Localization of Moving Microphone Arrays from Moving Sound Sources for Robot Audition

(Invited Paper)

Christine Evers, Alastair H. Moore, and Patrick A. Naylor
Department of Electrical and Electronic Engineering, Imperial College London
Exhibition Road, SW7 2AZ, London
Email: c.evers@imperial.ac.uk

***Abstract*—Acoustic Simultaneous Localization and Mapping (a-SLAM) jointly localizes the trajectory of a microphone array installed on a moving platform, whilst estimating the acoustic map of surrounding sound sources, such as human speakers. Whilst traditional approaches for SLAM in the vision and optical research literature rely on the assumption that the surrounding map features are static, in the acoustic case the positions of talkers are usually time-varying due to head rotations and body movements. This paper demonstrates that tracking of moving sources can be incorporated in a-SLAM by modelling the acoustic map as a Random Finite Set (RFS) of multiple sources and explicitly imposing models of the source dynamics. The proposed approach is verified and its performance evaluated for realistic simulated data.**

I. Introduction

The ability to roam freely within the surrounding environment is a fundamental requirement for robot technologies targeted at the provision of assistance to humans, including search-and-rescue, socially assistive, and hospitality applications. Any sensors, such as microphone arrays, installed in or along the robotic body are therefore subject to rotations in orientation and displacements in space over time.

The movement of microphone arrays can be exploited constructively to triangulate the positions of surrounding sound sources from bearing-only measurements. In order to exploit spatial diversity of the moving platform, accurate knowledge of the microphone array positions along the robot trajectory is required. However, many robots, such as the humanoid NAO by Aldebaran Robotics, are not equipped with self-localization sensors, such that the robot positions are unknown and must be estimated.

The robot trajectory can be estimated by exploiting the apparent displacement of surrounding sound sources observed from multiple positions along the robot trajectory. As both the acoustic scene map [1] as well as the robot trajectory are unknown and desired, a-SLAM is necessary to simultaneously localize the robot trajectory whilst mapping the surrounding map features, i.e., the sound sources. By exploiting spatial diversity of the sensor, a-SLAM uses instantaneous Sound Source Localization (SSL) estimates [2], [3] as measurements in order to track the three-dimensional Cartesian source positions in time. As the measurements, and hence the estimated source tracks, are relative to the microphone array center, the robot position can be estimated as the anchor point between subsequent source measurements.

Traditional SLAM approaches often used for visual or optical sensors, such as FActored Solution To Simultaneous Localization and Mapping (FastSLAM) [4], are based on the fundamental assumption that the surrounding features to be mapped are static. However, human speakers are highly non-stationary due to body movements and head rotations. Moreover, sound emitted in enclosed spaces such as rooms is subject to reverberation and background noise [5], leading to SSL errors and spurious clutter measurements.

A new approach for SLAM using acoustic sensors recently proposed in [6], [7] addresses the challenges of clutter, missing detections, and SSL errors encountered when mapping sound sources from reverberant speech signals. By formulating the multi-source states as a RFS, surviving and newborn sources, clutter, and missing detections can be explicitly modelled and exploited for estimation of the source positions using a Probability Hypothesis Density (PHD) filter [8]. Results in [6] demonstrate that the proposed a-SLAM framework successfully infers the sensor path as well as the range and positions of static sources from the bearing-only measurements.

In this paper we propose a novel extension of the a-SLAM approach in [6], [7] by incorporating dynamical models to capture the movement of sound sources typically encountered in practice. The resulting a-SLAM approach captures the movement of walking human speakers and models the time-varying positions of static speakers due to head and body rotations. Realistic simulations will investigate the performance of the proposed framework for varying numbers of static and moving sources, as well as the effect of a sound source's walking speed.

In the following, Section II introduces the system model and Section III details the proposed approach. Simulation results are presented in Section IV and conclusions are drawn in Section V.

II. Problem Formulation

The system model for a-SLAM captures positions and trajectories of the robot and surrounding sources.

The research leading to these results has received funding from the European Unions Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465.

A. Source dynamical model

The state[1] of source $n \in \mathcal{N}_t$ at time t is defined as $\mathbf{s}_{t,n} \triangleq \left[x_{t,n}, y_{t,n}, z_{t,n}, \dot{x}_{t,n}, \dot{y}_{t,n}, \dot{z}_{t,n}\right]^T$, containing the source position and velocity relative to $\mathbf{r}_t$. In order to allow for small local variations in the source position due to head movements, process noise is added to the source position, such that $\mathbf{s}_{t,n}$ is modelled as

$$\mathbf{s}_{t,n} = \boldsymbol{D}_{t,n}\,\mathbf{s}_{t-1,n} + \mathbf{n}_{t,n}, \qquad \mathbf{n}_{t,n} \sim \mathcal{N}\left(\mathbf{0},\, \boldsymbol{Q}_{t,n}\right), \tag{1}$$

where $\mathbf{n}_{t,n}$ is the process noise with covariance $\boldsymbol{Q}_{t,n}$ used to model small deviations due to head and body rotations, and $\boldsymbol{D}_{t,n}$ is the dynamical model, defined

$$\boldsymbol{D}_{t,n} \triangleq \begin{bmatrix} 1 & 0 & 0 & \Delta_T & 0 & 0 \\ 0 & 1 & 0 & 0 & \Delta_T & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta_T \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

where Δ_T is the time delay between $t-1$ and t.

By introducing moving sound sources and randomizing the positions of static sources, the source positions therefore become time-varying. This model violates the assumptions of traditional SLAM approaches [9] often used for visual and optical sensors, as approaches such as FastSLAM [4] require static map features in order to decouple estimation of the features from the robot trajectory.

To address the challenge of SLAM of moving sensor arrays and moving sources typical for acoustic sensors, we propose to utilize the a-SLAM approach in [6], [7]. In this framework, the map of sources in (1) is captured in a RFS, $\boldsymbol{S}_t \triangleq \{\mathbf{s}_{t,n}\}_{n=1}^{N_t}$, where each source position is considered as a random state variable and where the set cardinality, N_t, itself is a random variable. Using this formulation, the multi-source state model can explicitly account for source initialisation, survival between time steps, and termination, such that

$$\boldsymbol{S}_t = \left[\bigcup_{n=1}^{N_{t-1}} P(\mathbf{s}_{t-1,n})\right] \cup B_t, \tag{3}$$

where B_t is a birth process. Furthermore, $P(\mathbf{s}_{t-1,n}) = \mathbf{s}_{t,n}$ if source n is persistent between $t-1$ to t, and $P(\mathbf{s}_{t-1,n}) = \emptyset$ otherwise, where $\emptyset$ is the empty set.

B. Source measurement model

In this paper, we assume that range-bearing measurements of the source positions in (1) are available from, e.g., Direction-of-Arrival (DoA) estimation using a spherical anthropomorphic head array [3] and range estimation using a robomorphic array of microphones attached to the robot's limbs [10]. In reverberation, the range-bearing measurements are subject to localization error due to reverberation and noise. Moreover, early reflections of sources often lead to spurious clutter detections, such that the number of measurements, M_t, is often not equal to the number of sources, N_t. The measurements are therefore modelled as

$$\boldsymbol{\omega}_{t,m} = g(\mathbf{s}_{t,n}) + \mathbf{m}_{t,m}, \qquad \mathbf{m}_{t,m} \sim \mathcal{N}\left(\mathbf{0},\, \boldsymbol{R}\right), \tag{4}$$

where $\boldsymbol{\omega}_{t,m} \triangleq \left[r_{t,m}, \theta_{t,m}, \phi_{t,m}\right]^T$, for $m \in \mathcal{M}_t$, with range, $r = \sqrt{x^2+y^2+z^2}$, azimuth $\phi = \arctan(y/x)$ and inclination $\theta = \cos^{-1}\left(z/\sqrt{x^2+y^2+z^2}\right)$. The function $g(\cdot)$ denotes the Cartesian-to-spherical transformation, and $\mathbf{m}_{t,m}$ is the measurement noise with covariance $\boldsymbol{R}$.

Similar to (3), the point process of the M_t source measurements is expressed as a RFS, $\boldsymbol{\Omega}_t \triangleq \{\boldsymbol{\omega}_{t,m}\}_{m=1}^{M_t}$. Each source measurement corresponds to either the direct path between the robot and a source, a reverberant reflection, or a noise source. Multi-source measurement models should hence account for source detection, missed detection and clutter, such that

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

where C_t is the RFS of clutter measurements and the localization of sources is modelled as $D(\mathbf{s}_{t,n}) = \boldsymbol{\omega}_{t,m}$ if $\mathbf{s}_{t,n}$ is detected and $D(\mathbf{s}_{t,n}) = \emptyset$ otherwise.

C. Robot model

The robot state, $\mathbf{r}_t = \left[x_{t,r}, y_{t,r}, z_{t,r}, v_t, \gamma_t\right]^T$, at time t contains the three-dimensional Cartesian position, $(x_{t,r}, y_{t,r}, z_{t,r})$, robot speed, v_t, and orientation, γ_t. The robot position and velocity can be modelled as a linear Gaussian state space that is non-linear in the orientation. Augmenting the linear variables in the vector $\mathbf{p}_t \triangleq \left[x_{t,r}, y_{t,r}, z_{t,r}, v_t\right]^T$,

$$\mathbf{p}_t = \boldsymbol{F}_t\,\mathbf{p}_{t-1} + \mathbf{v}_{t,\mathbf{p}}, \qquad \mathbf{v}_{t,\mathbf{p}} \sim \mathcal{N}\left(\mathbf{0}_{4\times 1},\, \boldsymbol{\Sigma}_{t,\mathbf{v}}\right) \tag{6a}$$

$$\gamma_t = \gamma_{t-1} + v_{t,\gamma}, \qquad v_{t,\gamma} \sim \mathcal{N}\left(0,\, \sigma^2_{v_{t,\gamma}}\right) \tag{6b}$$

where the matrix $\boldsymbol{F}_t$ modelling the robot dynamics is general but defined in this paper as a constant velocity and constant height model, such that

$$\boldsymbol{F}_t = \begin{bmatrix} 1 & 0 & 0 & \Delta_T \sin\gamma_t \\ 0 & 1 & 0 & \Delta_T \cos\gamma_t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{7}$$

Robots typically move within an environment based on path planning algorithms. However, due to physical imperfections of the robot's motors and its environment, the executed orientation, γ_t, and velocity, v_t, typically diverge from the planned orientation, $y_{t,\gamma}$, and velocity, $y_{t,v}$. The planned path instructions, $\mathbf{y}_t \triangleq \begin{bmatrix} y_{t,\gamma} & y_{t,v} \end{bmatrix}^T$, referred to hereafter as the robot measurements, can therefore be modelled as

$$y_{t,v} = \mathbf{h}\,\mathbf{p}_t + w_{t,v}, \qquad w_{t,v} \sim \mathcal{N}\left(0,\, \sigma^2_{w_{t,v}}\right) \tag{8a}$$

$$y_{t,\gamma} = \gamma_t + w_{t,\gamma}, \qquad w_{t,\gamma} \sim \mathcal{N}\left(0,\, \sigma^2_{w_{t,\gamma}}\right) \tag{8b}$$

where $\mathbf{h} \triangleq \begin{bmatrix} 0, 0, 0, 1 \end{bmatrix}$ and where the discrepancy between the planned and executed velocity and orientation is modelled as White Gaussian Noise (WGN) noise, $w_{t,v}$ and $w_{t,\gamma}$ with variance $\sigma^2_{w_{t,v}}$ and $\sigma^2_{w_{t,\gamma}}$.

[1] The notation $k \in \mathcal{K}_t$ is used for compactness to denote $k \in \{1, \ldots, K_t\}$.

III. System Estimation

a-SLAM estimates the joint states, $\boldsymbol{X}_t \triangleq \{(\mathbf{r}_t, (N_t, \boldsymbol{S}_t))\}$, consisting of the robot states and the N_t multi-source states, from the measurements, $\boldsymbol{Z}_t = \{(\mathbf{y}_t, (M_t, \boldsymbol{\Omega}_t))\}$, containing the measurements of the robot position and the M_t source measurements, $\boldsymbol{\Omega}_t$.

In a Bayesian framework, $\boldsymbol{X}_t$ can be estimated sequentially by maximising the filtering density, $p(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t})$, at each time step t. However, in practice, the multi-object densities are combinatorially intractable. Instead of estimating the full Probability Density Function (pdf), its first-order moment, the PHD, $\lambda(\mathbf{x}_t|\boldsymbol{Z}_{1:t})$, can be estimated instead. Accordingly, the PHD of the joint state, $\boldsymbol{X}_t$, can be expressed as,

$$\lambda(\mathbf{x}_t|\boldsymbol{Z}_{1:t}) = p(\mathbf{r}_t|\mathbf{y}_{1:t})\,\lambda(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t}) \tag{9}$$

where $\lambda(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t})$ is the multi-source PHD, defined in detail in Section III-B. The density $p(\mathbf{r}_t|\mathbf{y}_{1:t})$ is the robot posterior pdf, given by [6]

$$p(\mathbf{r}_t|\mathbf{y}_{1:t}) = \frac{\mathcal{L}(\boldsymbol{\Omega}_t|\mathbf{r}_t)\,p(\mathbf{r}_t|\mathbf{y}_{1:t-1})p(\mathbf{y}_t|\mathbf{r}_t)}{\int \mathcal{L}(\boldsymbol{\Omega}_t|\mathbf{r}_t)\,p(\mathbf{r}_t|\mathbf{y}_{1:t-1})p(\mathbf{y}_t|\mathbf{r}_t)d\mathbf{r}_t} \tag{10}$$

where $p(\mathbf{r}_t|\mathbf{y}_{1:t-1})$ denotes the predicted robot pdf, and $p(\mathbf{y}_t|\mathbf{r}_t)$ is the likelihood of the robot instructions. The term $\mathcal{L}(\boldsymbol{\Omega}_t|\mathbf{r}_t)$ denotes the multi-source likelihood defined as

$$\mathcal{L}(\boldsymbol{\Omega}_t|\mathbf{r}_t) \triangleq e^{-N_{t,c}-p_d\,N_{t|t-1}} \prod_{m=1}^{M_t} \ell(\boldsymbol{\omega}_{t,m}|\mathbf{r}_t) \tag{11}$$

and where $N_{t,c}$ is the estimated number of clutter measurements, $N_{t|t-1}$ is the predicted number of sources, and $\ell(\boldsymbol{\omega}_{t,m}|\mathbf{r}_t)$ is the single-source likelihood and will be defined in Section III-B2.

A. Estimation of sensor path

The unknown variable, $\mathbf{p}_t$, in (6) is modelled as a linear Gaussian state space, but is non-linearly dependent on the orientation, γ_t. A Rao-Blackwellized particle filter [11] is therefore used to propagate $p(\mathbf{r}_t|\mathbf{y}_{1:t})$ in time, such that

$$p(\mathbf{r}_t|\mathbf{y}_{1:t-1}) = p(\mathbf{p}_t|\mathbf{y}_{1:t-1,v})\,p(\gamma_t|\mathbf{y}_{1:t-1,\gamma}). \tag{12}$$

Assuming that the orientation, γ_{t-1} at $t-1$, is described by a cloud of I_{t-1} particles, $\hat{\gamma}_{t-1}^{(i)}$, the orientation at t can be obtained by drawing P importance samples for each particle from $t-1$ from a proposal distribution, $\pi(\gamma_t|\gamma_{t-1}^{(i)})$, such that

$$\hat{\gamma}_t^{(i,p)} \sim \pi(\gamma_t|\gamma_{t-1}^{(i)}) \tag{13}$$

for all $p \in \mathcal{P}$ and $i \in \mathcal{I}_{t-1}$. The robot pose, $\mathbf{p}_t$, is described by a linear Gaussian state-space conditional on the orientation. Therefore, for each robot particle in (13), the Kalman Filter (KF) equations [12] are evaluated to propagate the robot pose in time. The pdf of the robot state, $\mathbf{r}_t$ in (12) can therefore be expressed as

$$p(\mathbf{r}_t|\mathbf{y}_{1:t}) = \sum_{i=1}^{I_{t-1}} \sum_{p=1}^{P} \alpha_t^{(i,p)}\,\delta_{\hat{\gamma}_t^{(i,p)}}(\gamma_t)\,\mathcal{N}(\mathbf{p}_t|\boldsymbol{\psi}_t^{(i,p)}, \boldsymbol{\Psi}_t^{(i,p)}) \tag{14}$$

where $\delta_{\hat{\gamma}_t^{(i,p)}}(\gamma_t)$ denotes the Dirac-Delta function of γ_t centered about $\hat{\gamma}_t^{(i,p)}$, the updated sensor position mean, $\boldsymbol{\psi}_t^{(i,p)}$ and covariance, $\boldsymbol{\Psi}_t^{(i,p)}$, are given by the KF update equations [12], and the weights, $\alpha_t^{(i,p)}$, are defined as

$$\alpha_t^{(i,p)} \triangleq \frac{\hat{\alpha}_t^{(i,p)}}{\sum_{j=1}^{I_{t-1}} \sum_{q=1}^{P} \hat{\alpha}_t^{(j,q)}} \tag{15}$$

where the unnormalised weights, $\hat{\alpha}_t^{(j,q)}$, are

$$\begin{aligned}\hat{\alpha}_t^{(i,p)} \triangleq\ & \alpha_{t|t-1}^{(i,p)}\,\mathcal{L}(\boldsymbol{\Omega}_t|\mathbf{r}_t^{(i,p)})\,\mathcal{N}\left(y_{t,v}|\mathbf{h}\,\boldsymbol{\psi}_{t|t-1}^{(i,p)}, \sigma_{t,s}^2\right) \\ & \times \mathcal{N}\left(y_{t,\gamma}|\gamma_t^{(i,p)}, \sigma_{t,w_\gamma}^2\right).\end{aligned} \tag{16}$$

It is important to note that the weight of each Rao-Blackwellised robot state particle, $\mathbf{r}_t^{(i,p)}$, accounts for the multi-source likelihood, $\mathcal{L}(\boldsymbol{\Omega}_t|\mathbf{r}_t^{(i,p)})$. As the set of source measurements, $\boldsymbol{\Omega}_t$, is relative to the robot state, unlikely particles, $\mathbf{r}_t^{(i,p)}$, result in rotations and displacements of the measurements. The discrepancy between the source estimates $\boldsymbol{S}_{t-1}$ and their measurements, $\boldsymbol{\Omega}_t$, at t is therefore increased, such that the likelihood, $\mathcal{L}(\boldsymbol{\Omega}_t|\mathbf{r}_t^{(i,p)})$ decreases. The likelihood is therefore used as a weighting of the robot particles in (16) in order to triangulate the robot position using the acoustic map of estimated sources.

B. Estimation of source trajectories

The single-source states in (1) are modelled as a linear Gaussian state space. Therefore, a Gaussian Mixture PHD (GM-PHD) [13] is used to estimate the multi-source states, $\boldsymbol{S}_t$. For each sensor estimate, $\mathbf{r}_t^{(i,p)}$, one GM-PHD filter is propagated to estimate the source trajectories. For readability the superscript (i,p) is dropped in the following.

1) Source Prediction: The predicted PHD is given by a Gaussian sum of $J_{t|t-1} = J_{t-1} + L$ components, with Gaussian Mixture (GM) weights, $w_{t|t-1}^{(j)}$, mean, $\mathbf{m}_{t|t-1}^{(j)}$, and covariance, $\boldsymbol{\Sigma}_{t|t-1}^{(j)}$, such that

$$\lambda(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t-1}) = \sum_{j=1}^{J_{t|t-1}} w_{t|t-1}^{(j)}\,\mathcal{N}\left(\mathbf{s}_t\,|\,\mathbf{m}_{t|t-1}^{(j)},\,\boldsymbol{\Sigma}_{t|t-1}^{(j)}\right). \tag{17}$$

The GM components can be grouped into components due to newborn sources, $\lambda_b(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_t)$, and persistent components due to sources surviving from $t-1$ to t, $\lambda_s(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t})$, such that

$$\lambda(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t-1}) = \lambda_b(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_t) + \lambda_s(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t}). \tag{18}$$

The predicted PHD of surviving sources is given by the GM,

$$\lambda_s(\mathbf{s}_t|\mathbf{r}_t, \boldsymbol{\Omega}_{1:t-1}) = \sum_{j=1}^{J_{t-1}} w_{s,t|t-1}^{(j)}\,\mathcal{N}\left(\mathbf{s}_t\,|\,\mathbf{m}_{s,t|t-1}^{(j)},\,\boldsymbol{\Sigma}_{s,t|t-1}^{(j)}\right), \tag{19}$$

where the predicted GM weight, $w_{s,t|t-1}^{(j)} = p_s\, w_{t-1}$, and the mean, $\mathbf{m}_{s,t|t-1}^{(j)}$, and covariance, $\mathbf{\Sigma}_{s,t|t-1}^{(j)}$, terms are given by the KF prediction [13].

Similar to (19), the PHD of newborn sources is given

$$\lambda_b(\mathbf{s}_t|\mathbf{r}_t, \mathbf{\Omega}_t) = \sum_{\ell=1}^{L} w_{b,t}^{(\ell)}\, \mathcal{N}\left(\mathbf{s}_t \,|\, \mathbf{m}_{b,t}^{(\ell)},\, \mathbf{\Sigma}_{b,t}\right). \tag{20}$$

where $\mathbf{m}_{b,t}^{(\ell)}$ are the GM mean terms, the GM weights are $w_{b,t}^{(\ell)} = \frac{N_b}{L}$ for N_b expected source births per time step, and the covariance, $\mathbf{\Sigma}_{b,t}$, is constant and known *a priori*. The birth process in this paper is measurement-driven, such that the GM mean terms, $\mathbf{m}_{b,t}^{(\ell)}$, are constructed from the measurements, $\boldsymbol{\omega}_{t,m}$, by drawing $M_t\, L$ birth GM components from

$$\mathbf{m}_{b,t}^{(\ell)} \sim \mathcal{N}\left(\boldsymbol{\omega}_{t,m},\, \boldsymbol{R}\right). \tag{21}$$

for all $m \in \mathcal{M}_t$, where $\ell = (m-1)L+1, \ldots, mL$ and $\boldsymbol{R}$ is the measurement covariance in (4).

2) Source update: The updated source PHD, $\lambda(\mathbf{s}_t|\mathbf{r}_t, \mathbf{\Omega}_{1:t})$, corrects the predicted components using the measurements whilst accounting for the probability of source detection, p_d. Since both the prediction of surviving sources and the birth PHD are expressed by GMs (19) and (20), the updated source PHD is equivalent to

$$\begin{aligned}\lambda(\mathbf{s}_t|\mathbf{r}_t, \mathbf{\Omega}_{1:t}) &= (1-p_d)\, \lambda_s(\mathbf{s}_t|\mathbf{r}_t, \mathbf{\Omega}_{1:t-1}) \\ &+ p_d \sum_{m=1}^{M_t} \sum_{j=1}^{J_{t|t-1}} w_t^{(j,m)}\, \mathcal{N}\left(\mathbf{s}_t \,|\, \mathbf{m}_t^{(j,m)},\, \mathbf{\Sigma}_t^{(j,m)}\right)\end{aligned} \tag{22}$$

where the updated mean, $\mathbf{m}_t^{(j,m)}$, and covariance, $\mathbf{\Sigma}_t^{(j,m)}$, of the GM components are given by the Extended Kalman Filter (EKF) update equations [13], $\lambda_s(\mathbf{s}_t|\mathbf{r}_t, \mathbf{\Omega}_{1:t-1})$ is given in (19) and the updated weights, $w_t^{(j,m)}$, are

$$w_t^{(j,m)} \triangleq w_{t|t-1}^{(j)}\, \frac{q(\boldsymbol{\omega}_{t,m}|\mathbf{r}_t)}{\ell(\boldsymbol{\omega}_{t,m}|\mathbf{r}_t)}. \tag{23}$$

with $q(\boldsymbol{\omega}_{t,m}|\mathbf{r}_t) \triangleq \mathcal{N}\left(\boldsymbol{\omega}_{t,m} \,|\, g\left(\mathbf{m}_{t|t-1}^{(j)}\right),\, \boldsymbol{S}_t^{(j)}\right)$. The single-source likelihood, $\ell(\boldsymbol{\omega}_{t,m}|\mathbf{r}_t)$, is given by

$$\ell(\boldsymbol{\omega}_{t,m}|\mathbf{r}_t) = \kappa(\boldsymbol{\omega}_{t,m}) + p_d \sum_{j=1}^{J_{t|t-1}} w_{t|t-1}^{(j)}\, q(\boldsymbol{\omega}_{t,m}|\mathbf{r}_t) \tag{24}$$

where $\kappa(\boldsymbol{\omega}_{t,m})$ is the PHD of clutter.

The source PHD can also be used to obtain an estimate of the number of sources, $\hat{N}_t$, where [13]

$$\hat{N}_t = \sum_{j=1}^{J_{t|t-1}} w_{t|t-1}^{(j)} + \sum_{m=1}^{M_t} \sum_{\ell=1}^{J_{t|t-1}} w_t^{(\ell,m)}. \tag{25}$$

C. Extraction of point estimates

A point estimate of the robot state at each t is extracted as the weighted average of the particles, $\mathbf{r}_t^{(i)}$ and their importance weights, $\alpha_t^{(i)}$ for all $i \in \mathcal{I}_t$, where $I_t = I_{t-1}\, P$. Source point estimates can be extracted as described in, e.g., [13].

IV. Results

To test the performance of the approach for walking sound sources the algorithm was tested in two experiments for data generated from the following simulation for a room of dimension $6 \times 6 \times 2.5$ m. The robot path is generated from (6) for an initial orientation of $\gamma_0 = 0$ deg and $\mathbf{p}_0 = \left[2\text{ m}, 2\text{m}, 1.5\text{m}, 0.5\text{m/s}\right]$ respectively, for process noise variance $\mathbf{\Sigma}_{t,\mathbf{v}} = 10^{-9} \times \boldsymbol{I}_4$ and with $\sigma_{v_{t,\gamma}} = 45$ deg on the orientation. If the robot position lies within 1 m of any wall at any $t = k$, the orientation is forced to $\gamma_k = \gamma_{k-1}\, \sigma_{v_{t,\gamma}}$ in order to reflect the position back into the room. With $\sigma_{v_{t,\gamma}} = 45$ deg, the robot can therefore perform a three-point turn. In order to initialise the sources, the room region is divided into four quadrants of equal area with origin at $(3, 3, 2.5)$ m. The $N_t = 3$ sources are assigned to quadrants by randomly choosing a unique quadrant per source. The source trajectories are generated from (1) with initial orientation of 0 rad. Head and body rotations of static talkers are modelled using process noise with standard deviation in position of $0.1\boldsymbol{I}_3$ m and 0 m/s in velocity. Walking sources are modelled for different velocities as detailed below. If a walking source reaches within 1 m of a wall at any t, its orientation is rotated by π rad for the remaining trajectory. Each source is initialized to a uniformly selected position within the assigned quadrant.

Two experiments were conducted: Experiment 1 investigates the influence of the number of sources. Experiment 2 investigates the effect of increasing source velocity. The number of robot particles is limited to $I_t = 50$ after each recursion with $P = 10$. The number of source components is limited to $J_t = 450$. Pruning as described in [13] was applied to the source GM-PHD, whilst systematic resampling was used for the robot particles [14].

A. Experiment 1: Number of walking sources

The scenario is evaluated for four different cases: 1) 3 static sources, 2) 3 walking sources, 3) 1 walking and 2 static sources, as well as 4) 2 walking and 1 static source.. Each case is evaluated for 5 independent Monte Carlo runs. The

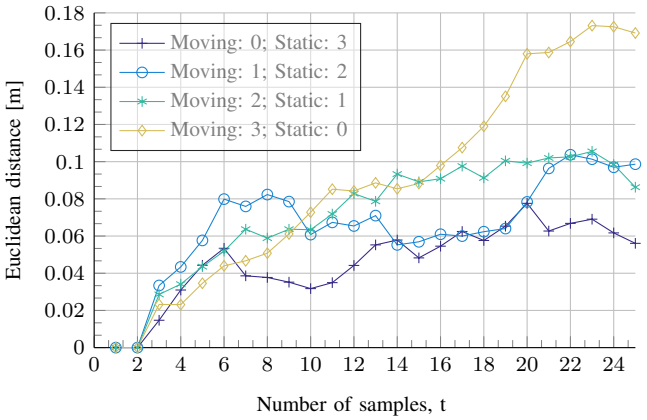

Fig. 1: Robot localization performance for increasing numbers of walking talkers.

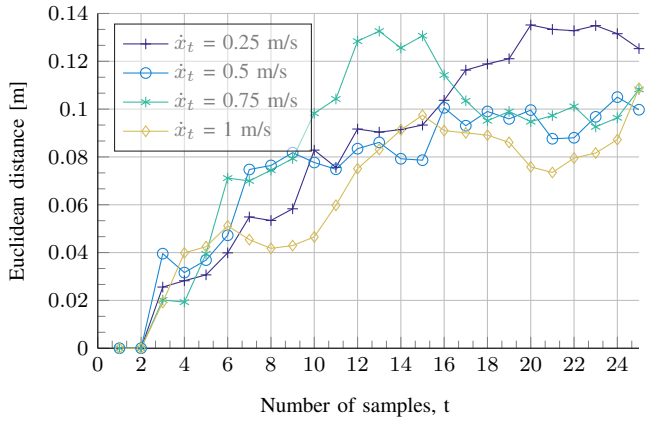

Fig. 2: Robot localization performance for increasing source velocity.

source speed is 0.25 m/s, such that walking sources cover a distance of 1 m during each experiment. The performance of the proposed a-SLAM approach for robot localization is evaluated using the Euclidean distance between the true and estimated robot trajectories shown in Fig. 1.

The results show that the a-SLAM approach for a single moving source converges to a localization accuracy of between $6 - 10$ cm after $t = 6$. The variation in this value is due to the dependency of a-SLAM on informative robot paths [7]. Smoother curves could hence be obtained by averaging over a large number of Monte Carlo runs to counteract this path-dependency. The results in Fig. 1 indicate the general performance trend of the proposed approach.

When considering 3 static sources, an average Euclidean distance of 7 cm is achieved, i.e., an improvement of up to 4 cm. The modest decrease of up to 3 cm in localization accuracy of the robot trajectory demonstrate that two sources are sufficient to triangulate the robot from the acoustic map.

When considering 2 moving speakers and a single static source, the robot path quickly converges to a Euclidean distance of 10 cm between the true and estimated robot position (see Fig. 1). For 3 moving sources, the maximum Euclidean distance diverges to 18 cm.

The results of Experiment 1 therefore demonstrate that a-SLAM accuracy within 5 cm Euclidean distance can be achieved if at least two sources are static. The static sources act as reference points – or anchors – for the a-SLAM algorithm. The results also indicate that a-SLAM is robust against additional moving sources.

B. Experiment 2: Source velocity

To further investigate the robustness of the approach for moving sources, Experiment 2 investigates the effect of the source velocity. The scenario is evaluated for 2 static sources and 1 walking source. The velocity of the walking source is increased from 0.25 m/s to 1 m/s. The proposed approach reaches a steady state after approximately $t = 16$. The Euclidean distance between the true and estimated robot trajectories are plotted in Fig. 2. The figure indicates that robot localization performance can in fact be improved by increasing the velocity of the moving source. The large displacement of fast sources allows for the easier detection of the source's motion. Therefore faster convergence of the source model can be achieved, leading to an overall improvement in a-SLAM performance.

V. Conclusions

This paper proposed an approach to a-SLAM for moving sound sources from moving microphone arrays. Models were proposed to capture the natural head and body rotations of static human talkers as well the motion of walking talkers. The models were incorporated within a framework for a-SLAM based on a RFS formulation. Simulation results demonstrated that two out of three static sources are sufficient to accurately localize the robot trajectory.

References

[1] C. Evers, J. Sheaffer, A. H. Moore, B. Rafaely, and P. A. Naylor, "Bearing-only acoustic tracking of moving speakers for robot audition," in *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*, Singapore, Jul. 2015.

[2] 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.

[3] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Nice, France, Jul. 2014.

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

[5] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer, 2010.

[6] 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 Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016.

[7] ——, "Informative path planning for acoustic simultaneous localization of microphone arrays and mapping of surrounding sound sources (a-SLAM)," in *DAGA*, Aachen, Germany, mar 2016.

[8] 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.

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

[10] H. Barfuss and W. Kellermann, "An adaptive microphone array topology for target signal extraction with humanoid robots," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, September 2014.

[11] A. Doucet, N. d. Freitas, K. P. Murphy, and S. J. Russell, "Rao-Blackwellised particle filtering for dynamic bayesian networks," in *Proc. Conf. on Uncertainty in Artificial Intelligence*, San Francisco, CA, 2000, pp. 176–183.

[12] S. Gannot and A. Yeredor, "The Kalman filter," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer-Verlag, 2008, ch. 8, part B.

[13] 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.

[14] 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.