,t,m^{\Omega}}, \boldsymbol{\Sigma}^{\Omega}_b), \quad (29a)$$

$$\lambda_b(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Psi}_t) = \sum_{m^{\Psi}=1}^{M_t^{\Psi}} w_{b,t,m^{\Psi}} \mathcal{N}(\mathbf{s}_t|\mathbf{m}_{b,t,m^{\Psi}}, \boldsymbol{\Sigma}^{\Psi}_b), \quad (29b)$$

$$\lambda_b(\mathbf{s}_t|\mathbf{r}_t,\mathbf{\Omega}_t,\mathbf{\Psi}_t)=\sum_{\mathcal{M}}^{|\boldsymbol{\mathcal{M}}|} w_{b,t,f}\,\lambda_b^{\boldsymbol{\mathcal{M}}}(\mathbf{s}_t|\mathbf{r}_t,\boldsymbol{\mathcal{M}}_f). \tag{30}$$

Where, $w_{b,t,f}$, are the birth GMM weights across modalities, $w_{b,t,f}=\frac{1}{|\boldsymbol{\mathcal{M}}|}$, $w^{(j)}_{b,t,m^\Omega}$, are the birth GMM weights for acoustic modality, $w^{(j)}_{b,t,m^\Omega}=\frac{1}{M_t^\Omega J_b}$, and, w_{b,t,m^Ψ}, are the birth GMM weights for visual modality, $w_{b,t,m^\Psi}=\frac{1}{M_t^\Psi}$.

2) Missed and Detected PHDs - Current Predicted and Detected Targets: We propose the use of a modified GM-PHD filter [16] for use towards two measurement spaces, representative of the acoustic and visual modalities. In this method, the class-conditional independence between acoustic and visual measurement modalities is exploited. Each modality runs through an independent update of the GM-PHD Filter in order to account for observed states in each modality. Gaussian mixture weights are then used to fuse the independent updates as determined through the standard weighting scheme in Eq. (25).

For implementation of the GM-PHD filter, the source dynamics are defined in familiar formation to [1],

$$\mathbf{s}^a_{t,n}\triangleq[x^a_{t,n},y^a_{t,n},z^a_{t,n}]^T, \tag{31a}$$

$$\mathbf{s}^a_{t,n}=\mathbf{s}^a_{t-1,n}+\mathbf{n}_{t,n},\qquad \mathbf{n}_{t,n}\sim\mathcal{N}(\mathbf{0}_{3\times1},\mathbf{Q}), \tag{31b}$$

$$\mathbf{s}_{t,n}=\mathbf{\Gamma}(\gamma_t)\mathbf{s}^a_{t,n}+[x_{t,r},\,y_{t,r},\,z_{t,r}]^T, \tag{31c}$$

$$\mathbf{\Gamma}(\gamma_t)\triangleq\begin{bmatrix} cos\gamma_t & -sin\gamma_t & \mathbf{0}_{2\times1}\\ sin\gamma_t & +cos\gamma_t & \\ \mathbf{0}_{1\times2} & & 0\end{bmatrix}. \tag{31d}$$

Where, $s^a_{t,n}$, is the absolute source position represented in the AV-SLAM coordinate frame in Fig. 2, and, $\mathbf{\Gamma}(\cdot)$, is the relative-to-absolute coordinate transform.

The GM-PHD Filter is performed for each modality, after first transforming the previous target PHD to be expressed relative to the current robot position, $\mathbf{r}_t$.

$$\lambda(\mathbf{s}_{t-1}|\mathbf{r}_t,\mathbf{\Omega}_{1:t-1},\mathbf{\Psi}_{1:t-1})=\sum_{j=1}^{J_{t-1}} w^{(j)}_{t-1}\mathcal{N}(\mathbf{s}_t|\tilde{\mathbf{m}}^{(j)}_{t-1},\tilde{\mathbf{\Sigma}}^{(j)}_{t-1}), \tag{32a}$$

$$\tilde{\mathbf{m}}^{(j)}_{t-1}=\mathbf{\Gamma}(\gamma_t)\mathbf{\Gamma}^{-1}(\gamma_{t-1})\,(\mathbf{m}^{(j)}_{t-1}-\mathbf{r}_{t-1})\,+\mathbf{r}_t, \tag{32b}$$

$$\tilde{\mathbf{\Sigma}}^{(j)}_{t-1}=\mathbf{\Gamma}(\gamma_t)\mathbf{\Gamma}^{-1}(\gamma_{t-1})\mathbf{\Sigma}^{(j)}_{t-1}[\mathbf{\Gamma}^{-1}(\gamma_{t-1})]^T\mathbf{\Gamma}(\gamma_t)^T. \tag{32c}$$

The realized predicted PHD, $\lambda(\mathbf{s}_t|\mathbf{r}_t,\mathbf{\Omega}_{1:t-1},\mathbf{\Psi}_{1:t-1})$ is synonymous to the missed PHD of Eq. (23) and is provided by,

$$\mathbf{m}^{(j)}_{t|t-1}=\tilde{\mathbf{m}}^{(j)}_{t-1}, \tag{33}$$

$$\mathbf{\Sigma}^{(j)}_{t|t-1}=\tilde{\mathbf{\Sigma}}^{(j)}_{t-1}+\mathbf{Q}, \tag{34}$$

$$w^{(j)}_{t|t-1}=p_s w^{(j)}_{t-1}, \tag{35}$$

$$\lambda(\mathbf{s}_t|\mathbf{r}_t,\mathbf{\Omega}_{1:t-1},\mathbf{\Psi}_{1:t-1})=\sum_{j=1}^{J_{t-1}} w^{(j)}_{t|t-1}\mathcal{N}(\mathbf{s}_t|\mathbf{m}^{(j)}_{t|t-1},\mathbf{\Sigma}^{(j)}_{t|t-1}). \tag{36}$$

Where, p_s, is the survival probability of all targets, $w_{\Omega,\,t|t-1}$, are the predicted GM weights, and, J_{t-1}, is the number of GM components in the multi-source posterior GMM of the previous time point.

After determination of the predicted GM in Eq. (36), the J_{t-1}, components are updated by measurements from each modality to provide GMs representative of detected PHDs.

$$\mathbf{S}^{(j,k)}_{t,\,m^\Omega}=\hat{\mathbf{G}}^{(j,k)}_t\Sigma^{(j,k)}_{t|t-1}[\hat{\mathbf{G}}^{(j,k)}_t]^T+\mathbf{R}^\Omega_{t,m^\Omega}, \tag{37a}$$

$$\hat{\mathbf{G}}^{(j,k)}_t\triangleq\frac{\partial\hat{g}_k}{\partial\mathbf{s}_t}\Big|_{\mathbf{s}_t=\mathbf{m}^{(j)}_{t|t-1}},$$

$$\hat{g}_k(\mathbf{s}_t)\triangleq g(\mathbf{s}_t)-k[2\pi,\pi]^T,\qquad k=-1,0,1.$$

$$\mathbf{S}^{(j)}_{t,\,m^\Psi}=\mathbf{H}^{(j)}_t\Sigma^{(j)}_{t|t-1}[\mathbf{H}^{(j)}_t]^T+\mathbf{R}^\Psi_{t,m^\Psi}, \tag{37b}$$

$$\mathbf{H}^{(j)}_t\triangleq\frac{\partial h}{\partial\mathbf{s}_t}\Big|_{\mathbf{s}_t=\mathbf{m}^{(j)}_{t|t-1}}.$$

$$\mathbf{K}^{(j,k)}_{t,m^\Omega}=\mathbf{\Sigma}^{(j)}_{t|t-1}[\hat{\mathbf{G}}^{(j,k)}_t]^T[\mathbf{S}^{(j,k)}_{t,\,m^\Omega}]^{-1}, \tag{38a}$$

$$\mathbf{K}^{(j)}_{t,m^\Psi}=\mathbf{\Sigma}^{(j)}_{t|t-1}[\mathbf{H}^{(j)}_t]^T[\mathbf{S}^{(j)}_{t,\,m^\Psi}]^{-1}. \tag{38b}$$

$$\mathbf{m}^{(j,k)}_{t,m^\Omega}=\mathbf{m}^{(j)}_{t|t-1}+\mathbf{K}^{(j,k)}_{t,m^\Omega}(\boldsymbol{\omega}_{t,m^\Omega}-\hat{g}_k(\mathbf{m}^{(j)}_{t|t-1})), \tag{39a}$$

$$\mathbf{m}^{(j)}_{t,m^\Psi}=\mathbf{m}^{(j)}_{t|t-1}+\mathbf{K}^{(j)}_{t,m^\Psi}(\boldsymbol{\psi}_{t,m^\Psi}-h(\mathbf{m}^{(j)}_{t|t-1})). \tag{39b}$$

$$\mathbf{\Sigma}^{(j,k)}_{t,m^\Omega}=(\mathbf{I}_3-\mathbf{K}^{(j,k)}_{t,m^\Omega}\hat{\mathbf{G}}^{(j,k)}_t)\mathbf{\Sigma}^{(j)}_{t|t-1}, \tag{40a}$$

$$\mathbf{\Sigma}^{(j,k)}_{t,m^\Psi}=(\mathbf{I}_3-\mathbf{K}^{(j)}_{t,m^\Psi}\mathbf{H}^{(j)}_t)\mathbf{\Sigma}^{(j)}_{t|t-1}. \tag{40b}$$

The representative updated target GMMs, $\lambda_{\boldsymbol{\mathcal{M}}}(\mathbf{s}_t|\mathbf{r}_t)$, are synonymous with the detected PHDs in (25), and are provided by the following equation,

$$\lambda_\Omega(\mathbf{s}_t|\mathbf{r}_t)=\sum_{m^\Omega=1}^{M_{\Omega,t}}\sum_{k=-1}^{1}\sum_{j=1}^{J_{t-1}} w^{(j,k)}_{t,m^\Omega}\mathcal{N}(\mathbf{s}_t|\mathbf{m}^{(j,k)}_{t,m^\Omega},\mathbf{\Sigma}^{(j,k)}_{t,m^\Omega}), \tag{41a}$$

$$\lambda_\Psi(\mathbf{s}_t|\mathbf{r}_t)=\sum_{m^\Psi=1}^{M_{\Psi,t}}\sum_{j=1}^{J_{t-1}} w^{(j)}_{t,m^\Psi}\mathcal{N}(\mathbf{s}_t|\mathbf{m}^{(j)}_{t,m^\Psi},\mathbf{\Sigma}^{(j)}_{t,m^\Psi}). \tag{41b}$$

$$w^{(j,k)}_{t,m^\Omega}=w^{(j)}_{t|t-1}\Big(\frac{\mathcal{N}(\boldsymbol{\omega}_{t,m^\Omega}|\hat{g}_k(\mathbf{m}^{(j)}_{\Omega,t|t-1}),\mathbf{S}^{(j,k)}_{t,\,m^\Omega})}{\kappa(\boldsymbol{\omega}_{t,m^\Omega}|\mathbf{r}_t)+\sum_{j=1}^{J_{t-1}}\mathcal{N}(\boldsymbol{\omega}_{t,m^\Omega}|\hat{g}(\mathbf{m}^{(j)}_{\Omega,t|t-1}),\mathbf{S}^{(j,k)}_{t,m^\Omega})}\Big), \tag{42a}$$

$$w^{(j)}_{t,m^\Psi}=w^{(j)}_{t|t-1}\Big(\frac{\mathcal{N}(\boldsymbol{\psi}_{t,m^\Psi}|(\mathbf{m}^{(j)}_{\Psi,t|t-1},\mathbf{S}^{(j)}_{t,\,m^\Psi})}{\kappa(\boldsymbol{\psi}_{t,m^\Psi}|\mathbf{r}_t)+\sum_{j=1}^{J_{t-1}}\mathcal{N}(\boldsymbol{\psi}_{t,m^\Psi}|h(\mathbf{m}^{(j)}_{\Psi,t|t-1}),\mathbf{S}^{(j)}_{t,m^\Psi})}\Big). \tag{42b}$$

Where the denominator in Eq. (42), are the models accounting for false measurements in the acoustic and visual modalities, $g(\mathbf{s}_t)$, is the Cartesian-to-spherical coordinate transformation, and, $g(\mathbf{s}_t)$, is the Cartesian-to-cylindrical coordinate transformation.

Fig. 6: Data recording setup.

Using the detected PHDs given in Eq. (41) for the acoustic and visual modalities, we can express the source states as a singular detection PHD by the realization of the standard weighting algorithm in Eq. (25), for the updated GMMs.

$$\lambda_e(\mathbf{s}_t|\mathbf{r}_t) = \Big(\big(\sum_{m^{\Omega}=1}^{M_{\Omega,t}}\sum_{k=-1}^{1}\sum_{j=1}^{J_{t-1}} w_{t,m^{\Omega}}^{(j,k)}\mathcal{N}(\mathbf{s}_t|\mathbf{m}_{t,m^{\Omega}}^{(j,k)},\boldsymbol{\Sigma}_{t,m^{\Omega}}^{(j,k)})\big)^{\alpha} \big(\sum_{m^{\Psi}=1}^{M_{\Psi,t}}\sum_{j=1}^{J_{t-1}} w_{t,m^{\Psi}}^{(j)}\mathcal{N}(\mathbf{s}_t|\mathbf{m}_{t,m^{\Psi}}^{(j)},\boldsymbol{\Sigma}_{t,m^{\Psi}}^{(j)})\big)^{\beta}\Big). \quad (43)$$

The resulting posterior target PHD at the current time point is then expressed as the superposition of the birth GMM, (birth PHD) in Eq. (30), the predicted GMM, (missed PHD) in Eq. (36), and the updated GMM, (detected PHD) in Eq. (43). The number of components in the realized posterior target GMM of the current time point is then,

$$J_t = (M_t^{\Omega}, J_b + M_t^{\Psi}) + (J_{t-1}) + \big((3M_{t,\Omega}J_{t-1}) + (M_t^{\Psi}, J_{t-1})\big). \quad (44)$$

And the posterior target PHD is a GMM given by,

$$\lambda(\mathbf{s}_t|\mathbf{r}_t,\boldsymbol{\Omega}_{1:t},\boldsymbol{\Psi}_{1:t}) = \sum_{j=1}^{J_t} w_t^{(j)}\mathcal{N}(\mathbf{s}_t|\mathbf{m}_t^{(j)},\boldsymbol{\Sigma}_t^{(j)}). \quad (45)$$

E. AV-SLAM Observer Localization

The proposed implementation of AV-SLAM observer localization is performed through the use of the EKF [18], [19]. For evaluation, we assume the use of a oracle localizer to provide representations of absolute observer position and orientation in lieu of odometry sensors. Observer localization is performed with similar foundations to our proposed target mapping, i.e. the EKF diverges after the prediction step to obtain two updated means and covariances, $\bar{\boldsymbol{\mu}}_t^{\boldsymbol{\mathcal{M}}}$, $\bar{\boldsymbol{\Sigma}}_t^{\boldsymbol{\mathcal{M}}}$, derived from the estimated measurements from the measurement models of each modality. The resulting observer location Gaussian is given by the familiar weighted combination scheme,

$$\mathcal{N}(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t) = \mathcal{N}(\bar{\boldsymbol{\mu}}_t^{\Omega},\boldsymbol{\Sigma}_t^{\Omega})^{\alpha}\mathcal{N}(\bar{\boldsymbol{\mu}}_t^{\psi},\boldsymbol{\Sigma}_t^{\Psi})^{\beta}. \quad (46)$$

IV. Evaluation

A. Environmental Setup

Real data was obtained in a well-lit room, with 2m anechoic padded walls surrounding two-of-four sides of a 6x6, 0.5m visual-grid. Robot motion path, and target positions were planned ahead of data acquisition and placed in the visual-grid accordingly. Measurements of ground-truth robot and target positions during the experiment were derived from the visual-grid.

B. Mock-Up Robot

To conduct real-world experiments and data acquisition, we created a mock-up robot to emulate the motion and data-collection of a mobile blimp drone; intended to act as a personal partner agent for human users [11], [20]. To satisfy hardware requirements, we equipped our robot with a 2D, circular, 8-channel microphone array, i.e. the TAMAGO-03 manufactured by System in Frontier Co., Ltd., and a monocular wide-angle camera. The RTFs correspondent to TAMAGO-03 are readily obtained from the HARK database for 72 directions around the azimuth of the microphone array [21]. Robot motion was emulated by the human-assisted movement of visual and audio acquisition hardware mounted on tripod and dolly. The recording apparatus maintained a consistent height as the tripod was pushed by a human experimenter seen in Fig. 6. Control reports of the robot motion were interpolated from observed robot position, and orientations based off the visual-grid of the environmental set-up.

C. Feature Acquisition Tuning

1) Acoustic Modality: Acoustic signal is acquired at a sampling rate of 16kHz. STFT coefficients are extracted using a 512-point discrete Fourier transform (DFT) and a 400-point (25 ms) Hanning window with a shift of 10 ms. To compute the MUSIC spectrogram (3), we set $w_u = 17$ and $w_l = 91$, corresponding to 500 Hz and 2812.5 Hz, respectively, as in the HARK software [21].

2) Visual Modality: Visual signal is acquired as wide-angle RGB-video, of 640x480 net resolution and captured at a rate of 30 frames-per-second. Distorted frames of the recorded monocular video are calibrated through the use of 60 self-collected test patterns and the OpenCV Library [22].

D. AV-SLAM Parameters

For evaluation, we apply AV-SLAM formulation, but flatten dimensions to the 2-D plane with X-Y axes, to consider the lack of elevation RTFs. GMM components are run though a pruning algorithm where they are merged, and extracted after target mapping, based off equations in [23].

E. Experimental Results

As proof-of-concept to our work, we evaluate the AV-SLAM framework on the experimental configuration shown in Fig: 7. We evaluate target mapping components with an oracle localizer at $t = 0$, and $t = 50$, which correspond to 0s, and 1.5s respectively. At $t = 0$, speech is not yet present in the environment, and the newborn, and pruned GMM components are scattered in the environment as a result. However, we see a heavily weighted component at (x=2.4, y=0.48), as derived from the single RTP estimate at $t = 0$. Over the consecutive time points to $t = 50$, the stationary human sound source remains in the visual FoV, and RTP estimates are born at each point. Due to the consistent RTP estimates, and birthed DoA estimates from $t = 0$ onwards, we see a slow convergence

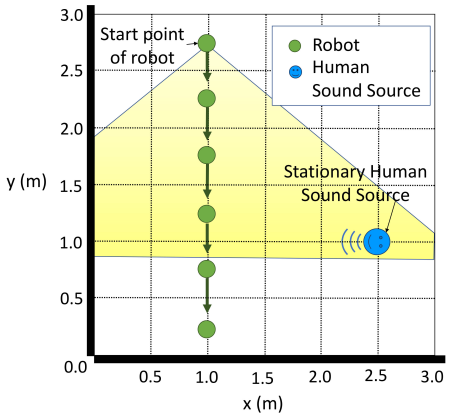
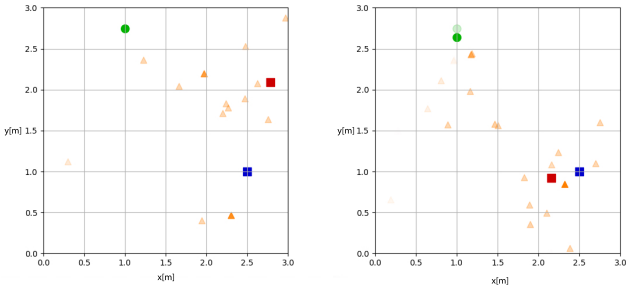


Fig. 7: Experimental configuration.



(a) $t = 0$. (b) $t = 50$.

Fig. 8: Experimental result. Yellow triangles indicate newborn and pruned GMM components, with opacity indicative of individual weighting. The green circles, red square, and blue square, are the robot position, estimated target position, and ground-truth target position, respectively.

of GMM components in the ground-truth sound source local area.

V. Conclusion

This paper presents a method for the robust acquisition for acoustic and visual localization features for use towards a multi-modality, audio-visual based SLAM algorithm. Classical methods in the study of SLAM often do not exploit signal present in the acoustic environment for localization of desired targets, and often are dependent upon features obtained through various optical sensors. Existing methods in acoustic-based SLAM rely upon temporally triangulated source-observer ranges for the convergence of multi-target source locations, and observer position. We first propose the degradation-resilient feature acquisition framework for acoustic and visual modalities by joining existing techniques in spectral mask estimation, multi-target source localization, human pose estimation, and projection plane models. We then propose to use acquired features for modality-reliability estimates, which in turn, allow the early fusion of acoustic and visual environment features in a multi-stream, GM-PHD, and EKF SLAM realization. In an environment of multiple human targets, characterized by intermittent speech in audio-FoVs, and intermittent presence inside visual-FoVs, we hope to see our proposed framework provide the foundation for robot perception, navigation, and interaction with acoustic-visual environments and human targets.

References

[1] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1484–1498, Sep. 2018.

[2] 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.

[3] J. Mullane, B. Vo, M. Adams, and B. Vo, "A random-finite-set approach to Bayesian SLAM," *IEEE Trans. Robot.*, vol. 27, no. 2, pp. 268–282, Apr. 2011.

[4] C. Evers and P. A. Naylor, "Optimized self-localization for SLAM in dynamic scenes using probability hypothesis density filters," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 863–878, Feb. 2018.

[5] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE ICASSP*, Mar. 2016, pp. 196–200.

[6] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *Proc. IEEE/RSJ IROS*, Oct. 2012, pp. 694–699.

[7] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[8] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE CVPR*, Jun. 2016, pp. 4724–4732.

[9] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE Trans. Robot. Autom.*, vol. 13, no. 2, pp. 99–110, Jun. 2006.

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

[11] N. Yao, E. Anaya, Q. Tao, S. Cho, H. Zheng, and F. Zhang, "Monocular vision-based human following on miniature robotic blimp," in *Proc. IEEE ICRA*, May 2017, pp. 3244–3249.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, Sep. 2014, pp. 740–755.

[13] J. R. Movellan and G. Chadderdon, "Channel separability in the audio-visual integration of speech: A Bayesian approach," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, D. G. Stork and M. E. Hennecke, Eds. Springer, 1996, pp. 473–487.

[14] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration: A 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech," *American Scientist*, vol. 86, no. 3, pp. 236–244, 1998.

[15] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 11, pp. 1260–1273, Nov. 2002.

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

[17] S. Gannot and A. Yeredor, "The Kalman filter," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds. Springer, 2008, pp. 135–160.

[18] T. Bailey, J. Nieto, J. Guivant, M. Stevens, and E. Nebot, "Consistency of the ekf-slam algorithm," in *Proc. IEEE/RSJ IROS*, Oct. 2006.

[19] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. The MIT Press, 2005.

[20] K. Funakoshi, H. Shimazaki, T. Kumada, and H. Tsujino, "Personal partner agents for cooperative intelligence," in *Proc. ACM/IEEE HRI*, Mar. 2019, pp. 570–571.

[21] K. Nakadai, H. G. Okuno, and T. Mizumoto, "Development, deployment and applications of robot audition open source software HARK," *J. Robot. Mechatronics*, vol. 29, no. 1, pp. 16–25, Feb. 2017.

[22] G. Bradski, "The OpenCV library," *Dr. Dobb's Journal of Software Tools*, vol. 25, no. 11, pp. 120–125, Nov. 2000.

[23] D. J. Salmond, "Mixture reduction algorithms for point and extended object tracking in clutter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 45, no. 2, pp. 667–686, Apr. 2009.