Acoustic SLAM

Christine Evers, *Senior Member, IEEE*, and Patrick A. Naylor, *Senior Member, IEEE*

***Abstract*—An algorithm is presented that enables devices equipped with microphones, such as robots, to move within their environment in order to explore, adapt to, and interact with sound sources of interest. Acoustic scene mapping creates a three-dimensional (3D) representation of the positional information of sound sources across time and space. In practice, positional source information is only provided by Direction-of-Arrival (DoA) estimates of the source directions; the source-sensor range is typically difficult to obtain. DoA estimates are also adversely affected by reverberation, noise, and interference, leading to errors in source location estimation and consequent false DoA estimates. Moreover, many acoustic sources, such as human talkers, are not continuously active, such that periods of inactivity lead to missing DoA estimates. Withal, the DoA estimates are specified relative to the observer's sensor location and orientation. Accurate positional information about the observer therefore is crucial. This paper proposes Acoustic Simultaneous Localization and Mapping (aSLAM), which uses acoustic signals to simultaneously map the 3D positions of multiple sound sources while passively localizing the observer within the scene map. The performance of aSLAM is analyzed and evaluated using a series of realistic simulations. Results are presented to show the impact of the observer motion and sound source localization accuracy.**

***Index Terms*—Bayes methods, reverberation, robot audition, simultaneous localization and mapping.**

I. Introduction

SIMULTANEOUS Localization and Mapping (SLAM) localizes an unknown, moving observer and jointly maps the 3D positions of objects of interest in the vicinity. Simultaneous Localization and Mapping (SLAM) is classically applied in applications such as robotics using visual and optical sensors [1]–[3]. In contrast to the classical context of SLAM, this paper presents an algorithm to perform SLAM using only acoustic signals. Acoustic SLAM has the potential for wide application in areas including autonomous robots, hearing aids, smart cars, and virtual reality devices.

Acoustic scene mapping typically applies source DoA estimators [4]–[6] as a pre-processing step to the acoustic measurements. In realistic acoustic environments, the resulting instantaneous DoA estimates are adversely affected by several factors. Due to the compact aperture of typical microphone arrays, for example when installed in the head of a robot [7], it is not normally feasible to determine the source-observer range [8]. Therefore, inference of Cartesian source positions from the DoA estimates is an underdetermined problem. Furthermore, reverberation [9] and noise cause estimation errors and false estimates of the source positions. Periods of inactivity, e.g., for human talkers, lead to missing source estimates.

For moving observers, spatio-temporal diversity of the sensors can be exploited for kinematic ranging of the 3D source positions from the DoA estimates [10], [11]. However, the DoA estimates are specified relative to the observer. Therefore, accurate knowledge of the observer's positional information is crucial in order to update the absolute source position estimates with the DoA estimates relative to the instantaneous observer location. In many cases, the observer has access to information on its own location obtained from, e.g., motor control data or inertial sensors. Nevertheless, in practice, such information is reported subject to errors due to physical and mechanical limitations [12]. Hence, dead reckoning [13], i.e., the propagation of the initial observer position using the reported motion information, leads to position estimates that diverge from the ground truth over time.

The observer location can also be anchored by identifying the position that best aligns the DoA estimates with the mapped sources. Therefore, observer localization and source mapping represent the joint estimation problem of SLAM [14], [15]. SLAM has received extensive attention in the robotics community [16], [17], predominantly for machine vision. Very few contributions address the application of SLAM to audio signals, and can be broadly classified into two categories. The first category [18]–[19] applies visual SLAM techniques, e.g., FastSLAM [20], to acoustic Times-of-Arrivals (TOAs). By virtue of the universal presence of immovable fixtures in visual scenes, FActored Solution To Simultaneous Localization and Mapping (FastSLAM) aligns the observer using permanently visible landmarks. However, this prerequisite is fundamentally conflicting with acoustic signals, affected by speech inactivity and reverberation. The second category [21], [22] localizes an active acoustic observer, equipped with a loudspeaker for actively probing the room and microphones for Room Impulse Response (RIR) measurements, and estimates the room dimensions from the TOAs of early reflections. However, TOA estimation from RIRs is ambiguous [23], and emission of controlled sound stimuli for RIR measurements is highly intrusive to people nearby, and therefore unacceptable in many important use-cases. Moreover, to avoid interference with the RIR measurements, the environment cannot contain sound sources other than the measurement stimuli.

Manuscript received October 10, 2017; revised March 1, 2018; accepted April 3, 2018. Date of publication April 18, 2018; date of current version May 21, 2018. This work was supported by the U.K. EPSRC Fellowship under Grant EP/P001017/1 and in part by the European Union's Seventh Framework Programme (FP7/2007–2013) under Grant 609465. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Roland Badeau. *(Corresponding author: Christine Evers.)*

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: c.evers@imperial.ac.uk; p.naylor@imperial.ac.uk).

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

Digital Object Identifier 10.1109/TASLP.2018.2828321

In this paper, we propose a novel approach, named Acoustic SLAM (aSLAM), to map the positions of sound sources passively, and simultaneously localize a moving observer in realistic acoustic environments. To avoid the need of permanent sound sources to act as landmarks, aSLAM is based on the theoretical foundations of Random Finite Sets (RFSs) [24] for mapping multiple, intermittent sources subject to erroneous, false and missing DoA estimates. Moreover, to avoid the active emission of intrusive sound stimuli, aSLAM passively infers the 3D Cartesian source positions from the 2D DoA estimates, by exploiting constructively the spatio-temporal diversity of the observer for probabilistic source triangulation. The observer is localized by probabilistic anchoring that best aligns the DoA estimates with the mapped source positions. Therefore, the novel features of the proposed aSLAM approach are 1) the joint estimation of the unknown observer path and the positions of multiple interfering sound sources, that is 2) robust against reverberation, noise, and periods of source inactivity, and uses 3) passive acoustic sensor arrays. Performance results for controlled and realistic room acoustics are presented to analyze the theoretical and practical behaviour of aSLAM. Specifically, it is shown that aSLAM is robust to the adverse effects of reverberation on DoA estimation, as well as uncertainties in the motion reports of the observer.

The paper is structured as follows: Section II formulates the problem and Section III summarizes the required background on SLAM using RFSs. Sections IV and V derive the proposed aSLAM algorithm. Section VI and VII present the experimental setup and results. Conclusions are drawn in Section VIII.

II. Problem Formulation

A. Observer Dynamics

Consider an observer with positional state, $\mathbf{r}_t$, at time step t. As a property specific to moving microphone arrays, the observer state is characterized by its Cartesian position, $(x_{t,r}, y_{t,r}, z_{t,r})$, orientation, γ_t, and speed, v_t. For readability, it is assumed in this paper that the observer moves in the direction of its orientation. As the observer position is a non-linear function of the orientation, the observer state is modelled as $\mathbf{r}_t \triangleq [\mathbf{p}_t^T, \gamma_t]^T$, where $\mathbf{p}_t \triangleq [x_{t,r}, y_{t,r}, z_{t,r}, v_t]^T$ and the orientation are given by

$$\mathbf{p}_t = \mathbf{F}_t\, \mathbf{p}_{t-1} + \mathbf{v}_{t,\mathbf{p}}, \quad \mathbf{v}_{t,\mathbf{p}} \sim \mathcal{N}(\mathbf{0}_{4\times 1}, \mathbf{\Sigma}_{t,\mathbf{v}}) \tag{1a}$$

$$\gamma_t = \vartheta(\gamma_{t-1} + v_{t,\gamma}), \quad v_{t,\gamma} \sim \mathcal{N}\left(0, \sigma^2_{v,\gamma_t}\right) \tag{1b}$$

where $\mathbf{0}_{I\times J}$ denotes the $I \times J$ zero matrix and $\vartheta(\alpha) = \mathrm{mod}\,(\alpha, 2\pi)$ is the wrapping operator that ensures that $\gamma_t \in [0, 2\pi)$, and $\mathbf{v}_{t,\mathbf{p}}$ and $v_{t,\gamma}$ are the process noise terms with covariance $\mathbf{\Sigma}_{t,\mathbf{v}}$ and σ^2_{v,γ_t} respectively. The matrix $\mathbf{F}_t$ is the dynamical model [25], given by

$$\mathbf{F}_t = \begin{bmatrix} \mathbf{I}_3 & \Delta_t \left[\sin\gamma_t, \cos\gamma_t, 0\right]^T \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix}, \tag{2}$$

where Δ_t is the delay between $t-1$ and t, and $\mathbf{I}_L$ denotes the $L \times L$ identity matrix.

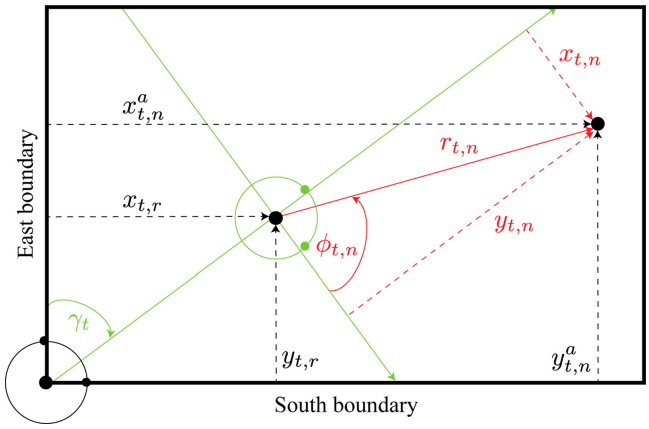


Fig. 1. Absolute (black) and observer-relative (green) source coordinates.

The motion reports of the observer speed and orientation are defined as $\mathbf{y}_t \triangleq [y_{t,v}, y_{t,\gamma}]^T$, and are modelled by

$$y_{t,v} = \mathbf{h}\,\mathbf{p}_t + w_{t,v}, \quad w_{t,v} \sim \mathcal{N}\left(0, \sigma^2_{w,v_t}\right) \tag{3a}$$

$$y_{t,\gamma} = \vartheta(\gamma_t + w_{t,\gamma}), \quad w_{t,\gamma} \sim \mathcal{N}\left(0, \sigma^2_{w,\gamma_t}\right) \tag{3b}$$

where $\mathbf{h} \triangleq [0, 0, 0, 1]$ and $w_{t,v}$ and $w_{t,\gamma}$ are noise terms with variance σ^2_{w,v_t} and σ^2_{w,γ_t} respectively.

B. Source Dynamics

The absolute positional state, $\mathbf{s}^a_{t,n} \triangleq [x^a_{t,n}, y^a_{t,n}, z^a_{t,n}]^T$, of source $n = 1, \ldots, N_t$ at time step t and position $(x^a_{t,n}, y^a_{t,n}, z^a_{t,n})$ in the global reference frame is described by

$$\mathbf{s}^a_{t,n} = \mathbf{s}^a_{t-1,n} + \mathbf{n}_{t,n}, \qquad \mathbf{n}_{t,n} \sim \mathcal{N}(\mathbf{0}_{3\times 3}, \mathbf{Q}), \tag{4}$$

where $\mathbf{n}_{t,n}$ is the process noise with covariance $\mathbf{Q}$. The source state relative to the observer is given by $\mathbf{s}_{t,n}$, defined as:

$$\mathbf{s}_{t,n} = \mathbf{\Gamma}(\gamma_t)\, \mathbf{s}^a_{t,n} + \left[x_{t,r}, y_{t,r}, z_{t,r}\right]^T, \tag{5a}$$

$$\mathbf{\Gamma}(\gamma_t) \triangleq \begin{bmatrix} +\cos\gamma_t & -\sin\gamma_t & \mathbf{0}_{2\times 1} \\ +\sin\gamma_t & +\cos\gamma_t & \\ \mathbf{0}_{1\times 2} & & 0 \end{bmatrix}. \tag{5b}$$

The reference frames used in this paper are illustrated in Fig. 1.

As acoustic sources, such as human talkers, are not continuously active, the number of sources, N_t, is time-varying and unknown. The number of sources and their corresponding states are hence modelled by a RFS [25] with realizations, $\mathbf{S}_t$:

$$\mathbf{S}_t = \left[\bigcup_{n=1}^{N_{t-1}} P(\mathbf{s}_{t-1,n})\right] \cup B_t, \tag{6}$$

where B_t models any newborn sources, active for the first time at t, and $P(\mathbf{s}_{t-1,n})$ describes the process of sources that "survive" between $t-1$ and t as

$$P(\mathbf{s}_{t-1,n}) = \begin{cases} \{\mathbf{s}_{t,n}\} & \text{if source } n \text{ was previously active} \\ \emptyset & \text{otherwise,} \end{cases} \tag{7}$$

where $\emptyset$ denotes the empty set.

DoA estimation algorithms are used to infer the instantaneous directions of sound waves impinging on the observer's microphone array at time step t. However, reverberation and noise lead to estimation errors as well as missing and false estimates. The DoA estimates are hence modelled by a RFS with realizations, $\mathbf{\Omega}_t$, such that

$$\mathbf{\Omega}_t = \left[\bigcup_{n=1}^{N_t} D(\mathbf{s}_{t,n}) \right] \cup K_t, \tag{8}$$

where K_t denotes the Poisson point process of $N_{t,c}$ Independent and Identically Distributed (IID) false DoA estimates distributed uniformly over the unit sphere [26]. The process, $D(\mathbf{s}_{t,n})$, models the missing DoAs and estimation errors, i.e.,

$$D(\mathbf{s}_{t,n}) = \begin{cases} \{\boldsymbol{\omega}_{t,m}\} & \text{if source } n \text{ is detected} \\ \emptyset & \text{if source } n \text{ is undetected,} \end{cases} \tag{9}$$

where $\boldsymbol{\omega}_{t,m} = [\, \phi_{t,m}, \theta_{t,m} \,]^T$ for $m = 1, \ldots, M_t$. The azimuth, $\phi_{t,m} \in [0, 2\pi)$ rad rotates counter-clockwise, where $\phi_{t,m} = 0$ rad points along the positive x-axis (see Fig. 1). The inclination, $\theta_{t,m} \in [0, \pi]$ rad rotates from $\theta_{t,m} = 0$ rad defined along the positive z-axis, to $\theta_{t,m} = \pi$ rad along the negative z-axis.

Each of the M_t DoA estimates is modelled as

$$\boldsymbol{\omega}_{t,m} = \hat{\vartheta}\left(g(\mathbf{s}_{t,n}) + \mathbf{e}_{t,m}\right), \quad \mathbf{e}_{t,m} \sim \mathcal{N}\left(\mathbf{0}_{2\times 1},\, \mathbf{R}_{t,m}\right) \tag{10}$$

where $g(\cdot)$ is the Cartesian-to-spherical transformation, $\mathbf{e}_{t,m}$ is the DoA estimation error with covariance $\mathbf{R}_{t,m}$, and the wrapping operator, $\hat{\vartheta}(\hat{\boldsymbol{\omega}})$ for any $\hat{\boldsymbol{\omega}} \triangleq [\, \hat{\phi}, \hat{\theta} \,]^T$, is defined as

$$\hat{\vartheta}(\hat{\boldsymbol{\omega}}) \triangleq \begin{cases} \left[\, \mathrm{mod}(\hat{\phi}, 2\pi), \hat{\theta} \,\right]^T, & \text{if } \hat{\theta} \in [0, \pi], \\ \left[\, \mathrm{mod}(\hat{\phi} + \pi, 2\pi), \pi - (\hat{\theta} - \pi) \,\right]^T, & \text{if } \hat{\theta} > \pi, \\ \left[\, \mathrm{mod}(\hat{\phi} + \pi, 2\pi), |\hat{\theta}| \,\right]^T, & \text{if } \hat{\theta} < 0. \end{cases}$$

To derive an estimator of $\mathbf{r}_t$ and $\mathbf{S}_t$ given $\mathbf{\Omega}_t$ and $\mathbf{y}_t$, the following challenges need to be addressed.

1) The number of sources and their states relative to the observer as in (5) and (6) are unknown and time-varying.
2) The model in (4) and (10), mapping from the 2D DoAs to the 3D states, presents an underdetermined system.
3) The non-linear wrapping in (10) results in a non-Gaussian Probability Density Function (pdf) of the DoA estimates.
4) The observer model in (1) and (3) is non-linear, non-Gaussian.

The background theory necessary to address Challenge 1) is summarized in Section III. Challenge 2) and 3) are addressed in Section IV. Section V addresses Challenge 4). A block diagram overview of the proposed processing chain for aSLAM is provided in Fig. 2.

III. Background on SLAM Using RFSs

This section summarizes the background theory on SLAM using RFSs required for the derivation of aSLAM.

A. Posterior pdf for SLAM

The SLAM problem is fully described by the posterior pdf, $p(\mathbf{r}_t, \mathbf{S}_t \,|\, \mathbf{y}_{1:t}, \mathbf{\Omega}_{1:t})$, which can be factorized into the observer

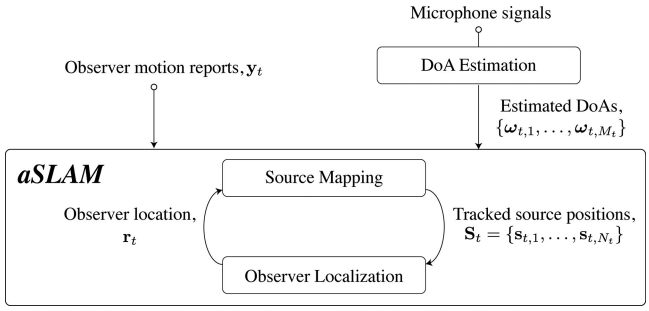


Fig. 2. Block diagram of the proposed processing chain.

posterior pdf, $p(\mathbf{r}_t \,|\, \mathbf{y}_{1:t}, \mathbf{\Omega}_{1:t})$, and conditional multi-source posterior pdf, $p(\mathbf{S}_t \,|\, \mathbf{r}_t, \mathbf{\Omega}_{1:t})$:

$$p\left(\mathbf{r}_t, \mathbf{S}_t \mid \mathbf{y}_{1:t}, \mathbf{\Omega}_{1:t}\right) = p\left(\mathbf{r}_t \mid \mathbf{y}_{1:t}, \mathbf{\Omega}_{1:t}\right) p\left(\mathbf{S}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t}\right). \tag{11}$$

The multi-source posterior pdf conditional on the observer state, $\mathbf{r}_t$, is propagated sequentially via Bayes's theorem:

$$p\left(\mathbf{S}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t}\right) = \frac{p\left(\mathbf{\Omega}_t \mid \mathbf{r}_t, \mathbf{S}_t\right) p\left(\mathbf{S}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1}\right)}{\int p\left(\mathbf{\Omega}_t \mid \mathbf{r}_t, \mathbf{S}_t\right)\, p\left(\mathbf{S}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1}\right) \delta\mathbf{S}_t}, \tag{12}$$

where $p(\mathbf{\Omega}_t \,|\, \mathbf{r}_t, \mathbf{S}_t)$ is the likelihood of the set of DoA estimates. The predicted pdf, $p(\mathbf{S}_t \,|\, \mathbf{r}_t, \mathbf{\Omega}_{1:t-1})$, is given by

$$p\left(\mathbf{S}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1}\right) = \int p\left(\mathbf{S}_t \mid \mathbf{r}_t, \mathbf{S}_{t-1}\right)\, p\left(\mathbf{S}_{t-1} \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1}\right) \delta\mathbf{S}_{t-1}, \tag{13}$$

where $p(\mathbf{S}_t \,|\, \mathbf{r}_t, \mathbf{S}_{t-1})$ models the transition of the set of sources between $t-1$ and t, and $p(\mathbf{S}_{t-1} \,|\, \mathbf{r}_t, \mathbf{\Omega}_{1:t-1})$ is the posterior pdf at $t-1$ and relative to $\mathbf{r}_t$. The set integral, $\int \cdot\, \delta\mathbf{W}$, in (12) and (13) is defined for any $\mathbf{W}$ as [25]

$$\int p(\mathbf{W})\, \delta\mathbf{W} = p(\emptyset) + \sum_{n=1}^{\infty} \frac{1}{n!} \int \ldots \int p(\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}) d\mathbf{w}_1 \ldots d\mathbf{w}_n. \tag{14}$$

The summation in (14) enumerates the hypotheses that any number of objects can be contained in $\mathbf{W}$. The set integral is hence equivalent to marginalizing over all subsets $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\} \subset \mathbf{W}$, $\forall n = 0, \ldots, \infty$ [27]. However, as a consequence, (12) is combinatorially intractable.

B. Probability Hypothesis Density

The multi-source pdf, $p(\mathbf{S}_t \,|\, \mathbf{r}_t, \mathbf{\Omega}_{1:t})$, can be approximated by its first-order moment, referred to as the Probability Hypothesis Density (PHD), $\lambda(\mathbf{s}_t \,|\, \mathbf{r}_t, \mathbf{\Omega}_{1:t})$. The PHD expresses the probability that one of the multiple objects in $\mathbf{S}_t$ has the state $\mathbf{s}_t$. Assuming the random finite set, $\mathbf{S}_t$, is a Poisson point process [27], i.e., the number of sources, N_t, is Poisson distributed and the source states are IID, the posterior pdf and its corresponding

PHD are related via [25], [28]

$$p\left(\mathbf{S}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t}\right) = e^{-N_t} \prod_{\mathbf{s}_t \in \mathbf{S}_t} \lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t}\right), \tag{15}$$

$$\lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t}\right) = \int \delta_{\mathbf{S}_t}(\mathbf{s}_t)\, p\left(\mathbf{S}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t}\right) \delta\mathbf{S}_t, \tag{16}$$

where $\delta_{\mathbf{S}_t}(\mathbf{s}_t) = \sum_{\mathbf{s}' \in \mathbf{S}_t} \delta_{\mathbf{s}'}(\mathbf{s}_t)$ is the sum of Dirac-Delta functions concentrated at $\mathbf{s}' \in \mathbf{S}_t$. It is important to note that the PHD is by definition not a pdf, but rather an intensity, since it integrates to [25]:

$$\int \lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t}\right) d\mathbf{s}_t = \mathbb{E}\left[N_t\right], \tag{17}$$

where $\mathbb{E}[\cdot]$ denotes the expected value. Nevertheless, (17) is an important property of the PHD: By estimating the posterior PHD, $\lambda(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t})$, an estimate of the number of sources is simultaneously obtained.

The general, sensor-agnostic posterior PHD, independent of models such as in Section II, accounts for the hypotheses that a source may 1) be "born" at time step t with a probability of p_b, 2) persist between $t-1$ and t and be detected with probability p_d by a DoA estimate, or 3) be missed by DoA estimation with probability $(1-p_d)$, i.e., [25]

$$\begin{aligned}\lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t}\right) &= p_b\, \lambda_b\left(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_t\right) \\ &\quad + p_d\, \lambda_d\left(\mathbf{s}_t \mid \mathbf{r}_t\right) + (1-p_d)\, \lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1}\right),\end{aligned} \tag{18}$$

where $\lambda_b(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_t)$ is the birth PHD. The predicted PHD, $\lambda(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1})$, propagates the previous PHD, $\lambda(\mathbf{s}_{t-1} \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1})$, from $t-1$ to t by applying the transition density, $p(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{s}_{t-1})$, of the source dynamical model, i.e.,

$$\begin{aligned}&\lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1}\right) \\ &\quad = p_s \int p\left(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{s}_{t-1}\right) \lambda\left(\mathbf{s}_{t-1} \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1}\right) d\mathbf{s}_{t-1},\end{aligned} \tag{19}$$

where p_s is the survival probability and $\lambda(\mathbf{s}_{t-1} \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1})$ is obtained by transformation from $\mathbf{r}_{t-1}$ to $\mathbf{r}_t$ using (5).

The detection PHD, $\lambda_d(\mathbf{s}_t \mid \mathbf{r}_t)$, in (18) updates the predicted PHD with new information inferred from each DoA estimate, $\boldsymbol{\omega}_{t,m}$, for all $m = 1, \ldots, M_t$, i.e.,

$$\lambda_d\left(\mathbf{s}_t \mid \mathbf{r}_t\right) \triangleq \sum_{m=1}^{M_t} \frac{p\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t, \mathbf{s}_t\right) \lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1}\right)}{\ell\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t\right)}, \tag{20}$$

where $p(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t, \mathbf{s}_t)$, is the likelihood function of $\boldsymbol{\omega}_{t,m}$. The denominator in (20) is a crucial term specific to the RFS formulation of Bayes's paradigm. The denominator incorporates an explicit model of false DoA estimates, i.e.,

$$\ell\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t\right) \triangleq \kappa\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t\right) + p_d\, p\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t\right), \tag{21}$$

where the evidence, $p(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t)$, and likelihood of a false DoA estimate, $\kappa(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t)$, are given by

$$p\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t\right) = \int \lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t-1}\right) p\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t, \mathbf{s}_t\right) d\mathbf{s}_t, \tag{22a}$$

$$\kappa(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t) = \lambda_c\, \mathcal{U}\left[\mathbf{0}_{2\times 1},\ \left[2\pi, \pi\right]^T\right], \tag{22b}$$

and $\mathcal{U}[\cdot]$ denotes the uniform distribution and λ_c is the expected rate of false DoA estimates. For high rates of false DoA estimates, e.g., in reverberant environments, (20) is scaled by large values of $\kappa(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t)$ and hence $\ell(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t)$. Depletion of the detection PHD, $\lambda_d(\mathbf{s}_t \mid \mathbf{r}_t)$, by large values of $\ell(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t)$ can only be avoided if the predicted PHD corresponds to large likelihood values in the numerator of (20). Section IV elaborates on the conditions required for acoustic sources to outweigh the contribution of false DoA estimates.

C. Posterior SLAM PHD

The general, model-independent SLAM PHD, $\lambda(\mathbf{r}_t, \mathbf{S}_t \mid \mathbf{y}_{1:t}, \mathbf{\Omega}_{1:t})$, corresponding to the pdf in (11), was derived in [27] as:

$$\begin{aligned}&\lambda\left(\mathbf{r}_t, \mathbf{S}_t \mid \mathbf{y}_{1:t}, \mathbf{\Omega}_{1:t}\right) \\ &\quad = \frac{\mathcal{L}\left(\mathbf{\Omega}_t \mid \mathbf{r}_t\right) p\left(\mathbf{r}_t \mid \mathbf{y}_{1:t}, \mathbf{\Omega}_{1:t}\right)}{\int \mathcal{L}\left(\mathbf{\Omega}_t \mid \mathbf{r}_t\right) p\left(\mathbf{r}_t \mid \mathbf{y}_{1:t}, \mathbf{\Omega}_{1:t}\right) d\mathbf{r}_t} \lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t}\right),\end{aligned} \tag{23}$$

where $\mathcal{L}(\mathbf{\Omega}_t \mid \mathbf{r}_t)$ is the evidence of the set of DoA estimates:

$$\mathcal{L}\left(\mathbf{\Omega}_t \mid \mathbf{r}_t\right) \triangleq e^{-N_{t,c} - p_d\, N_{t|t-1}} \prod_{m=1}^{M_t} \ell\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t\right), \tag{24}$$

and the predicted number of sources, $N_{t|t-1}$, is obtained through (17).

The SLAM PHD, $\lambda(\mathbf{r}_t, \mathbf{S}_t \mid \mathbf{y}_{1:t}, \mathbf{\Omega}_{1:t})$, expresses the probability that one of the many sources in the acoustic scene map takes the state, $\mathbf{s}_t$, and the observer takes the state, $\mathbf{r}_t$. The expression in (23) facilitates for full exploitation of the joint dependency between the sources and the observer. The sources are dependent on the observer through $\lambda(\mathbf{s}_t \mid \mathbf{r}_t, \mathbf{\Omega}_{1:t})$. Simultaneously, the observer pdf is dependent on the sources through $\mathcal{L}(\mathbf{\Omega}_t \mid \mathbf{r}_t)$.

It is important to note that the evidence of the set of DoA estimates in (24) corresponds to the intersection of the evidence terms over all DoA estimates. This intersection facilitates that the observer position is probabilistically anchored by identifying the positional state that best aligns the DoA estimates with the map of surviving sources (see Section V.).

In the following, Section IV and Section V present the novel realization of (23), specific to the source and observer models in Section II for acoustic SLAM.

IV. aSLAM Source Mapping

This section proposes a method to probabilistically triangulate the source positions from the DoA estimates by exploiting the observer's spatio-temporal diversity.

A. Range Induction at Source Initialization

A birth process driven by the DoA estimates is proposed in this subsection that induces source-observer range hypotheses to address the challenge of the underdetermined system.

In order to estimate the unmeasured source range, source states are initialized by drawing J_b range hypotheses, $\hat{r}_{t,m}^{(j)}$, along

each DoA, $\boldsymbol{\omega}_{t,m}$, for all $m = 1, \ldots, M_t$, i.e.,

$$\hat{r}_{t,m}^{(j)} \sim \mathcal{U}\left(r_{\min}, r_{\max}\right), \tag{25}$$

where $j = 1, \ldots, J_b$ and $\mathcal{U}(r_{\min}, r_{\max})$ denotes the uniform distribution between the minimum and maximum source-observer distance, $r_{\min}$ and $r_{\max}$ respectively. To account for DoA estimation errors, J_b hypotheses, $\hat{\boldsymbol{\omega}}_{t,m}^{(j)}$, of the source direction are drawn from a wrapped Gaussian distribution [29]:

$$\hat{\boldsymbol{\omega}}_{t,m}^{(j)} \sim \mathcal{N}^w\left(\boldsymbol{\omega}_{t,m}, \mathbf{R}_{t,m}\right), \tag{26}$$

where $\mathbf{R}_{t,m}$ is assumed known *a priori* in this paper. The initialized source states, $\mathbf{m}_{b,t,m}^{(j)}$, are constructed as

$$\mathbf{m}_{b,t,m}^{(j)} = \left[g^{-1}\left(\left[\hat{r}_{t,m}^{(j)}, \left[\hat{\boldsymbol{\omega}}_{t,m}^{(j)}\right]^T \right] \right) \right]^T, \tag{27}$$

where $g^{-1}(\cdot)$ is the spherical-to-Cartesian transformation.

From (4), the prior source pdf is Gaussian. To ensure that the posterior source PHD corresponds to a stable distribution, the birth PHD is modelled by a Gaussian Mixture Model (GMM):

$$\lambda_b\left(\mathbf{s}_t \mid \mathbf{r}_t, \boldsymbol{\Omega}_t\right) = \sum_{m=1}^{M_t} \sum_{j=1}^{J_b} w_{b,t,m}^{(j)} \mathcal{N}\Big(\mathbf{s}_t \,\Big|\, \mathbf{m}_{b,t,m}^{(j)}, \boldsymbol{\Sigma}_b\Big), \tag{28}$$

where $w_{b,t,m}^{(j)} = (J_b\, M_t)^{-1}$ are the Gaussian Mixture (GM) weights and $\boldsymbol{\Sigma}_b$ is constant and assumed known *a priori*.

At each time-step, (28) results in $M_t\, J_b$ source state hypotheses, randomly sampled along vectors in the direction of the DoA estimates. The fundamental principle is that a sufficient number of random hypotheses will result in candidates close to the true source position.

B. Probabilistic Source Triangulation

This subsection proposes an approach for probabilistic triangulation that distinguishes meaningful hypotheses by exploiting spatio-temporal diversity of the observer.

As the birth PHD is a GMM, it is assumed that the source PHD at $t-1$, $\lambda(\mathbf{s}_{t-1} \mid \mathbf{r}_{t-1}, \boldsymbol{\Omega}_{1:t-1})$, is a GM of J_{t-1} components with mean $\mathbf{m}_{t-1}^{(j)}$, covariance, $\boldsymbol{\Sigma}_{t-1}^{(j)}$, and weight, $w_{t-1}^{(j)}$, where the GM mean and covariance are relative to $\mathbf{r}_{t-1}$. As the observer state changes at t, the components relative to $\mathbf{r}_t$ are obtained using (5), such that

$$\begin{aligned}
&\lambda\left(\mathbf{s}_{t-1} \mid \mathbf{r}_t, \boldsymbol{\Omega}_{1:t-1}\right) = \sum_{j=1}^{J_{t-1}} w_{t-1}^{(j)} \mathcal{N}\left(\mathbf{s}_t \,\middle|\, \tilde{\mathbf{m}}_{t-1}^{(j)}, \tilde{\boldsymbol{\Sigma}}_{t-1}^{(j)}\right), \\
&\tilde{\mathbf{m}}_{t-1}^{(j)} = \boldsymbol{\Gamma}(\gamma_t)\boldsymbol{\Gamma}^{-1}(\gamma_{t-1})(\mathbf{m}_{t-1}^{(j)} - \mathbf{r}_{t-1}) + \mathbf{r}_t, \\
&\tilde{\boldsymbol{\Sigma}}_{t-1}^{(j)} = \boldsymbol{\Gamma}(\gamma_t)\boldsymbol{\Gamma}^{-1}(\gamma_{t-1})\boldsymbol{\Sigma}_{t-1}^{(j)} \left[\boldsymbol{\Gamma}^{-1}(\gamma_{t-1})\right]^T \boldsymbol{\Gamma}(\gamma_t)^T.
\end{aligned} \tag{29}$$

The state of each component in (29) can be predicted at t using the source dynamical model in (4). The predictions are subsequently updated by inferring information from the DoA estimates in (10). As derived in Appendix A, the predicted and detection PHDs in (19) and (20) thus are:

$$\begin{aligned}
&\lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \boldsymbol{\Omega}_{1:t-1}\right) = \sum_{j=1}^{J_{t-1}} w_{t|t-1}^{(j)} \mathcal{N}\left(\mathbf{s}_t \,\middle|\, \mathbf{m}_{t|t-1}^{(j)}, \boldsymbol{\Sigma}_{t|t-1}^{(j)}\right), \\
&\lambda_d\left(\mathbf{s}_t \mid \mathbf{r}_t\right) \approx \sum_{m=1}^{M_t} \sum_{k=-1}^{1} \sum_{j=1}^{J_{t-1}} w_{t,m}^{(j,k)} \mathcal{N}\left(\mathbf{s}_t \,\middle|\, \mathbf{m}_{t,m}^{(j,k)}, \boldsymbol{\Sigma}_{t,m}^{(j,k)}\right),
\end{aligned} \tag{30}$$

where $w_{t|t-1}^{(j)}$ and $w_{t,m}^{(j,k)}$ are the predicted and updated GM weights, and the mean and covariance terms are given by the Extended Kalman Filter (EKF) equations [30]:

$$\mathbf{m}_{t|t-1}^{(j)} = \tilde{\mathbf{m}}_{t-1}^{(j)} \quad \text{and} \quad \boldsymbol{\Sigma}_{t|t-1}^{(j)} = \tilde{\boldsymbol{\Sigma}}_{t-1}^{(j)} + \mathbf{Q}, \tag{31a}$$

$$\mathbf{m}_{t,m}^{(j,k)} = \mathbf{m}_{t|t-1}^{(j)} + \mathbf{K}_{t,m}^{(j,k)} \left(\boldsymbol{\omega}_{t,m} - \hat{g}_k(\mathbf{m}_{t|t-1}^{(j)})\right), \tag{31b}$$

$$\boldsymbol{\Sigma}_{t,m}^{(j,k)} = \left(\mathbf{I}_3 - \mathbf{K}_{t,m}^{(j,k)}\, \hat{\mathbf{G}}_t^{(j,k)}\right) \boldsymbol{\Sigma}_{t|t-1}^{(j)}, \tag{31c}$$

where $g(\mathbf{s}_t)$ is the Cartesian-to-spherical transformation and $\hat{g}_k(\mathbf{s}_t) \triangleq g(\mathbf{s}_t) - k\big[2\pi, \pi\big]^T$ for $k = -1, 0, 1$ accounts for the directional wrapping, and $\hat{\mathbf{G}}_t^{(j,k)} \triangleq \partial \hat{g}_k / \partial \mathbf{s}_t \big|_{\mathbf{s}_t = \mathbf{m}_{t|t-1}^{(j)}}$. The gain, $\mathbf{K}_{t,m}^{(j,k)}$, and innovation covariance, $\mathbf{S}_{t,m}^{(j,k)}$, are:

$$\mathbf{K}_{t,m}^{(j,k)} = \boldsymbol{\Sigma}_{t|t-1}^{(j)} \left[\hat{\mathbf{G}}_t^{(j,k)}\right]^T \left[\mathbf{S}_{t,m}^{(j,k)}\right]^{-1}, \tag{32a}$$

$$\mathbf{S}_{t,m}^{(j,k)} = \hat{\mathbf{G}}_t^{(j,k)} \boldsymbol{\Sigma}_{t|t-1}^{(j)} \left[\hat{\mathbf{G}}_t^{(j,k)}\right]^T + \mathbf{R}_{t,m}. \tag{32b}$$

The predicted and updated weights, $w_{t|t-1}^{(j)}$ and $w_{t,m}^{(j,k)}$, are

$$w_{t|t-1}^{(j)} = p_s\, w_{t-1}^{(j)}, \tag{33a}$$

$$w_{t,m}^{(j,k)} = w_{t|t-1}^{(j)} \frac{\mathcal{N}\left(\boldsymbol{\omega}_{t,m} \,\middle|\, \hat{g}_k(\mathbf{m}_{t|t-1}^{(j)}), \mathbf{S}_{t,m}^{(j,k)}\right)}{\ell\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t\right)}, \tag{33b}$$

where $\ell(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t)$ is obtained using (21) and (22a) as:

$$\begin{aligned}
&\ell\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t\right) \\
&\quad = \kappa\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t\right) + \sum_{j=1}^{J_{t-1}} \mathcal{N}\left(\boldsymbol{\omega}_{t,m} \,\middle|\, \hat{g}(\mathbf{m}_{t|t-1}^{(j)}), \mathbf{S}_{t,m}^{(j,k)}\right).
\end{aligned} \tag{34}$$

Using (28) and (30), the posterior source PHD in (18) hence reduces to a GMM of $J_t = J_b + J_{t-1} + 3\, J_{t-1}\, M_t$ components, i.e.,

$$\lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \boldsymbol{\Omega}_{1:t}\right) = \sum_{j=1}^{J_t} w_t^{(j)} \mathcal{N}\left(\mathbf{s}_t \,\middle|\, \mathbf{m}_t^{(j)}, \boldsymbol{\Sigma}_t^{(j)}\right). \tag{35}$$

The expected number of sources, $\hat{N}_t$, is obtained via (17) as:

$$\begin{aligned}
\hat{N}_t = &\sum_{j=1}^{J_{t-1}} \left((1 - p_d)\, w_{t-1}^{(j)}\right) + \sum_{m=1}^{M_t} \Bigg[\sum_{j=1}^{J_{b,t}} \left(p_b\, w_{b,t,m}^{(j)}\right) \\
&+ \sum_{j=1}^{J_{t-1}} \sum_{k=-1}^{1} \left(p_d\, w_{t,m}^{(j,k)}\right) \Bigg].
\end{aligned}$$

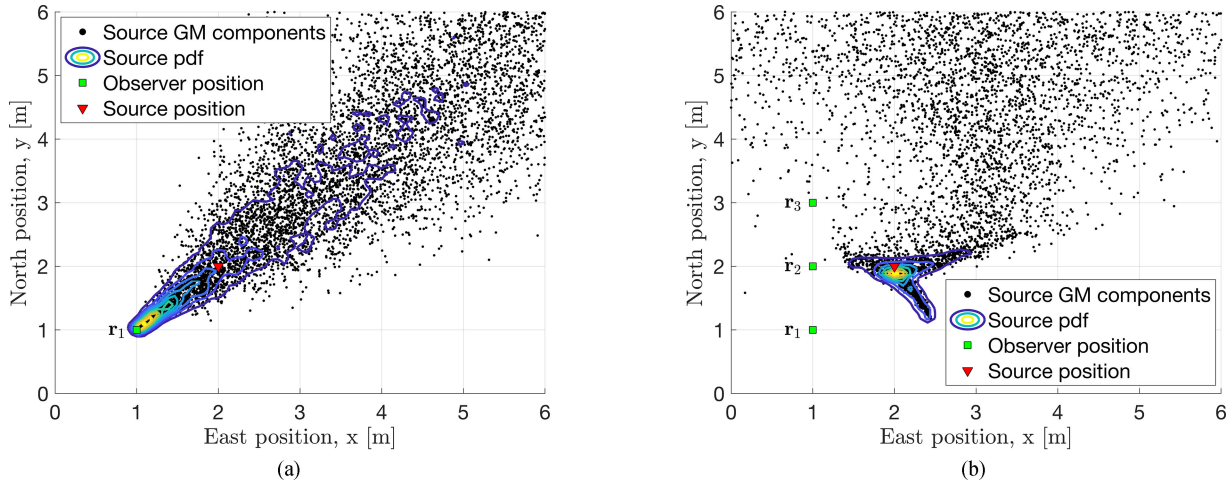


Fig. 3. Illustrative example of probabilistic triangulation, showing the distribution of GM components is as follows. (a) At the initial position. (b) At the final observer position.

From (33b), the GM weight, $w_{t,m}^{(j,k)}$, is scaled by $\ell(\boldsymbol{\omega}_{t,m} \,|\, \mathbf{r}_t)$ in (34), and hence by the likelihood of false DoA estimates, $\kappa(\boldsymbol{\omega}_{t,m} \,|\, \mathbf{r}_t)$. Therefore, the GM weight decreases as the number of false DoA estimates increases. To ensure stochastically relevant GM components, the numerator of (33b) must increase to counteract the scaling by the false DoA estimates. This is achieved if $\mathcal{N}(\boldsymbol{\omega}_{t,m} \,|\, \hat{g}_k(\mathbf{m}_{t|t-1}^{(j)}), \mathbf{S}_{t,m}^{(j,k)})$ is narrow and its mean is centred about the true source position. For $\mathbf{S}_{t,m}^{(j,k)}$ to be narrow, the GM component must be in the vicinity of DoA estimates over multiple, consecutive waypoints. In contrast, false DoA estimates due to reverberation are typically incoherent across time for moving observers. Therefore, (33b) enforces that components close to true sources become increasingly probable over time, whereas components initialized from false DoA estimates deplete.

It is crucial to note explicit modelling of $\kappa(\boldsymbol{\omega}_{t,m} \,|\, \mathbf{r}_t)$ facilitates robustness against false DoA estimates, e.g., due to strong reflections directed away from the sound sources.

C. Illustrative Example

To illustrate probabilistic triangulation, the proposed PHD filter for source mapping was evaluated for an observer path of three waypoints at $(1, 1, 1.8)$ m, $(1, 2, 1.8)$ m and $(1, 3, 1.8)$ m, within a $6 \times 6 \times 2.5$ m^3 room. A static source is placed at $(2, 2, 1.8)$ m. DoA estimates are simulated from (10) with an error of 10 deg in azimuth and inclination. For each time-step, the continuous single-source pdf is approximated by the kernel density estimate [33] of the samples.

Fig. 3(a) shows that the initial GM components are distributed as a cone along the DoA estimates. At the final waypoint, the GMM is concentrated about a small area surrounding the source position Fig. 3(b). Therefore, spatio-temporal diversity of the observer leads to convergence of the source components to a reduced area of stochastically likely positions.

V. aSLAM Observer Localization

This section proposes a method to probabilistically anchor the observer by fusing the motion reports and source map.

A. Particle Filter for Nonlinear, NonGaussian Observer States

As outlined in the Challenges in Section II, the observer is modelled by a non-linear, non-Gaussian state space. Therefore, the posterior pdf, $p(\mathbf{r}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\Omega}_{1:t})$ in (23) is analytically intractable. Particle filters [34] can be used to approximate the posterior pdf by sampling I particles, $\{\hat{\mathbf{r}}_t^{(i)}\}_{i=1}^{I}$, from a pre-specified Importance Sampling (IS) function [35], such that

$$p\left(\mathbf{r}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\Omega}_{1:t}\right) \approx \sum_{i=1}^{I} \alpha_t^{(i)} \, \delta_{\hat{\mathbf{r}}_t^{(i)}}(\mathbf{r}_t), \tag{36}$$

where $\alpha_t^{(i)}$ are the importance weights, and $\delta_{\hat{\mathbf{r}}_t^{(i)}}(\mathbf{r}_t)$ is the Dirac-Delta function centered at $\mathbf{r}_t$ and evaluated at $\hat{\mathbf{r}}_t^{(i)}$.

The prior pdf is often used as the IS function [35], but requires large numbers of particles for models affected by uncertainty. Optimal IS minimizes the estimation variance [35], and hence the computational overhead, but is itself intractable for estimation of $\mathbf{r}_t$. Nevertheless, as shown in (1), the observer state can be separated into the orientation, γ_t, and $\mathbf{p}_t$. Thus, the posterior pdf can be factorized as

$$p\left(\mathbf{r}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\Omega}_{1:t}\right) = p\left(\gamma_t \mid \mathbf{y}_{1:t,\gamma}, \boldsymbol{\Omega}_{1:t}\right) p\left(\mathbf{p}_t \mid \gamma_t, \mathbf{y}_{1:t,v}, \boldsymbol{\Omega}_{1:t}\right). \tag{37}$$

Although $\mathbf{p}_t$ is non-linearly dependent on the orientation through the dynamical model in (2), the state-space model of $\mathbf{p}_t$ in (1a) and (3a) is linear-Gaussian given a realization of γ_t. We therefore propose a marginalized IS scheme [36] that exploits the linear-Gaussian substructure for optimal importance sampling of $\mathbf{p}_t$, conditional on orientation particles drawn from an approximate IS function of γ_t.

The dynamical model and motion reports of the orientation presented in (1) and (3) correspond to wrapped Gaussian distributions. Therefore, the orientation particles, $\hat{\gamma}_t^{(i)}$, are sampled from the wrapped Kalman Filter (KF) [37]:

$$\hat{\gamma}_t^{(i)} \sim \mathcal{N}^w\left(\gamma_t \,|\, \mu_{t,\gamma}^{(i)}, \varsigma_{t,\gamma}^{(i)}\right), \qquad i = 1, \ldots, I \tag{38}$$

where the mean, $\mu_{t,\gamma}^{(i)}$, and variance, $\varsigma_{t,\gamma}^{(i)}$, in (38) are given by

$$\mu_{t,\gamma}^{(i)} = \vartheta\left(\mu_{t-1,\gamma}^{(i)} + \frac{\varsigma_{t-1,\gamma}^{(i)} + \sigma_{v,\gamma_t}^2}{\varsigma_{t-1,\gamma}^{(i)} + \sigma_{v,\gamma_t}^2 + \sigma_{w,\gamma_t}^2}\hat{\mu}_{t-1,\gamma}^{(i)}\right),$$

$$\varsigma_{t,\gamma}^{(i)} = \frac{\left(\varsigma_{t-1,\gamma}^{(i)} + \sigma_{v,\gamma_t}^2\right)\sigma_{w,\gamma_t}^2}{\varsigma_{t-1,\gamma}^{(i)} + \sigma_{v,\gamma_t}^2 + \sigma_{w,\gamma_t}^2},$$

and $\hat{\mu}_{t-1,\gamma}^{(i)}$ is the weighted average of wrapped hypotheses:

$$\hat{\mu}_{t-1,\gamma}^{(i)} = \sum_{k=-1}^{1} \nu_t^{(i,k)}\left(y_{t,\gamma} - \mu_{t-1,\gamma}^{(i)} + 2\pi\, k\right),$$

$$\nu_t^{(i,k)} = \mathcal{N}\left(y_{t,\gamma} \,\middle|\, \mu_{t-1,\gamma}^{(i)} - 2\pi\, k,\, \varsigma_{t-1,\gamma}^{(i)} + \sigma_{v,\gamma_t}^2 + \sigma_{w,\gamma_t}^2\right).$$

Conditional on each $\hat{\gamma}_t^{(i)}$, one particle, $\hat{\mathbf{p}}_t^{(i)}$ of $\mathbf{p}_t$, is sampled from the KF:

$$\hat{\mathbf{p}}_t^{(i)} \sim \mathcal{N}\left(\mathbf{p}_t \,\middle|\, \boldsymbol{\mu}_{t,\mathbf{p}}^{(i)},\, \boldsymbol{\Sigma}_{t,\mathbf{p}}^{(i)}\right), \qquad i = 1, \ldots, I \tag{39}$$

where $\boldsymbol{\mu}_{t,\mathbf{p}}^{(i)}$ and $\boldsymbol{\Sigma}_{t,\mathbf{p}}^{(i)}$, are the KF mean and covariance [31] corresponding to the model in (1a) and (3a).

Using (38) and (39), $\alpha_t^{(i)}$ in (36) is given by

$$\alpha_t^{(i)} = \alpha_{t-1}^{(i)}\,\mathcal{N}\left(y_{t,v} \,\middle|\, \mathbf{h}\,\boldsymbol{\mu}_{t,\mathbf{p}}^{(i)},\, \mathbf{h}\,\boldsymbol{\Sigma}_{t,\mathbf{p}}^{(i)}\,\mathbf{h}^T\right)\sum_{k=-1}^{1}\nu_t^{(i,k)}. \tag{40}$$

where $\hat{\mathbf{r}}_t^{(i)} \triangleq [\,[\hat{\mathbf{p}}_t^{(i)}]^T, \hat{\gamma}_t^{(i)}\,]^T$.

B. Importance Weights for Probabilistic Anchoring

The joint aSLAM posterior PHD is obtained by solving (23) using the observer posterior pdf in (36) and the multi-source PHD in (35). For this derivation, a closed-form expression of the integral in the denominator of (23) is required. Inserting (36) into (23), the integral in the denominator reduces to

$$\iint \sum_{i=1}^{I} \alpha_t^{(i)}\delta_{\hat{\gamma}_t^{(i)}}(\gamma_t)\,\delta_{\hat{\mathbf{p}}_t^{(i)}}(\mathbf{p}_t)\,\mathcal{L}\left(\boldsymbol{\Omega}_t \mid \gamma_t, \mathbf{p}_t\right) d\gamma_t\, d\mathbf{p}_t$$

$$= \sum_{i=1}^{I} \alpha_t^{(i)}\,\mathcal{L}\left(\boldsymbol{\Omega}_t \mid \hat{\gamma}_t^{(i)}, \hat{\mathbf{p}}_t^{(i)}\right). \tag{41}$$

Therefore, although the evidence of the set of DoA estimates in (24) is non-linearly dependent on the observer states through (34) and (29), the point mass approximation in (36) facilitates a closed-form solution of the integrals in (41).

Using (41), (36), (35) in (23), the posterior SLAM PHD is

$$\lambda\left(\mathbf{r}_t, \mathbf{S}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\Omega}_{1:t}\right) = \sum_{i=1}^{I} \beta_t^{(i)}\,\delta_{\hat{\mathbf{r}}_t^{(i)}}(\mathbf{r}_t)\,\lambda\left(\mathbf{s}_t \mid \hat{\mathbf{r}}_t^{(i)}, \boldsymbol{\Omega}_{1:t}\right),$$

where the weights, $\beta_t^{(i)}$, are given by

$$\beta_t^{(i)} \triangleq \frac{\alpha_t^{(i)}\,\mathcal{L}\left(\boldsymbol{\Omega}_t \mid \hat{\gamma}_t^{(i)}, \hat{\mathbf{p}}_t^{(i)}\right)}{\sum_{j=1}^{I}\alpha_t^{(j)}\,\mathcal{L}\left(\boldsymbol{\Omega}_t \mid \hat{\gamma}_t^{(j)}, \hat{\mathbf{p}}_t^{(j)}\right)}. \tag{42}$$

Algorithm 1: aSLAM.

for $i = 1, \ldots, I$ **do**
  Sample $\hat{\mathbf{r}}_t^{(i)}$ using (38),(39);
  for $j = 1, \ldots, J_{t-1}$ **do**
    Evaluate $\mathbf{m}_{t-1}^{(i,j)}, \boldsymbol{\Sigma}_{t-1}^{(i,j)}$ relative to $\hat{\mathbf{r}}_t^{(i)}$ (29)
    Predict $\mathbf{m}_{t|t-1}^{(i,j)}, \boldsymbol{\Sigma}_{t|t-1}^{(i,j)}, w_{t|t-1}^{(i,j)}$ (31a);
  end for
  for $m = 1, \ldots, M_t$ **do**
    for $j = 1, \ldots, J_b$ **do**
      Sample $\hat{r}_{t,m}^{(i,j)}, \hat{\boldsymbol{\omega}}_{t,m}^{(i,j)}$ (25), (26);
      Evaluate $\mathbf{m}_{b,t,m}^{(i,j)}, \boldsymbol{\Sigma}_{b,t,m}^{(i,j)}, w_{b,t,m}^{(i,j)}$ (27);
    end for
    for $j = 1, \ldots, J_{t-1}$ **do**
      for $k = -1, 0, 1$ **do**
        Update $\mathbf{m}_{t,m}^{(i,j,k)}, \boldsymbol{\Sigma}_{t,m}^{(i,j,k)}, w_{t,m}^{(i,j,k)}$ (31), (33b);
      end for
    end for
    Evaluate $\ell(\boldsymbol{\omega}_{t,m}|\hat{\mathbf{r}}_t^{(i)})$ (34);
  end for
  Evaluate $\mathcal{L}(\boldsymbol{\Omega}_t|\hat{\mathbf{r}}_t^{(i)})$ (24);
  GM reduction [31];
  Evaluate $\beta_t^{(i)}$ (42);
end for
Resampling [32];

Each observer particle is therefore weighted by 1) the report evidence terms in (40), accounting for the uncertainty in the observer motion and reports, and 2) the evidence of the set of DoA estimates, aligning the observer-relative DoAs with the source tracks. aSLAM is summarized in Algorithm 1.

C. Illustrative Example

The following illustrative example provides insight into the physical significance of the terms in (42). Two static sources are positioned at $\mathbf{s}_1 = (4, 3, 1.8)$ m and $\mathbf{s}_2 = (2, 2, 1.8)$ m in a $6 \times 6 \times 2.5$ m^3 room. The observer is positioned at $(5, 2, 1.8)$ m at $t-1$ and moves to $(5, 3.5, 1.8)$ m at t. The report errors and DoA errors are arbitrarily chosen as $\sigma_{w,\gamma_t} = 30$ deg, $\sigma_{w,v_t} = 1$ m/s, and 25 deg measurement noise in the detected azimuth and inclination respectively. Furthermore, $p_d = 1$ and $\lambda_c = 0$. Candidate observer positions are sampled in x- and y-coordinates on a deterministic grid of 0.05 m resolution within the room boundaries. Assuming that the source positions at $t-1$ are known with $\boldsymbol{\Sigma}_{t-1}^{(j)} = \text{diag}[0.2, 0.2, 0.05]$ for $j = 1, 2$, the source EKF, and the evidence terms of the single DoA estimates as well as the set of DoA estimates are evaluated from (31)–(34). Using the results, (34), (24), (40) and (42) are evaluated for each observer candidate in the grid.

A contour plot of the evidence of $\boldsymbol{\omega}_{t,1}$ in (34) across all observer candidates is plotted in Fig. 4(a). The "confidence area", i.e., the area within the contour corresponding to evidence values ≥ 0.1, is 8.98 m^2.[1] The results are compared in Fig. 4(b)

[1]The confidence area is evaluated based on the distances between vertices along the contour line.

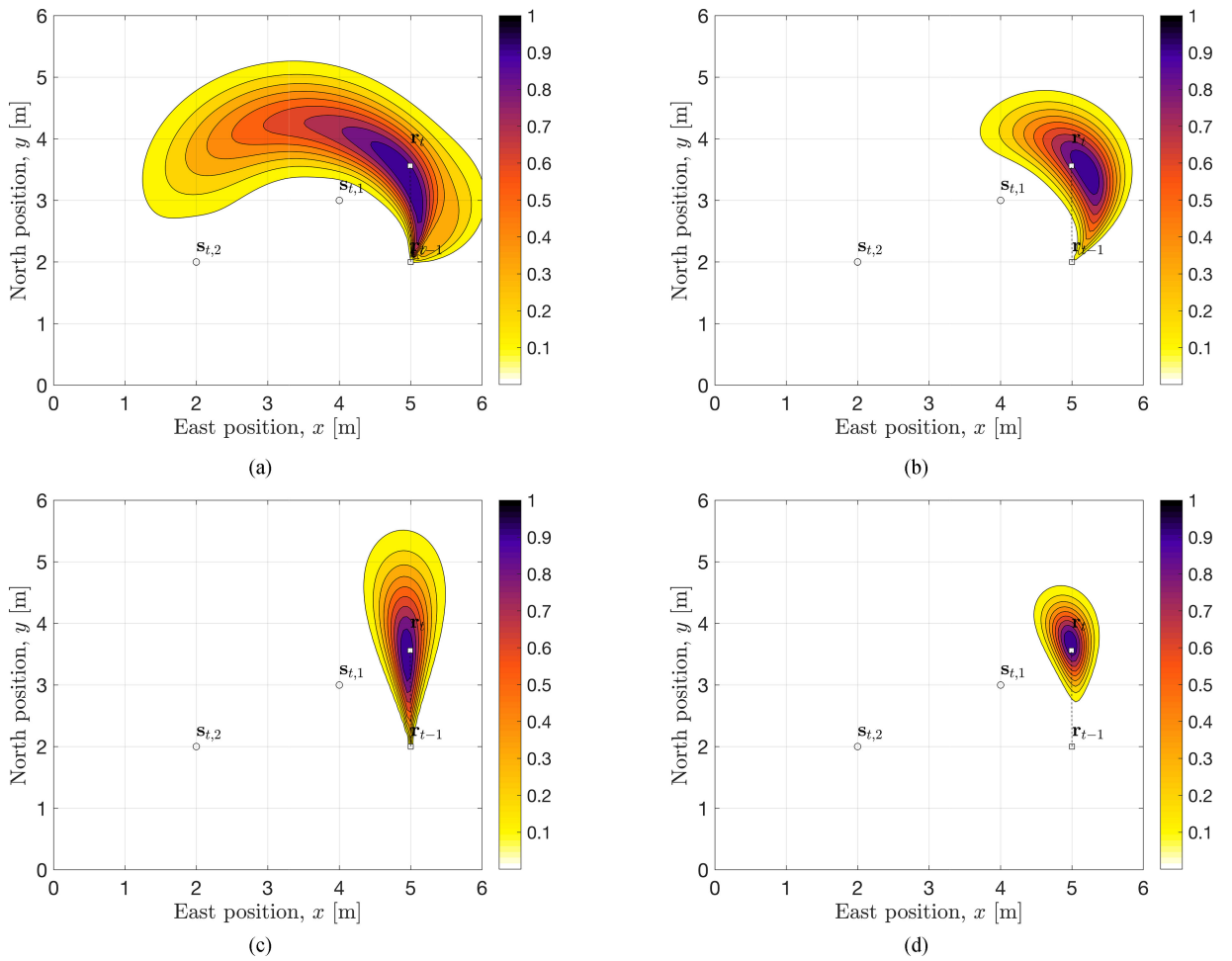


Fig. 4. Probabilistic observer anchoring: Areas of probable observer positions using the evidence of the following. (a) Single DoA estimate. (b) Set of DoA estimates. (c) Motion report of the observer. (d) aSLAM importance weights, equivalent to the intersection of (b) and (c). Contour colors indicate corresponding normalized values of the pdf.

to the evidence of the *set* of DoA estimates in (24), and corresponding to a confidence area of 3.14 m^2. The area of probable observer positions is therefore reduced by 5.84 m^2 by combining the observer information inferred from all sources in the scene map. Fig. 4(c) shows the contour plot corresponding to the observer evidence in (40), which corresponds to a confidence area of 2.56 m^2. The confidence area corresponding to the aSLAM weights in (42), i.e., the intersection of the evidence terms of the source DoAs and the observer motion report, is shown in Fig. 4(d). The results highlight an area of probable observer positions of 1.17 m^2. Therefore, by combining the observer information gleaned from multiple sources with models accounting for the observer motion and expected report error, probabilistic anchoring results in a reduction of 7.81 m^2 compared to Fig. 4(a).

VI. Experimental Setup

A. Application Scenario

aSLAM presents a general framework, suitable for any application involving mobile acoustic sensors for scene mapping. The following considers the application of robotic systems, typically equipped with a microphone array and Inertial Measurement Unit (IMU). Drift in gyroscopes results in cumulative errors in the orientation and velocity reports. In addition, errors between the instructed and executed observer motion are caused by, e.g., wear of mechanical parts with robot age [[38], Chapter 15]. Navigation errors of 33% over a travelled distance of 0.75 m are reported for a commercial robot in [12].

The first set of experiments is designed to decouple the aSLAM performance from any specific DoA estimator. An 'oracle localizer' is used to control and investigate the impact of the observer motion, motion report and DoA estimation errors. The second set of experiments evaluates aSLAM performance in realistic acoustic environments using room simulations of a moving spherical microphone array integrated in the robot head [7].

The observer path is simulated from (1) for 100 time steps with $\Delta_t = 0.25$ s and $\sigma_{v,\gamma_t} = 45$ deg in a $6 \times 6 \times 2.5$ m^3 room. The initial x- and y-coordinates of the observer are at the room center. The height of the robot's sensor is 1.20 m, is assumed known, and corresponds to the height of a Pepper robot by Softbank Robotics. The initial sensor state is assumed known with an arbitrarily chosen accuracy of 0.1 m in position and

3 deg in orientation. To ensure that the sensor is positioned within the room volume, a maximum turn is enforced by letting $v_{t,\gamma} = \sigma^2_{v,\gamma_t}$ at any t for which $\mathbf{p}_{t-1}$ is within 1 m of the room boundaries. Three sources are simulated from (4) for randomly sampled source heights between $[1.6, 1.95]$ m, representative of human height. Each source is placed at the centre of a randomly selected quadrant in the room.

B. Oracle Localizer

The observer reports are generated from (3). DoA estimates are generated from (10) with $\lambda_c = 0$ and $p_d = 1$. The covariance terms, σ^2_{w,γ_t}, σ^2_{w,v_t}, and $\mathbf{R}_{t,m}$ are assumed known.

Experiment 1: Experiment 1 evaluates the robustness of aSLAM against errors in the reported observer velocity. The experiment is evaluated for $\sigma_{w,v_t} = 0.25, 0.75, 1.5$ m/s. The sensor speed is set to a typical walking speed of $v_t = 1.5$ m/s, and $\sigma_{w,\gamma_t} = 0$ deg. The DoA estimation error is 5 deg in azimuth and inclination, reflecting state-of-the-art performance [6]. Since σ_{w,γ_t} is negligible, (1)–(3) reduce to a linear system. Experiment 1 thus is evaluated using $I = 1$ particles.

Experiment 2: Experiment 1 is repeated for $\sigma_{w,\gamma_t} = 2.5, 5, 10$ deg and $\sigma_{w,v_t} = 0.75$ m/s, chosen to reflect the findings in [12]. As $\sigma_{w,\gamma_t} \neq 0$, the observer model is non-linear. Although particle filters are guaranteed to converge as $I \to \infty$ [39], computational constraints enforce finite I. The experiment also investigates the number of required particles.

Experiment 3: The observer path dictates the spatio-temporal diversity between consecutive waypoints. As detailed in Section IV, the disparity between waypoints is fundamental for range inference. However, in many practical applications, the observer speed is limited by hardware constraints as well as the available space for manœuvers within the environment. Experiment 3 thus investigates the effect of increasing $v_t = 0.5, 1.5, 2$ m/s with $\sigma_{w,v_t} = 0.75$ m/s and $\sigma_{w,\gamma_t} = 5$ deg.

Experiment 4: DoA estimation accuracy depends on the choice of the DoA estimator, reverberation / noise levels, and the sensor geometry. Experiment 4 thus investigates the impact of increasing DoA errors for $\sigma_{\text{DoA}} = 5, 10, 15$ deg.

C. Room Simulations

The response of a rigid spherical array [40], [41] based on the mh-acoustics Eigenmike is simulated using the image source method [42], [43] with $T_{60} = 0.5$ s at sampling frequency $f_s = 4$ kHz and spherical harmonic order of $N_{\text{harm}} = 3$. The sensor is moved in a "stop-perceive-act" motion [44]: the observer moves for 0.5 s to a new waypoint, from which the reverberant signals are simulated over a measurement period of 0.5 s. The simulated RIRs are convolved with TIMIT sentences [45]. For DoA estimation, Direct-Path Dominance Multiple Signal Classification (DPD-MUSIC) [5] is used, which maximizes the spatial spectra of the signal subspace in the spherical harmonic domain. The angular resolution is 2 deg, with a Direct-Path Dominance (DPD) threshold of 2, and smoothing across 2 frames in time and 15 bins in frequency [46]. The reverberant signals are bandlimited between $0.5 - 1.5$ kHz to avoid noise amplification due to mode strength compensation at low frequencies and spatial aliasing at high frequencies. The Multiple Signal Classification (MUSIC) spectrum is evaluated and its highest peak extracted for each time-frequency bin in the measurement period that passes the DPD-test. All peaks within the measurement period are clustered by fitting a von Mises-Fisher (vMF) mixture model [47] using the algorithm in [48]. As the number of sources is unknown but required for model fitting, the following model selection scheme [49] is used: a vMF mixture model is evaluated for each hypothesis in $N = 1, \ldots, N_{\max}$, where $N_{\max} = 15$ is chosen to avoid undermodelling. The DoA estimates are given by the mixture model that results in the minimum message length [50].

D. aSLAM Implementation Aspects

Systematic resampling is applied to the sensor particles to avoid particle depletion [33]. Mixture reduction [[51], Table II] is used to avoid an exponential explosion of the GMM. The observer point estimate is extracted as the weighted average of all particles. Point estimates of the source positions are obtained by clustering the GM components into $\hat{N}_t + 1$ clusters using [49]. The additional cluster is a diffuse cluster that absorbs outliers. The centroid of the diffuse cluster is the room center and its covariance corresponds to the room volume.

E. Performance Metrics

Accuracy of the observer estimates compared to the ground truth is evaluated using the Euclidean distance, defined as $d(\mathbf{x}_t, \mathbf{y}_t) = \|\mathbf{x}_t - \mathbf{y}_t\|$ between two positions, $\mathbf{x}_t$ and $\mathbf{y}_t$. The Optimal Subpattern Assignment (OSPA) distance [52], [53], $\Delta(\mathbf{X}_t, \mathbf{Y}_t)$, is used as a measure of source mapping accuracy, continuity, and false track initialization. Given $\mathbf{X}_t \triangleq \{\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,N}\}$ and $\mathbf{Y}_t \triangleq \{\mathbf{y}_{t,1}, \ldots, \mathbf{y}_{t,M}\}$, the OSPA is defined as

$$\Delta(\mathbf{X}_t, \mathbf{Y}_t) \triangleq \left[\frac{1}{N} \min_{\pi \in \mathbf{\Pi}_N} \sum_{i=1}^{M} d_c(\mathbf{x}_{t,i}, \mathbf{y}_{t,\pi(i)})^p + (N - M)c^p\right]^{\frac{1}{p}}$$

where $1 \leq p < \infty$ is the OSPA metric order parameter, $\mathbf{\Pi}_N$ is the set of permutations of length M with elements from $\{1, \ldots, N\}$, and $d_c(\mathbf{x}_{t,i}, \mathbf{y}_{t,\pi(i)}) \triangleq \min(c, d(\mathbf{x}_{t,i}, \mathbf{y}_{t,\pi(i)}))$ is the distance between two tracks with cutoff value, c. A first-order metric of $p = 1$ is used to facilitate that the total metric corresponds to the sum of the mapping and cardinality errors [54]. A cutoff value of $c = 1$ m is used. The cutoff was selected so that the aSLAM metrics are not affected, but the maximum error of the benchmark algorithm can be limited for improved readability of the graphs in Section VII.

VII. Results

A. Oracle Localizer

1) Experiment 1 - Observer Velocity Error: Fig. 5 shows the estimated scene map for $\sigma_{w,v_t} = 1.5$ m/s after 0.75 s, 3.25 s, and 25 s. The results after 0.75 s in Fig. 5(a) show that the source GM components are distributed in a conical volume along each DoA. Averaged across all Monte Carlo (MC) runs, the source

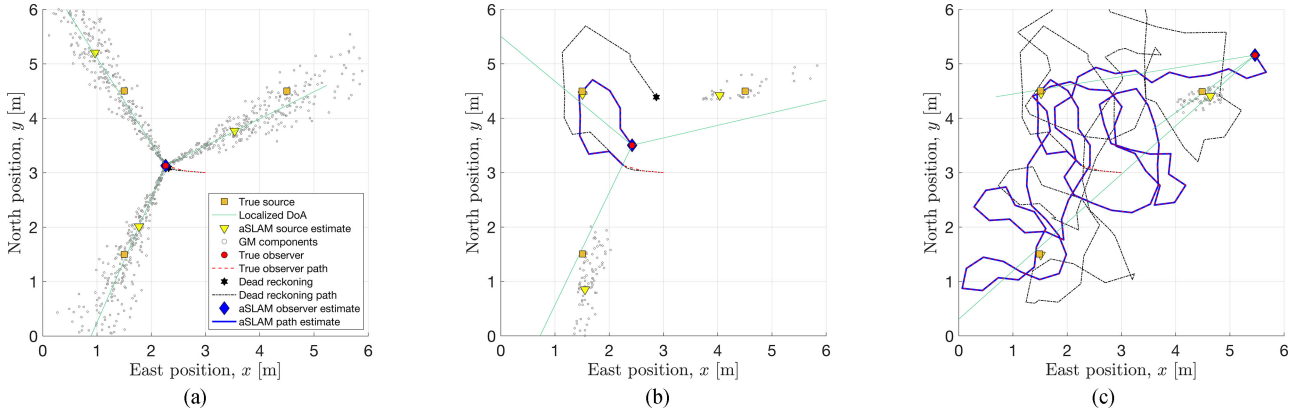


Fig. 5. Experiment 1: Evolution of aSLAM map estimate for $\sigma_{w,v_t} = 1.5$ m/s at the following time durations. (a) 0.75 s. (b) 3.25 s. (c) 25 s.

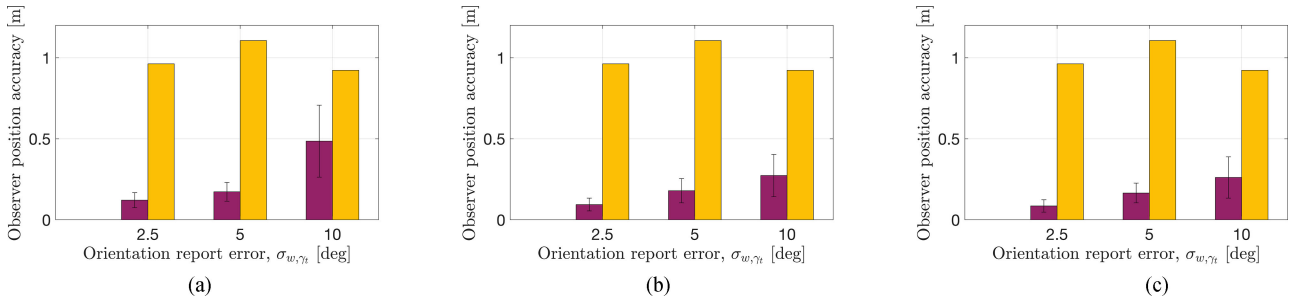


Fig. 6. Experiment 2: Observer localization accuracy, averaged over time, and Monte Carlo runs, of aSLAM (purple) and dead reckoning (orange). (a) $I = 5$. (b) $I = 50$. (c) $I = 100$.

estimates correspond to an OSPA metric of 0.56 m at 0.75 s. After 3.25 s [Fig. 5(b)], the source estimates lead to an improved OSPA of 0.26 m, and converge to within 0.15 m of the ground truth after 25 s.

The results in Fig. 5 also compare the estimated observer trajectory to the ground truth and dead reckoning [13] estimates. The results after 25 s in Fig. 5(c) highlight that the estimated observer trajectory is identical to the ground truth, whereas the dead reckoning trajectory diverges. Across all settings of σ_{w,v_t}, the Euclidean distance, averaged across time and MC runs, between the observer estimates and ground truth is found as 0 m. The dead reckoning estimates correspond to average errors of 0.37 m, 0.71 m, and 1.32 m for $\sigma_{w,v_t} = 0.25, 0.75, 1.5$ m/s respectively. Therefore, optimal estimation of the velocity allows for perfect localization of the observer if reports are affected by velocity errors only.

2) Experiment 2 - Observer Orientation Error: The results in Fig. 6 compare the aSLAM observer localization accuracy against the dead reckoning performance, both averaged across time and MC runs. In Fig. 6(a), the Euclidean distance between the observer estimates and ground truth, is provided for $I = 5$ particles. The average Euclidean distance values correspond to 0.12 m, 0.17 m and 0.49 m for $\sigma_{w,\gamma_t} = 2.5, 5, 10$ deg respectively. Compared to dead reckoning, aSLAM provides an improvement of 0.84 m, 0.94 m, and 0.43 m for $\sigma_{w,\gamma_t} = 2.5, 5, 10$ deg respectively.

The experiment is repeated for 50 and 100 particles to investigate if a further improvement can be achieved. The results for $I = 50$ particles are given in Fig. 6(b), showing that the localization accuracy remains approximately constant for $\sigma_{w,\gamma_t} = 2.5, 5$ deg, whereas the Euclidean distance for $\sigma_{w,\gamma_t} = 10$ deg is decreased by 55% to 0.27 m. These results remain unchanged for $I = 100$ [see Fig. 6(c)], i.e., steady-state for the simulated scenario is reached for $I \leq 50$.

Therefore, an adequate number of particles is required to sample the area of likely observer states with sufficient resolution. The reason for an increasing number of required particles is that the range of stochastically likely hypotheses of γ_t widens as σ_{w,γ_t} increases [see Fig. 4(c)].

3) Experiment 3 - Observer Speed: The results in Fig. 7(a)–(c) compare the estimated source range relative to the observer estimates to the ground truth range for $v_t = 0.5, 1.5, 2$ m/s. Fig. 7(a) shows that for $v_t = 0.5$ m/s all three sources achieve temporary periods of convergence, e.g., between 23.75 s for source 1 (purple) and between $12.5 - 16.25$ s for source 2 (yellow). The most distant source 3 (orange) temporarily converges between $9 - 10.75$ s, but is affected by an estimation bias of $0.15 - 0.26$ m. All three sources are affected by outliers and periods of missing DoA estimates, leading to mean range estimation errors of 0.43 m for source 1, up to 0.55 m for source 2, and up to 0.48 m for source 3.

As shown in Fig. 7(b), increasing the observer speed to $v_t = 1.5$ m/s leads to improved range estimates for close-range sources. Source 3, the nearest source towards the beginning of the simulation, converges after 7.75 s, with a mean range error of 0.14 m. Source 2 converges to within 0.25 m on average of the ground truth. The most distant source leads to an average error of 0.47 m. By further increasing the observer speed to 2.0 m/s,

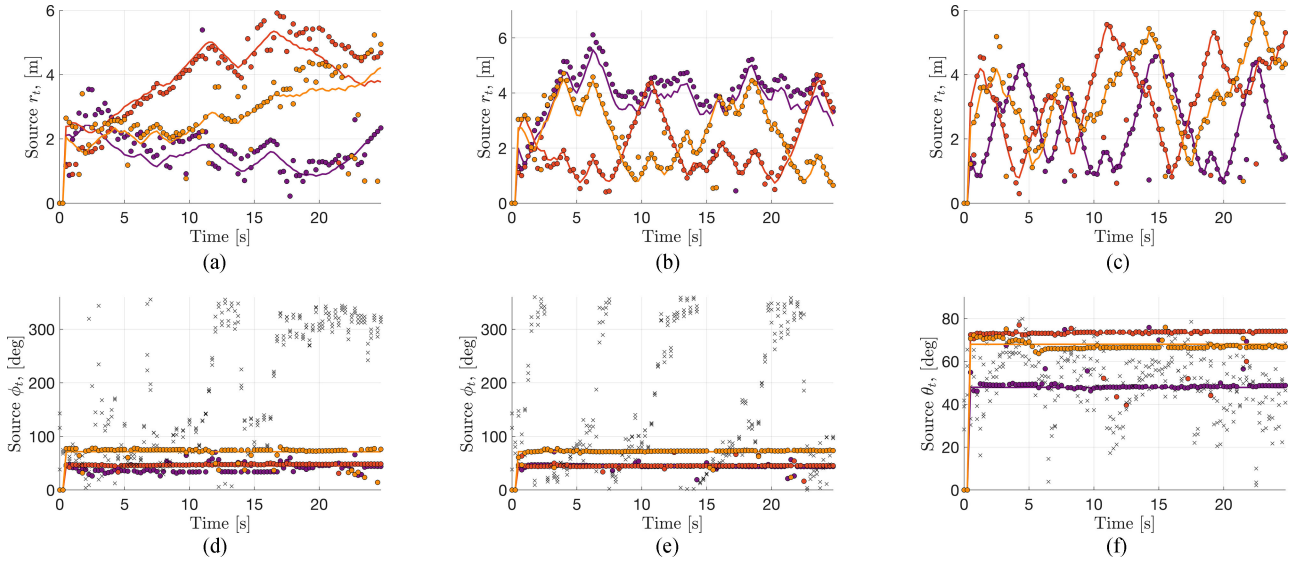


Fig. 7. Experiment 3: Comparison of ground truth (line), aSLAM (circles), and DoA estimates (crosses) for one Monte Carlo run is as follows. (a)–(c) Range for $v_t = 0.5, 1.5, 2$ m/s respectively. (d)–(e) Azimuth for $v_t = 0.5, 2$ m/s respectively. (f) Inclination for $v_t = 2$ m/s. Source 1: Purple; source 2: Yellow; source 3: Orange.

all three sources converge to within 0.3 m of the ground truth, with an average error of 0.2 m for source 1, 0.3 m for source 3, and 0.22 m for source 2 [Fig. 7(c)].

Fig. 7(d) and (e) compare the source azimuth aSLAM estimates with the ground truth in the global reference frame. For $v_t = 0.5$ m/s, source 1 corresponds to an average estimation error of 8.22 deg, source 2 results in 7.47 deg and source 3 corresponds to 2.42 deg mean errors. By increasing the speed to $v_t = 2$ m/s, the azimuth estimation error for source 1 and 2 are reduced by 5.64 deg and 4.35 deg respectively.

As also shown in Fig. 7(d) and (e), the azimuth corresponding to the mapped source positions is significantly improved over the DoA estimates, whose values are spread over the entire unit circle.[2] A similar trend is observed in Fig. 7(f), showing DoA estimation errors in inclination of up to 45 deg.

Therefore, the results show that by exploiting the observer's spatio-temporal diversity, aSLAM successfully solves the underdetermined problem of 3D source position estimation from the 2D DoA estimates. Moreover, informative planning of the observer motion can be used to optimize source mapping.

4) Experiment 4 - Source DoA error: We now discuss the results for Experiment 4, investigating the impact of increasing DoA errors for $\sigma_{\text{DoA}} = 5, 10, 15$ deg. The results in Fig. 9 depict the estimated azimuth accuracy for sources 1 and 2. The discussion of the results requires additional explanation for the underlying effects to be clear. Therefore, a birdseye view of the ground truth observer path and source positions is shown in Fig. 8. The observer initially moves away from source 1 between the period between $[0, 2.5]$ s, resulting in increasing azimuth errors for source 1 for all three settings of σ_{DoA}. Between the period of $[2.5, 10]$ s, the observer approaches and moves around source 1, leading to an improvement in azimuth accuracy, e.g., from 13.88 deg to 3.17 deg for $\sigma_{\text{DoA}} = 10$ deg.

[2]The DoA estimates were transformed to the global reference frame using the corresponding dead reckoning observer state.

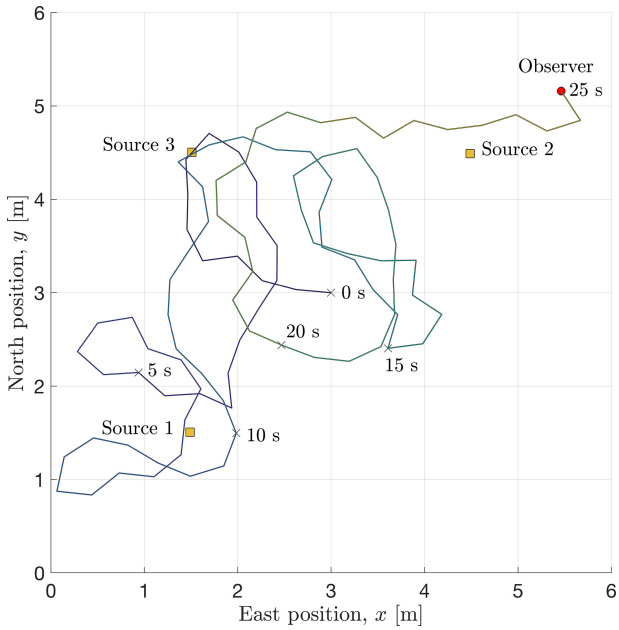


Fig. 8. Experiment 4: Birdseye view of scenario. Color gradient of observer path indicates progression time.

After 10 s, the observer moves away from source 1 and towards source 2, leading to gradually increasing azimuth errors for all three settings of σ_{DoA}. For example, the source estimates for $\sigma_{\text{DoA}} = 5$ deg deteriorate from 2.35 deg to 3.67 deg. Therefore, for DoA estimates that are only affected by estimation errors, the azimuth accuracy of the aSLAM estimates improves with decreasing source-sensor range. The impact of false and missing DoA estimates combined with estimation errors, as encountered in realistic acoustic scenarios, are discussed in Section VII-B.

The azimuth accuracy is also correlated with the spatial diversity between waypoints of the observer path relative to the source. Contrary to source 1, source 2 in Fig. 9(b) results in

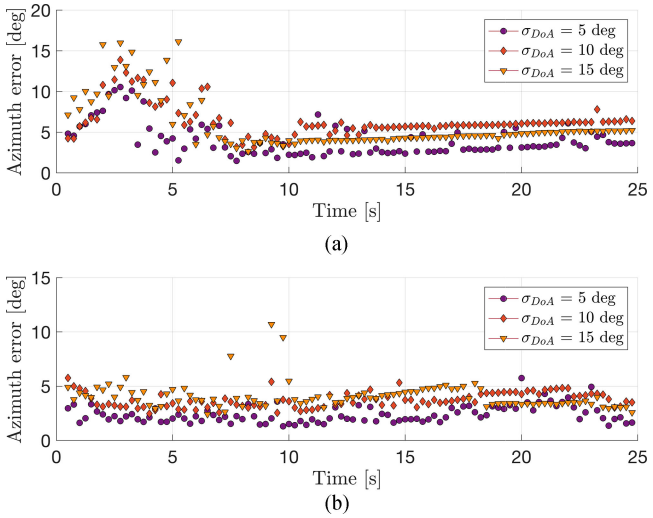


Fig. 9. Experiment 4: Azimuth error. (a) Source 1. (b) Source 2.

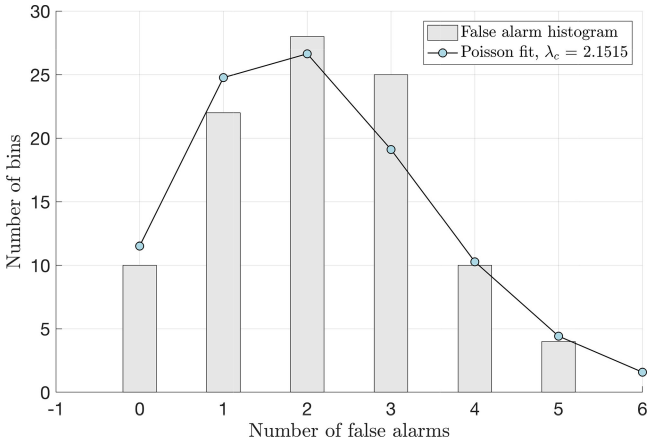


Fig. 10. Room Simulation: Histogram of the distribution of false DoA estimates (bar plot) compared to fit by Poisson distribution (line).

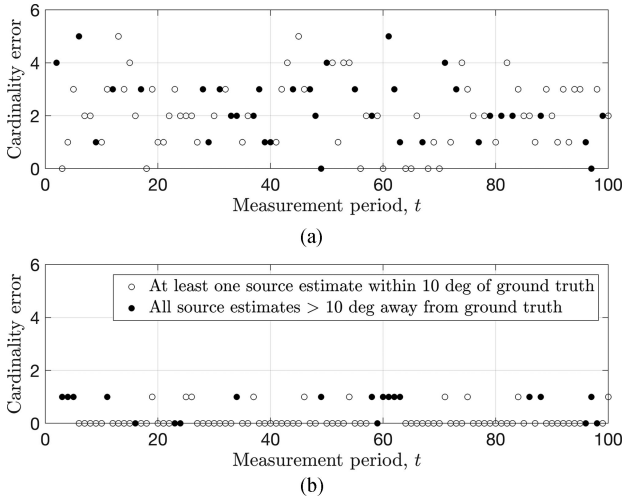


Fig. 11. Room simulation: Comparison of cardinality error between the single ground truth source and the following. (a) Number of DPD-MUSIC DoA estimates. (b) aSLAM source estimates.

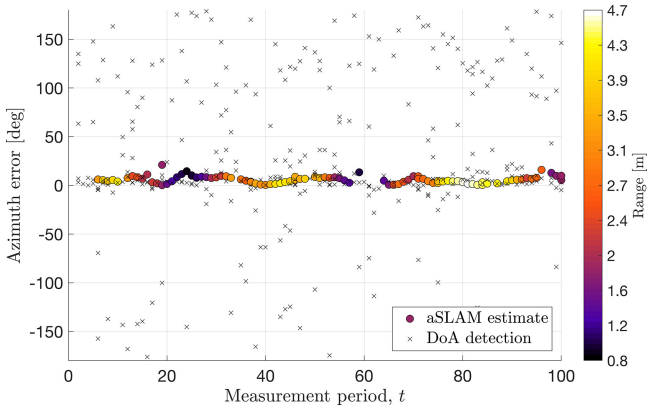


Fig. 12. Room Simulation: Azimuth error of the DoA and aSLAM estimates compared to the ground truth. Fill color: Ground truth source-sensor range.

improving azimuth accuracies in the period between $[0, 2]$ s, e.g., for $\sigma_{\text{DoA}} = 10$ deg from 5.76 deg to 3.12 deg. Although the observer path starts at equal range to all three sources, the observer path corresponds to a tangential motion with respect to source 2 for the first 2 s of the experiment. In contrast, between $[1, 2]$ s, the observer moves in an approximately straight line away from source 1. Therefore, the azimuth accuracy of the aSLAM estimates is also sensitive to the angle of approach towards, or away from, a source.

Due to the time-varying source-sensor geometry in scenarios involving moving observers, the observer trajectory also affects convergence of the source position estimates. For example, in the period between $[13.5, 18.5]$ s, the observer path corresponds to a tangential motion relative to source 2, followed by moving away from the source. Hence, the azimuth error of the source 2 estimates for $\sigma_{\text{DoA}} = 15$ deg increases by 1.13 deg from 4.15 deg to 5.28 deg. As the observer re-enters a tangential motion at 18.5 s, the improved spatial diversity between waypoints facilitates correction of the diverging source estimates, resulting in an abrupt change in azimuth accuracy from 5.28 deg to 3.12 deg. In contrast, due to the improved DoA estimation accuracy for $\sigma_{\text{DoA}} = 10$ deg compared to $\sigma_{\text{DoA}} = 15$ deg, the source estimates are less affected by the motion of the observer, leading to a degradation by 0.47 deg in azimuth errors from 3.81 deg to 4.28 deg between $[13.5, 18.5]$ s. However, the divergence is not sufficiently severe to result in an abrupt correction. As a consequence, the source 2 estimates for $\sigma_{\text{DoA}} = 15$ deg outperform the estimates for $\sigma_{\text{DoA}} = 10$ deg after 18.5 s. A similar trend is observed for source 1 after 10.75 s.

B. Room Simulations

This subsection analyzes the aSLAM performance for the room simulations of a moving spherical microphone array using DPD-MUSIC for DoA estimation. The histogram of the false DoA estimates are shown in Fig. 10, highlighting that DPD-MUSIC results in $1 - 5$ false DoA estimates in 88% of the measurement periods. To validate the model of false DoA estimates in (22b), a Poisson distribution was fit to the false DoA estimates, resulting in the distribution in Fig. 10, corresponding to $\lambda_c = 2.15$. To evaluate the goodness-of-fit, the χ^2-test [55], using the standard choice of the significance level at 0.05, does

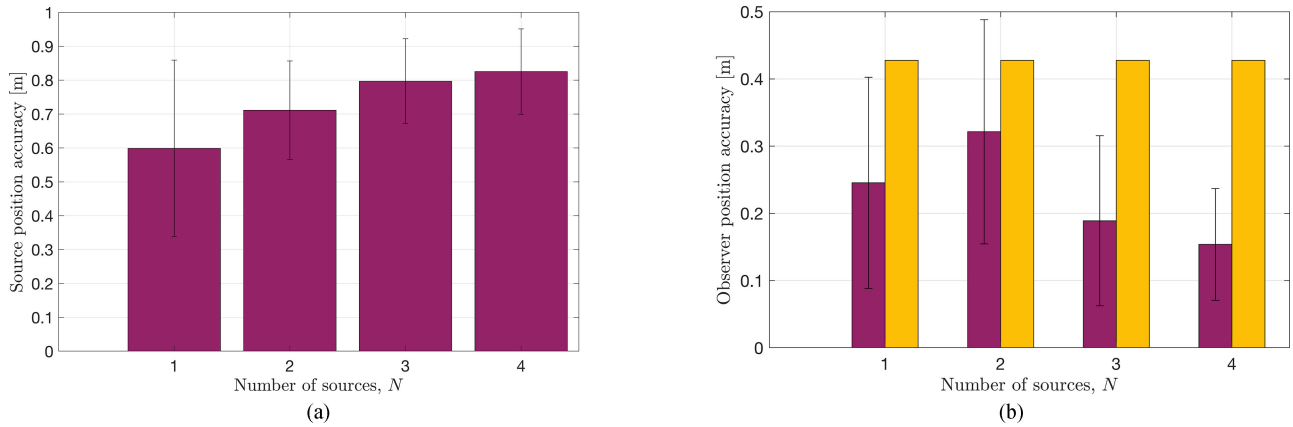


Fig. 13. Room simulation: Accuracy averaged across time for $N = 1, \ldots, 4$ sources of the following. (a) aSLAM observer estimates (purple) compared to dead reckoning (orange). (b) aSLAM source map estimate.

not reject the null-hypothesis that the false DoA estimates are Poisson distributed.

To emphasize the challenging conditions of the scenario, Fig. 11(a) shows the cardinality error between the single-source ground truth and the number of DPD-MUSIC DoA estimates for each measurement period. The figure also highlights the measurement periods in which at least one of the DPD-MUSIC estimates is within 10 deg of the true source direction. The threshold of 10 deg was chosen according to the worst-case localization performance for recordings from a spherical microphone array as reported in [6]. The probability of source detection hence corresponds to $p_d = 65.66\%$.

aSLAM is evaluated using p_d and λ_c as per the findings above. The results in Fig. 12 show the error of the source azimuth aSLAM estimate for $N = 1$, which is also compared to the error of the DPD-MUSIC DoA estimates. Despite the large number of false and missing DoA estimates, the results highlight that the source azimuth is estimated with an average error of 5.99 deg. The mean OSPA distance of the source position estimates corresponds to 0.6 m [see Fig. 13(a)].

The results in Fig. 12 also provide the ground truth source-sensor range for each measurement period using the marker fill color. The results in Section VII-A4 showed that - among other factors - an increase in source-sensor range is correlated to some extent with an improvement in azimuth accuracy of the aSLAM estimates. This finding is supported by the results in Fig. 12 during the measurement periods between, e.g., $t = [13, 20]$ and $t = [50, 57]$. As shown in Section VII-A for the oracle localizer, the mapping accuracy for an observer moving towards a source also depends on the observer speed and angle of approach towards the source. In addition, in realistic acoustic conditions, proximity to walls or other reflecting surfaces may affect the number of missing and false DoA estimates. Therefore, due to the spatio-temporal variation of the observer position within the acoustic enclosure, the azimuth error may temporarily increase for decreasing source-sensor range, as observed in, e.g., $t = [21, 24]$ of Fig. 12.

Furthermore, the source mapping accuracy is decreased by increasing the number of sources, indicated by the increase in OSPA in Fig. 13(a). This finding is intuitively expected since the inter-source distance decreases, rendering the sources less resolvable. Furthermore, increasingly many sources may be located in the far-field, affected by signal diffuseness. Nevertheless, the observer localization accuracy increases with larger choices of N, highlighted by the decreasing Euclidean distance in Fig. 13(b). Since an increasingly large number of sources bears additional spatial information about the observer, additional sources can be used to anchor the observer using the DoA estimates as captured in (42) and (24). Therefore, the tradeoff between source mapping and observer localization accuracy is affected by the number of sources in the scene.

VIII. Conclusion

This paper proposed a novel aSLAM approach that passively localizes a moving observer and simultaneously maps the positions of surrounding sound sources. Using DoA estimates at its input, aSLAM maps the source positions across time and space by probabilistic triangulation that exploits the robustness of the PHD filter against false DoA estimates in reverberation. The observer position is probabilistically anchored using a marginalized particle filter by fusing the motion reports with observer information inferred from the acoustic source map.

aSLAM performance was evaluated using simulations based on an oracle localizer in order to decouple aSLAM from any specific DoA estimator or acoustic sensor / array. Furthermore, realistic room simulations for the example application of a moving robot were used to evaluate and verify aSLAM performance in realistic acoustic conditions using a spherical microphone array and DPD-MUSIC for DoA estimation.

It was shown that aSLAM accurately reconstructs the observer path and jointly infers the 3D source positions from the 2D DoA estimates by exploiting the observer's spatio-temporal diversity. The results demonstrated that the source-sensor range for sources at $1 - 4$ m distance can be estimated with 0.14 m accuracy for an observer moving at human walking speed. Furthermore, the results demonstrated that aSLAM is robust to DoA estimation errors as well as false and missing DoA

estimates, caused by reverberation, interference and inactivity of non-continuous sources. Observer localization is further improved as the number of sources increases.

Interference from multiple, simultaneously active sources leads to significant performance degradation for classical systems that involve static arrays and rely only on DoA estimation. This degradation has severe consequences on algorithms, such as blind source separation, that require accurate estimates of the source directions. In contrast, the results in this paper highlighted that systems using moving microphone arrays and aSLAM exploit the spatio-temporal diversity of acoustic arrays in order to resolve multiple interfering sources. The performance of such a system is closely coupled with the observer motion, i.e., acoustic scene mapping can be optimized by approaching sources affected by uncertainty. Therefore, aSLAM in combination with informative motion planning allows autonomous machines to embrace, rather than tackle, the multi-source nature of everyday acoustic environments.

APPENDIX A
SOURCE MAPPING DERIVATION

The updated PHD in (30) is derived by inserting the predicted PHD in (13) and likelihood, $p(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t, \mathbf{s}_t)$, in (26) into (20). By probability transformation of (10), the likelihood of a single DoA estimate is a wrapped Gaussian pdf [30, Chapter 3], i.e.,

$$p\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t, \mathbf{s}_t\right) = \mathcal{N}^w\left(\boldsymbol{\omega}_{t,m} \,\middle|\, g(\mathbf{s}_t),\, \mathbf{R}_{t,m}\right) = \sum_{k=-\infty}^{\infty} \mathcal{N}\left(\boldsymbol{\omega}_{t,m} \,\middle|\, \hat{g}_k(\mathbf{s}_t),\, \mathbf{R}_{t,m}\right), \quad \text{(A.1)}$$

Using (13) and (A.1), the numerator in (20) is:

$$p\left(\boldsymbol{\omega}_{t,m} \mid \mathbf{r}_t, \mathbf{s}_t\right) \lambda\left(\mathbf{s}_t \mid \mathbf{r}_t, \boldsymbol{\Omega}_{1:t-1}\right) = \sum_{j=1}^{J_{t-1}} w_{t-1}^{(j)} \sum_{k=-\infty}^{\infty} \mathcal{N}\Big(\mathbf{s}_t \,\Big|\, \mathbf{m}_{t|t-1}^{(j)},\, \boldsymbol{\Sigma}_{t|t-1}^{(j)}\Big) \mathcal{N}\big(\boldsymbol{\omega}_{t,m} \big| \hat{g}_k(\mathbf{s}_t),\, \mathbf{R}_{t,m}\big).$$

By rearranging Bayes's theorem:

$$\mathcal{N}\left(\mathbf{s}_t \,\middle|\, \mathbf{m}_{t|t-1}^{(j)},\, \boldsymbol{\Sigma}_{t|t-1}^{(j)}\right) \mathcal{N}\left(\boldsymbol{\omega}_{t,m} \,\middle|\, \hat{g}_k(\mathbf{s}_t),\, \mathbf{R}_{t,m}\right) = \mathcal{N}\left(\boldsymbol{\omega}_{t,m} \,\middle|\, \hat{g}(\mathbf{m}_{t|t-1}^{(j)}),\, \mathbf{S}_{t,m}^{(j,k)}\right) \mathcal{N}\left(\mathbf{s}_t \,\middle|\, \mathbf{m}_{t,m}^{(j,k)},\, \boldsymbol{\Sigma}_{t,m}^{(j,k)}\right), \quad \text{(A.2)}$$

with terms as defined in (31)–(32). Inserting (A.2) into (20), the detected PHD is equivalent to:

$$\lambda_d\left(\mathbf{s}_t \mid \mathbf{r}_t\right) \approx p_d \sum_{m=1}^{M_t} \sum_{k=-\infty}^{\infty} \sum_{j=1}^{J_{t-1}} w_{t,m}^{(j,k)} \mathcal{N}\Big(\mathbf{s}_t \,\Big|\, \mathbf{m}_{t,m}^{(j,k)},\, \boldsymbol{\Sigma}_{t,m}^{(j,k)}\Big),$$

where $w_{t,m}^{(j,k)}$ is defined in (33b). To capture only the relevant wrapping effects in $[\phi_{t,m} - 2\pi, \phi_{t,m} + 2\pi]$ and $[\theta_{t,m} - \pi, \theta_{t,m} + \pi]$ [38], the infinite series over k can be approximated by a finite sum over $k = -1, 0, 1$, thus leading to (30).

REFERENCES

[1] C. Cadena *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.

[2] G. Grisetti, R. Kummerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based SLAM," *IEEE Intell. Transp. Syst. Mag.*, vol. 2, no. 4, pp. 31–43, Winter 2010.

[3] S. Lowry *et al.*, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[4] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[5] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.

[6] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 178–192, Jan. 2017.

[7] V. Tourbabin and B. Rafaely, "Theoretical framework for the optimization of microphone array configuration for humanoid robot audition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1803–1814, Dec. 2014.

[8] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auton. Syst.*, vol. 55, no. 3, pp. 216–228, 2007.

[9] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. New York, NY: Springer-Verlag, 2010.

[10] P. N. Randall, C. J. Tucker, and G. A. Page, "Kinematic ranging for IRSTs," in *Proc. SPIE*, vol. 1950, 1993, pp. 96–104.

[11] C. Evers, Y. Dorfan, S. Gannot, and P. A. Naylor, "Source tracking using moving microphone arrays for robot audition," in *Proc. IEEE Intl. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 6145–6149.

[12] L. George and A. Mazel, "Humanoid robot indoor navigation based on 2D bar codes: Application to the NAO robot," in *Proc. IEEE-RAS Intl. Conf. Humanoid Robots*, Oct. 2013, pp. 329–335.

[13] C. Fischer and H. Gellersen, "Location and navigation support for emergency responders: A survey," *IEEE Pervasive Comput.*, vol. 9, no. 1, pp. 38–47, Jan. 2010.

[14] C. Evers, A. H. Moore, and P. A. Naylor, "Acoustic simultaneous localization and mapping (a-SLAM) of a moving microphone array and its surrounding speakers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Shanghai, China, Mar. 2016, pp. 6–10.

[15] C. Evers, A. H. Moore, and P. A. Naylor, "Localization of moving microphone arrays from moving sound sources for robot audition," in *Proc. Eur. Signal Proc. Conf.*, Budapest, Hungary, 2016, pp. 1008–1012.

[16] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robot. Autom. Mag.*, vol. 13, no. 2, pp. 99–110, Jun. 2006.

[17] A. Nüchter, *3D Robotic Mapping* (Springer Tracts in Advanced Robotics Series), vol. 52. Berlin-Heidelberg: Springer-Verlag, 2009.

[18] J.-S. Hu, C.-Y. Chan, C.-K. Wang, M.-T. Lee, and C.-Y. Kuo, "Simultaneous localization of a mobile robot and multiple sound sources using a microphone array," *Adv. Rbot.*, vol. 25, no. 1–2, pp. 135–152, 2011.

[19] S. Ogiso, T. Kawagishi, K. Mizutani, N. Wakatsuki, and K. Zempo, "Self-localization method for mobile robot using acoustic beacons," *ROBOMECH J.*, vol. 2, no. 1, p. 12, Sep. 2015.

[20] M. Montemerlo and S. Thrun, *FastSLAM*, ser. Springer Tracts in Advanced Robotics, vol. 27. Berlin-Heidelberg: Springer-Verlag, 2007.

[21] M. Kreković, I. Dokmanić, and M. Vetterli, "EchoSLAM: Simultaneous localization and mapping with acoustic echoes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 11–15.

[22] T. Wang, F. Peng, and B. Chen, "First order echo based room shape recovery using a single mobile device," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Proces.*, Mar. 2016, pp. 21–25.

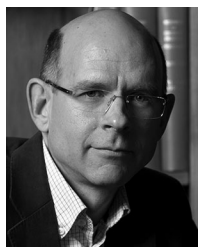
[23] I. Kelly and F. Boland, "Detecting arrivals in room impulse responses with dynamic time warping," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 7, pp. 1139–1147, Jul. 2014.

[24] 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.

[25] X.-R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. Part I: Dynamic models," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1333–1364, Oct. 2003.

[26] 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. 1, pp. 863–878, Feb. 2018.
[27] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*, vol. 1, Elementary Theory and Methods. Berlin, Germany: Springer-Verlag, 2003.
[28] R. P. S. Mahler, "Statistics 101 for multisensor, multitarget data fusion," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 19, no. 1, pp. 53–64, Jan. 2004.
[29] K. V. Mardia and P. E. Jupp, *Directional Statistics*. Hoboken, NJ: Wiley-Blackwell, 1999.
[30] S. Gannot and A. Yeredor, "The Kalman filter," in *Springer Handbook Speech Process.*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin, Germany: Springer-Verlag, 2008, ch. 8, Part B.
[31] 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.
[32] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.
[33] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *The Ann. Statist.*, vol. 38, no. 5, pp. 2916–2957, 2010.
[34] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist. Comput.*, vol. 10, no. 3, pp. 197–208, 2000.
[35] G. Casella and C. P. Robert, "Rao-Blackwellisation of sampling schemes," *Biometrika*, vol. 83, no. 1, pp. 81–94, 1996.
[36] T. Schon, F. Gustafsson, and P. J. Nordlund, "Marginalized particle filters for mixed linear/nonlinear state-space models," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2279–2289, Jul. 2005.
[37] J. Traa and P. Smaragdis, "A wrapped Kalman filter for azimuthal speaker tracking," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1257–1260, Dec. 2013.
[38] R. N. Jazar, *Theory of Applied Robotics: Kinematics, Dynamics, and Control*, 2nd ed. New York, NY: Springer-Verlag, 2010.
[39] J. Geweke, "Bayesian inference in econometric models using Monte Carlo integration," *Econometrica*, vol. 57, no. 6, pp. 1317–1339, 1989.
[40] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*. Berlin, Germany: Springer-Verlag, 2016.
[41] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.
[42] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven (TU/e), 2006, https://www.researchgate.net/publication/259991276_Room_Impulse_Response_Generator.
[43] E. Lehmann and A. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1429–1439, Aug. 2010.
[44] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proc. Nat. Conf. Artif. Intell.*, 2000, pp. 832–839.
[45] J. S. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium (LDC), Philadelphia, Corpus LDC93S1, 1993.
[46] B. Rafaely and D. Kolossa, "Speaker localization in reverberant rooms based on direct path dominance test statistics," in *Proc. IEEE Intl. Conf. Acoust., Speech Signal Process.*, 2017, pp. 6120–6124.
[47] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *J. Mach. Learn. Res.*, vol. 6, pp. 1345–1382, Sep. 2005.
[48] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud, "EM Algorithms for weighted-data clustering with application to audio-visual scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2402–2415, Dec. 2016.
[49] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer-Verlag, 2002.
[50] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length* (Information Science and Statistics Series). New York, NY: Springer-Verlag, 2005.
[51] B.-N. Vo and W.-K. Ma, "The Gaussian Mixture probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4091–4104, Nov. 2006.
[52] B. Ristic, *Particle Filters for Random Set Models*, 1st ed. New York, NY: Springer-Verlag, 2013.
[53] B. Ristic, B.-N. Vo, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3452–3457, Jul. 2011.
[54] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3447–3457, Aug. 2008.
[55] P. E. Greenwood and M. S. Nikulin, *A Guide to Chi-Squared Testing* (Wiley Series in Probability and Statistics Series). Hoboken, NJ: Wiley, 1996.

Christine Evers (M'14–SM'16) received the the B.Sc. degree in electrical engineering and computer science from the Jacobs University Bremen, Bremen, Germany, in 2005, the M.Sc. degree in signal processing and communications from the University of Edinburgh, Edinburgh, U.K., in 2006, and the Ph.D. degree from the University of Edinburgh, Edinburgh, U.K., in 2010. She is an Engineering and Physical Research Council (EPSRC) Research Fellow with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. After a position as a Research Fellow with the University of Edinburgh between 2009 and 2010, she worked as a Senior Systems Engineer with Selex ES, Edinburgh, U.K., between 2010 and 2014. In 2014, she joined the Imperial College as a Research Associate. In 2017, she was awarded a fellowship by the U.K. EPSRC to advance her research on acoustic signal processing and scene mapping for socially assistive robots. Her research focuses on statistical signal processing for audio applications, including sound source localization and tracking, acoustic simultaneous localization and mapping for robot audition, and sensor fusion. She is member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing.



Patrick A. Naylor (SM'07) received the B.Eng. degree in electronic and electrical engineering from the University of Sheffield, Sheffield, U.K., and the Ph.D. degree from the Imperial College London, London, U.K. He is a member of academic staff with the Department of Electrical and Electronic Engineering, Imperial College London.He has worked, in particular, on adaptive signal processing for speech dereverberation, blind multichannel system identification and equalization, acoustic echo control, speech quality estimation and classification, single- and multichannel speech enhancement, and speech production modelling with particular focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several fruitful links with industry in the U.K., USA, and in Europe. His research interests include speech, audio, and acoustic signal processing. He is the Past-Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and a Director of the European Association for Signal Processing. He was an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS and is currently a Senior Area Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.