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.
A. Source dynamical model

The state\(^1\) of source \(n \in \mathcal{N}_t\) at time \(t\) is defined as \(s_{t,n} \triangleq [x_{t,n}, \dot{x}_{t,n}, z_{t,n}, \dot{z}_{t,n}, \gamma_{t,n}, \dot{\gamma}_{t,n}]^T\), containing the source position and velocity relative to \(r_t\). 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 \(s_{t,n}\) is modelled as

\[
s_{t,n} = D_{t,n} s_{t-1,n} + n_{t,n}, \quad n_{t,n} \sim \mathcal{N}(0, Q_{t,n}),
\]

where \(n_{t,n}\) is the process noise with covariance \(Q_{t,n}\) used to model small deviations due to head and body rotations, and \(D_{t,n}\) is the dynamical model, defined

\[
D_{t,n} = \begin{bmatrix}
1 & 0 & 0 & \Delta_T & 0 & 0 \\
0 & 1 & 0 & 0 & \Delta_T & 0 \\
0 & 0 & 1 & 0 & 0 & \Delta_T \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
\end{bmatrix},
\]

where \(\Delta_T\) 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 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

\[
\Omega_t = \bigcup_{n=1}^{N_t} D(s_{t,n}) \cup C_t,
\]

where \(C_t\) is the RFS of clutter measurements and the localization of sources is modelled as \(D(s_{t,n}) = \omega_{t,m}\) if \(s_{t,n}\) is detected and \(D(s_{t,n}) = \emptyset\) otherwise.

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

\[
\omega_{t,m} = g(s_{t,n}) + m_{t,m}, \quad m_{t,m} \sim \mathcal{N}(0, R),
\]

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

Similar to (3), the point process of the \(M_t\) source measurements is expressed as a RFS, \(\Omega_t \triangleq \{\omega_{t,m}\}_{m=1}^{M_t}\). 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

\[
\Omega_t = \bigcup_{n=1}^{N_t} D(s_{t,n}) \cup C_t,
\]

where \(C_t\) is the RFS of clutter measurements and the localization of sources is modelled as \(D(s_{t,n}) = \omega_{t,m}\) if \(s_{t,n}\) is detected and \(D(s_{t,n}) = \emptyset\) otherwise.

C. Robot model

The robot state, \(r_t = [x_{t,r}, y_{t,r}, z_{t,r}, v_{t,r}, \gamma_t]^T\), at time \(t\) contains the three-dimensional Cartesian position, \((x_{t,r}, y_{t,r}, z_{t,r})\), robot speed, \(v_{t,r}\), and orientation, \(\gamma_t\). 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 \(p_t \triangleq [x_{t,r}, y_{t,r}, z_{t,r}, v_{t,r}, \gamma_t]^T\),

\[
p_t = F_t p_{t-1} + v_{t,p}, \quad v_{t,p} \sim \mathcal{N}(0_{4 \times 1}, \Sigma_{t,p})
\]

\[
\gamma_t = \gamma_{t-1} + v_{t,\gamma}, \quad v_{t,\gamma} \sim \mathcal{N}(0, \sigma_{v_{t,\gamma}}^2)
\]

where the matrix \(F_t\) modelling the robot dynamics is general but defined in this paper as a constant velocity and constant height model, such that

\[
F_t = \begin{bmatrix}
1 & 0 & 0 & \Delta_T \sin \gamma_t \\
0 & 1 & 0 & \Delta_T \cos \gamma_t \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}.
\]

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, \(\gamma_t\), and velocity, \(v_{t,r}\), typically diverge from the planned orientation, \(y_{t,v}\) and velocity, \(v_{t,v}\). The planned path instructions, \(y_t \triangleq [y_{t,v}, y_{t,\gamma}]^T\), referred to hereafter as the robot measurements, can therefore be modelled as

\[
y_{t,v} = h p_t + w_{t,v}, \quad w_{t,v} \sim \mathcal{N}(0, \sigma_{w_{t,v}}^2)
\]

\[
y_{t,\gamma} = \gamma_t + w_{t,\gamma}, \quad w_{t,\gamma} \sim \mathcal{N}(0, \sigma_{w_{t,\gamma}}^2)
\]

where \(h \triangleq [0, 0, 0, 1]\) and where the discrepancy between the planned and executed orientation and orientation is modelled as White Gaussian Noise (WGN) noise, \(w_{t,v}\) and \(w_{t,\gamma}\) with variance \(\sigma_{w_{t,v}}^2\) and \(\sigma_{w_{t,\gamma}}^2\).
III. System Estimation

a-SLAM estimates the joint states, $X_t \triangleq \{(r_i, (N_i, S_i))\}$, consisting of the robot states and the $N_i$ multi-source states, from the measurements, $Z_t = \{(y_t, (M_t, \Omega_t))\}$, containing the measurements of the robot position and the $M_t$ source measurements, $\Omega_t$.

In a Bayesian framework, $X_t$ can be estimated sequentially by maximising the filtering density, $p(X_t | Z_{1:t})$, 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 of the joint state, $X_t$, can be expressed as,

$$\lambda(x_t | Z_{1:t}) = p(r_t | y_{1:t}) \lambda(s_t | r_t, \Omega_{1:t})$$

where $\lambda(s_t | r_t, \Omega_{1:t})$ is the multi-source PHD, defined in detail in Section III-B. The density $p(r_t | y_{1:t})$ is the pdf posterior pdf, given by [6]

$$p(r_t | y_{1:t}) = \frac{\mathcal{L}(\Omega_t | r_t) p(r_t | y_{1:t-1}) p(y_t | r_t)}{\int \mathcal{L}(\Omega_t | r_t) p(r_t | y_{1:t-1}) p(y_t | r_t) dr_t}$$

where $p(r_t | y_{1:t-1})$ denotes the predicted robot pdf, and $p(y_t | r_t)$ is the likelihood of the robot instructions. The term $\mathcal{L}(\Omega_t | r_t)$ is the multi-source likelihood defined as

$$\mathcal{L}(\Omega_t | r_t) \triangleq e^{-N_t, \sigma_p N_{t|t-1}} \prod_{m=1}^{M_t} \ell(\omega_{t,m} | r_t)$$

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

A. Estimation of sensor path

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

$$p(r_t | y_{1:t-1}) = p(p_t | y_{1:t-1}, \gamma) p(\gamma_t | y_{1:t-1}, \gamma).$$

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

$$\gamma_t^{(i,p)} \sim \pi(\gamma_t | \gamma_{t-1}^{(i)})$$

for all $p \in \mathcal{P}$ and $i \in I_{t-1}$. The robot pose, $p_t$, 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, $r_t$ in (12) can therefore be expressed as

$$p(r_t | y_{1:t}) = \sum_{i=1}^{J_{t-1}} \sum_{p=1}^{P} \alpha_t^{(i,p)} \delta_{\gamma_t^{(i,p)}}(\gamma_t) \mathcal{N}(p_t | \psi_t^{(i,p)}, \Psi_t^{(i,p)})$$

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

$$\alpha_t^{(i,p)} \triangleq \frac{\hat{\alpha}_t^{(i,p)} \mathcal{I}_t^{(i,p)}}{\sum_{j=1}^{J_{t-1}} \sum_{q=1}^{Q} \hat{\alpha}_t^{(j,q)}}$$

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

$$\hat{\alpha}_t^{(i,p)} \triangleq \alpha_t^{(i,p)} \mathcal{L}(\Omega_t | r_t^{(i,p)}) \mathcal{N}(y_t | h(\psi_t^{(i,p)}, \sigma_t^2, \omega_{t,\gamma})$$

It is important to note that the weight of each Rao-Blackwellised robot state particle, $r_t^{(i,p)}$, accounts for the multi-source likelihood, $\mathcal{L}(\Omega_t | r_t^{(i,p)})$. As the set of source measurements, $\Omega_t$, is relative to the robot state, unlikely particles, $r_t^{(i,p)}$, result in rotations and displacements of the measurements. The discrepancy between the source estimates $S_{t-1}$ and their measurements, $\Omega_t$, at $t$ is therefore increased, such that the likelihood, $\mathcal{L}(\Omega_t | r_t^{(i,p)})$, 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, $S_t$. For each sensor estimate, $r_t^{(i,p)}$, 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, $m_{t|t-1}^{(j)}$, and covariance, $\Sigma_{t|t-1}^{(j)}$, such that

$$\lambda(s_t | r_t, \Omega_{t|t-1}) = \sum_{j=1}^{J_{t|t-1}} w_{t|t-1}^{(j)} \mathcal{N}(s_t | m_{t|t-1}^{(j)}, \Sigma_{t|t-1}^{(j)})$$

The GM components can be grouped into components due to newborn sources, $\lambda_b(s_t | r_t, \Omega_t)$, and persistent components due to sources surviving from $t-1$ to $t$, $\lambda_s(s_t | r_t, \Omega_{t|t-1})$, such that

$$\lambda(s_t | r_t, \Omega_{t|t-1}) = \lambda_b(s_t | r_t, \Omega_t) + \lambda_s(s_t | r_t, \Omega_{t|t-1})$$

The predicted PHD of surviving sources is given by the GM, $\lambda_s(s_t | r_t, \Omega_{t|t-1}) = \sum_{j=1}^{J_{t-1}} w_{s,t|t-1}^{(j)} \mathcal{N}(s_t | m_{s,t|t-1}^{(j)}, \Sigma_{s,t|t-1}^{(j)})$.
where the predicted GM weight, $w_{s,t-1}^{(j)} = p_s w_{t-1}$, and the mean, $\mathbf{m}_{s,t-1}^{(j)}$, and covariance, $\Sigma_{s,t-1}^{(j)}$, terms are given by the KF prediction [13].

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

$$\lambda_b(s_t | r_t; \Omega_t) = \sum_{l=1}^{L} w_{b,t}^{(l)} N\left(s_t | \mathbf{m}_{b,t}^{(l)}, \Sigma_{b,t}^{(l)} \right).$$

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

$$\mathbf{m}_{b,t}^{(l)} \sim N(\omega_{t,m}, \mathbf{R}).$$

for all $m \in M_t$, where $\ell = (m-1) L + 1, \ldots, m L$ and $\mathbf{R}$ is the measurement covariance.

2) Source update: The updated source PHD, $\lambda(s_t | r_t; \Omega_{1:t})$, 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

$$\lambda(s_t | r_t, \Omega_{1:t}) = (1 - p_d) \lambda_s(s_t | r_t, \Omega_{1:t-1}) + p_d \sum_{m=1}^{M_t} \sum_{j=1}^{J_{t-1}} w_{t,m}^{(j,m)} N\left(s_t | \mathbf{m}_{t}^{(j,m)}, \Sigma_{t}^{(j,m)} \right)$$

(22)

where the updated mean, $\mathbf{m}_{t}^{(j,m)}$, and covariance, $\Sigma_{t}^{(j,m)}$, of the GM components are given by the Extended Kalman Filter (EKF) update equations [13], $\lambda_s(s_t | r_t, \Omega_{1:t-1})$ is given in (19) and the updated weights, $w_{t,m}^{(j,m)}$, are

$$w_{t,m}^{(j,m)} = w_{t-1,m}^{(j,m)} g(\omega_{t,m} | \mathbf{r}_t).$$

(23)

with $g(\omega_{t,m} | \mathbf{r}_t) = N(\omega_{t,m} | \mathbf{m}_{t}^{(j,m)}, \Sigma_{t}^{(j,m)})$. The single-source likelihood, $\ell(\omega_{t,m} | r_t)$, is given by

$$\ell(\omega_{t,m} | r_t) = \kappa(\omega_{t,m}) + p_d \sum_{j=1}^{J_{t-1}} w_{t-1,m}^{(j,m)} g(\omega_{t,m} | \mathbf{r}_t)$$

(24)

where $\kappa(\omega_{t,m})$ 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]

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


---

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