An Audio-visual Solution to Sound Source Localization and Tracking with Applications to HRI

Brendan M. Emery†, Maani Ghaffari Jadidi†, Keisuke Nakamura‡ and Jaime Valls Miro†
† Centre for Autonomous Systems, University of Technology Sydney, Australia
brendan.emery@student.uts.edu.au,{maani.ghaffarijadidi, jaime.vallsmiro}@uts.edu.au
‡ Honda Research Institute Japan Co., Ltd., Japan
keisuke@jp.honda-ri.com

Abstract

Robot audition is an emerging and growing branch in the robotic community and is necessary for a natural Human-Robot Interaction (HRI). In this paper, we propose a framework that integrates advances from Simultaneous Localization And Mapping (SLAM), bearing-only target tracking, and robot audition techniques into a unified system for sound source identification, localization, and tracking. In indoors, acoustic observations are often highly noisy and corrupted due to reverberations, the robot egomotion and background noise, and the possible discontinuous nature of them. Therefore, in everyday interaction scenarios, the system requires accommodating for outliers, robust data association, and appropriate management of the landmarks, i.e. sound sources. We solve the robot self-localization and environment representation problems using an RGB-D SLAM algorithm, and sound source localization and tracking using recursive Bayesian estimation in the form of the extended Kalman filter with unknown data associations and an unknown number of landmarks. The experimental results show that the proposed system performs well in the medium-sized cluttered indoor environment.

1 Introduction

Natural Human-Robot Interaction (HRI) necessitates a robot to be able to communicate with humans using its multitude of senses. Among these senses, there has been an oversight in exploiting audioception in robotics in comparison to visual perception. Robot Audition has been a growing field in the past decade. However, the efforts have mainly been focused on the development of audition systems for scene understanding [Nakadai *et al.*, 2000; Valin *et al.*, 2003; Yamamoto *et al.*, 2006].

Figure 1: An example of the HRI scenario where different participants speak simultaneously and the robot is required to be responsive to the correct speaker.

A robot audition system provides measurements such as bearings by processing received sound signals. The problem of estimating a non-stationary sound source position from noisy measurements (bearings) is sound source localization and tracking (SSLT). While this problem at its core is bearing-only target tracking [Bar-Shalom, 1987], the highly noisy nature of acoustic signals, as well as reverberant indoor environments, make finding a reliable solution to this problem challenging. In this paper, we propose a system that can detect, localize, and track sound sources in indoor environments and is suitable for indoor HRI applications.

Motivation

The primary motivation of this work is to construct a system for SSLT that can deal with spurious measurements in a systematic way. Even though solving the problem of bearing only target localization and tracking is studied considerably, such systems cannot be directly applied to the case of working with acoustic signals. This challenge is due to the influence of structural forms of an indoor environment on acoustic signal propagation. Therefore, in the absence of prior knowledge of the environment, a reliable solution to the SSLT is non-trivial.

Contributions

The main contributions of this paper are as follows.

- We integrate techniques from SLAM, target tracking, and robot audition into a unified system to solve the problem of SSLT.
- We evaluate the proposed system in two scenarios and present experimental results in environments populated with stationary and moving sound sources.

Notation

Throughout this paper, matrices are capitalized in bold, such as in $\boldsymbol{P}$, and vectors are in lower case bold type, such as in $\boldsymbol{p}$. Vectors are column-wise and $1\colon k$ means whole numbers from 1 to k. For any quantity at a discrete time-step k such as in $\boldsymbol{x}(k)$, its predicted value is denoted by $\boldsymbol{x}^{-}(k)$. The $(.)^T$ and $(.)^*$ operators denote the transpose and conjugate transpose, respectively, such as in $\boldsymbol{P}^T$ and $\boldsymbol{P}^*$.

Outline

We review the relevant works related to this paper in the next section. In Section 3, the robot world modeling including solving SLAM and acoustic signals processing is explained. In Section 4, we present the proposed SSLT algorithm. The experimental results are presented in Section 5 and Section 6 concludes the paper.

2 Related Work

Application of audio signals in robotic navigation cannot be compared with the popularity of range-finders and visual sensors. However, there have been efforts to incorporate audio measurements for localization [Valin *et al.*, 2004; Hu *et al.*, 2011], mapping [Sasaki *et al.*, 2009; Kagami *et al.*, 2009; Even *et al.*, 2014], or navigation purposes [Huang *et al.*, 1999; Wang *et al.*, 2004; Martinson and Fransen, 2011]. The reason for this relative lack of interest arises from the nature of these types of observations, since they are not necessary for an autonomous mobile robot in its common, basic form. However, for an intelligent robot to be able to interact with humans, a robust audition system is inevitable. Audio features are important due to their complementary role with other sensor modalities.

Robot audition involves recognition and analysis of multiple coexisting sound sources. There are significant challenges in real-world environments such as tracking moving sound sources, reverberation or background noises which make it non-trivial for an autonomous robot to interact with a surrounding environment. In [Nakadai *et al.*, 2000], an active audition system for humanoid robots is proposed. The method improves localization through aligning microphones orthogonally to a sound source and capturing the possible sound sources using vision. In [Nakadai *et al.*, 2010], separation of sound sources and speech recognition for moving objects are studied. The process is defined in four successive stages: Sound Source Localization (SSL) [Nakamura *et al.*, 2011], Sound Source Tracking (SST), Sound Source Separation (SSS) [Nakajima *et al.*, 2010], and Automatic Speech Recognition (ASR) modules. The most important known auditory uncertainties such as reverberation or unknown number of sources are studied in [Otsuka *et al.*, 2014] through using a hierarchical Dirichlet process [Teh *et al.*, 2006] to avoid the over-fitting problem and determine the number of sources globally. In a more robotic navigation related context in [Thrun, 2005], by neglecting the reverberation problem, assuming sound source locations are given and microphones are fully synchronized, the localization problem of a set of microphones is solved. In [Nakamura *et al.*, 2013], they developed a super-resolution robot audition framework capable of SSL, SSS and ASR. In the super-resolution SSL and SSS, the resolution surpasses the original resolution of the pre-measured Transfer Functions (TF); this is achieved through TF interpolation based on the integration of Frequency and Time Domain Linear Interpolation (FTDLI). The technique is based on the adjacent-points method [Watanabe *et al.*, 2003; Matsumoto *et al.*, 2003] to make it suitable for real-time processing. Note that the SSL problem in robot audition refers to the computation of measurements, e.g. bearings, using signal processing and optimization techniques. In this paper, without *a priori* knowledge about sound sources and the environment, we use these relative measurements to localize and track sound sources in the environment's global coordinates.

3 The Robot World Modeling

The robot requires building a representation of the environment for tasks such as navigation and HRI. We employ an RGB-D camera and a synchronized microphone array for online construction of such a representation. In the following, we explain each part.

3.1 Simultaneous Localization and Mapping

Probabilistic SLAM is formulated as a *Bayesian estimation* problem in which the goal is to estimate the robot pose or trajectory and the feature map of the environment using noisy observations and control inputs. In graph SLAM [Dellaert and Kaess, 2006; Thrun and Montemerlo, 2006], the robot trajectory is estimated through the observation of relative constraints between robot poses. The pose-pose constraints can be obtained using scan matching [Olson, 2009; Bosse and Zlot, 2009] or visual place recognition tech-

niques [Cummins and Newman, 2008] between arbitrary poses called loop-closures.

We use a horizontal 2D laser scanner to estimate laser odometry. The odometry information is passed to a graph-based, RGB-D SLAM algorithm [Labbe and Michaud, 2014] which corrects the odometry and generates the robot trajectory. The 3D occupancy grid of the environment is then constructed using registered images and depth data.

3.2 Sound Signal Processing

The robot uses a microphone array to process sound data. The employed technique is called multiple signal classification based on standard eigenvalue decomposition (SEVD-MUSIC) [Schmidt, 1986]. Using a microphone array with N_m microphones, a set of transfer functions between a sound source and a microphone array, i.e. a steering vector, is obtained by the geometrical time delay. The steering vector is described as $\boldsymbol{a}(\omega, \boldsymbol{\phi}) = [a_1(\omega, \boldsymbol{\phi}), ..., a_{N_m}(\omega, \boldsymbol{\phi})]^T \in \mathbb{C}^{N_m}$, where ω is frequency and $\boldsymbol{\phi} = [\theta_l, \varphi_l]^T$ is the sound source direction relative to the microphone array with azimuth θ_l and elevation φ_l.

To solve the SSL, a short-time Fourier transform of multi-channel acoustic signals, denoted by $\boldsymbol{\xi}(\omega, k) \in \mathbb{C}^{N_m}$, is obtained at time instant k. The correlation matrix of $\boldsymbol{\xi}(\omega, k)$ can be computed by averaging over T_r frames as

$$\mathbf{R}(\omega, k) = \frac{1}{T_r} \sum_{\tau_r=0}^{T_r-1} \boldsymbol{\xi}(\omega, k+\tau_r)\boldsymbol{\xi}^*(\omega, k+\tau_r) \qquad (1)$$

resulting in a more robust SSL against noise, where $\mathbf{R}(\omega, k) \in \mathbb{C}^{N_m \times N_m}$.

Standard eigenvalue decomposition of $\mathbf{R}(\omega, k)$ decomposes the signal space into the noise and signal subspaces:

$$\mathbf{R}(\omega, k) = \mathbf{E}(\omega, k)\boldsymbol{\Lambda}(\omega, k)\mathbf{E}^{-1}(\omega, k) \qquad (2)$$

in which $\boldsymbol{\Lambda}(\omega, k) = \mathrm{diag}(\lambda_1(\omega, k), ..., \lambda_{N_m}(\omega, k))$ in descending order and $\mathbf{E}(\omega, k) = [\boldsymbol{e}_1(\omega, k), ..., \boldsymbol{e}_{N_m}(\omega, k)]$ respectively denote eigen values and corresponding eigen vectors. The spatial spectrum for SSL can be written as follows:

$$\zeta(\omega, \boldsymbol{\phi}, k) = \frac{|\boldsymbol{a}^*(\omega, \boldsymbol{\phi})\boldsymbol{a}(\omega, \boldsymbol{\phi})|}{\sum_{j=N_s+1}^{N_m} |\boldsymbol{a}^*(\omega, \boldsymbol{\phi})\boldsymbol{e}_j(\omega, k)|} \qquad (3)$$

where N_s is an empirical parameter considered as the number of sound sources in the SSL process in order to remove the noise from the correlation matrix. We estimate the direction of arrival (DOA) over a range of frequencies with the lower and higher cut-off frequencies i_l and i_h, respectively. Therefore, summing out ω in (3) results in

$$\bar{\zeta}(\boldsymbol{\phi}, k) = \frac{1}{i_h - i_l + 1} \sum_{i=i_l}^{i_h} \zeta(\omega_i, \boldsymbol{\phi}, k) \qquad (4)$$

At every time instant k, the local maxima of $\bar{\zeta}(\boldsymbol{\phi}, k)$ with respect to $\boldsymbol{\phi}$ are obtained [Nakamura *et al.*, 2013]. Directions of local maxima which have larger values than a threshold are selected as bearing measurements. Hereinafter, the selected i-th direction $\boldsymbol{\phi}_i$ is denoted as measurement $\boldsymbol{z}_i(k) = [\theta_l, \varphi_l]^T$ and the sound signal processing algorithm is referred to as MUSIC.

4 Sound Source Localization and Tracking Algorithm

In this section, we explain the proposed algorithm to solve the SSLT problem.

4.1 Inverse Depth Parametrization

A major issue with bearing-only measurements is the difficulty in dealing with landmarks that exhibit no parallax during the robot motion due to their extreme depth. The corresponding depth uncertainty cannot be modeled by a standard Gaussian distribution, and these landmarks are considered to be at infinity. This unobservability of landmarks in an Extended Kalman Filter (EKF) framework with standard Cartesian state parametrization makes the landmark initialization non-trivial.

The Inverse Depth Parametrization (IDP) [Civera *et al.*, 2008] provides a suitable solution, allowing the filter to simultaneously track close and infinitely far landmarks. IDP assigns each landmark an inverse depth value, ρ, based on the apparent range to the landmark. By using inverse depth, landmarks at infinity have an inverse depth of zero and can be initialized and tracked. The corresponding depth uncertainty then has a Gaussian that covers uncertainty from nearby the robot to infinity.

4.2 EKF Formulation

Let $\boldsymbol{p}(k) = [x_r(k), y_r(k), z_r(k)]^T$, $\boldsymbol{p}(k) \in \mathbb{R}^3$, and $\boldsymbol{o}(k) = [\psi_r(k), \varphi_r(k), \theta_r(k)]^T$, $\boldsymbol{o}(k) \in SO(3)$, be respectively the robot position and orientation at time step k which are available through the SLAM algorithm. Let $\boldsymbol{l}_j(k) = [x_j, y_j, z_j, \theta_j, \varphi_j, \rho_j]^T$ be the inverse depth parametrization of landmark j at time step k. The state vector $\boldsymbol{x}(k) = [\boldsymbol{l}_1(k), \ldots, \boldsymbol{l}_N(k)]^T$ consists of N initialized landmarks. Assuming the process and measurement noise are additive white zero-mean Gaussian, an EKF is used to recursively estimate

$$\boldsymbol{\mu}(k) = \mathbb{E}[\boldsymbol{x}(k)] \qquad (5)$$

$$\boldsymbol{\Sigma}(k) = \mathbb{E}[(\boldsymbol{x}(k) - \boldsymbol{\mu}(k))(\boldsymbol{x}(k) - \boldsymbol{\mu}(k))^T] \qquad (6)$$

where $\boldsymbol{\mu}(k)$ and $\boldsymbol{\Sigma}(k)$ are the mean and covariance of the state vector estimate at time step k. Note that since the robot pose is not included in the state vector, the corresponding covariance matrix is block-diagonal, i.e. the landmarks are not correlated.

Prediction

The current location of the sound source landmarks relative to the robot cannot be predicted without a measurement, as the landmarks are not correlated with the robot pose. However, the uncertainty of the landmark estimates increases with the robot's motion. To account for this effect, we add a zero-mean white Gaussian noise to the state vector estimate at every time step. Therefore,

$$\boldsymbol{\mu}^{-}(k) = \boldsymbol{\mu}(k-1) \tag{7}$$

$$\boldsymbol{\Sigma}^{-}(k) = \boldsymbol{\Sigma}(k-1) + \boldsymbol{Q}(k) \tag{8}$$

where $\boldsymbol{Q}(k)$ is a diagonal covariance matrix.

Update

In the update step, the predicted state is updated based on the sensor measurements taken in that time step. The Kalman filter requires that observations be linear functions of the state and the next state be a linear function of the previous state, so that the state vector remains a Gaussian. Accordingly, the sensor model $\boldsymbol{h}(\boldsymbol{x})$ is linearized by approximation to a first order Taylor expansion at the current mean.

The recursion to correct the predicted state can be written as follows.

$$\boldsymbol{K}(k) = \boldsymbol{\Sigma}^{-}(k)\boldsymbol{H}(k)^{T}[\boldsymbol{H}(k)\boldsymbol{\Sigma}^{-}(k)\boldsymbol{H}(k)^{T} + \boldsymbol{R}(k)]^{-1} \tag{9}$$

$$\boldsymbol{\mu}(k) = \boldsymbol{\mu}^{-}(k) + \boldsymbol{K}(k)[\boldsymbol{z}(k) - \boldsymbol{h}(\boldsymbol{\mu}^{-}(k))] \tag{10}$$

$$\boldsymbol{\Sigma}(k) = [\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{H}(k)]\boldsymbol{\Sigma}^{-}(k) \tag{11}$$

where $\boldsymbol{H}(k) \triangleq \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}|_{\boldsymbol{\mu}^{-}(k)}$ is the Jacobian calculated to propagate the uncertainty from the observation space to the state space according to Appendix I, and $\boldsymbol{R}(k)$ is the measurement noise covariance.

4.3 Data Association

In this work, we use the Joint Compatibility Branch and Bound (JCBB) data association approach [Tard, 2001]. The JCBB algorithm considers all of the established data association pairings when associating an observation, to limit the possibility of accepting a spurious observation. As the number of pairings in a hypothesis increases, the probability that a spurious pairing is jointly compatible with the hypothesis reduces. The algorithm ensures robust data association when faced with a high density of features in the environment and imprecision of the vehicle location estimate and/or sensor being used. A system dealing with noisy acoustic signals clearly faces the issue of clutter due to sound reverberations creating spurious sound source landmarks in the environment. Therefore, JCBB is a powerful and computationally manageable data association approach to be used in this work.

The JCBB algorithm constructs an interpretation tree with nodes containing an interpretation of the possible associations of preceding measurements [Grimson, 1990]. The individual compatibilities between each observation and landmark are computed using a Mahalanobis Distance gating approach according to

$$d^2 = \boldsymbol{\nu}_{ij}(k)^{T}\boldsymbol{S}_i(k)^{-1}\boldsymbol{\nu}_{ij}(k) < \gamma \tag{12}$$

where observation i observes landmark j and,

$$\boldsymbol{\nu}_{ij}(k) = \boldsymbol{z}_i(k) - \boldsymbol{h}(\boldsymbol{\mu}^{-}(k)) \tag{13}$$

$$\boldsymbol{S}_i(k) = \boldsymbol{H}_j(k-1)\boldsymbol{\Sigma}^{-}(k)\boldsymbol{H}_j^{T}(k-1) + \boldsymbol{R}_i(k) \tag{14}$$

A threshold value, γ, is determined from statistical tables of a χ^2 distribution using the degree-of-freedom of the measurement and the desired confidence level. If any pairings in the tree are above the threshold, they are eliminated. The remaining pairings are then searched to find the maximal data association set; the branch of the tree that has the most number of associations made. If multiple, maximal data association sets exist, then the set with the maximum joint likelihood will be chosen. The Branch and Bound method is used to search all viable solutions in the tree while minimizing computational time and complexity [Cooper, 2005].

4.4 Landmark Initialization

Observations that are not associated with any existing sound source landmarks during the data association stage are used to initialize potential landmarks. This paper follows the landmark initialization approach outlined in [Civera *et al.*, 2008] to initialize new landmarks in a separate state vector and covariance matrix. If a potential landmark is associated with the required number of observations, the landmark is initialized in the state vector.

4.5 Map Management

The underlying observation noise is nonlinear, and EKF can only estimate up to the second moment of the posterior filtering distribution of the state vector. However, to solve the nonlinear filtering problem, moments higher than two are required. As such, the map management plays a key role in the performance of the EKF-based SSLT algorithm. We conceptually follow the map management steps proposed in [Dissanayake *et al.*, 2001]; first, to prevent spurious measurements being initialized as sound source landmarks, and then to ensure that all the confirmed landmarks are of sufficient quality.

Landmark Initialization Management
Two landmark lists are maintained. The state vector stores N confirmed landmarks $\boldsymbol{l}_j$, $j = 1, \ldots, N$. Another list stores M potential landmarks, $\boldsymbol{l}_{p,j}$, $j = 1, \ldots, M$. When a set of observations is received:

1. The observations are associated with the confirmed landmarks in the state vector using the JCBB algorithm described in Section 4.3. If an observation is associated with a landmark, then it is used to update the estimated position of the landmark in the EKF.
2. The unassociated observations from the previous step are associated with the landmarks in the potential list using the JCBB algorithm. If an observation is associated with the jth potential landmark, then the counter c_j corresponding to the landmark is incremented.
3. If an observation is not associated with a confirmed landmark, then the observation is used to initialize a potential landmark according to section 4.4 and a counter c_{M+1} and timer t_{M+1} is initialized.
4. The potential landmark list is examined according to the following criteria
   (a) If the counter corresponding to a landmark is above a user defined threshold, i.e. $c_j > c_{min}$, then the potential landmark is confirmed and moved to the state vector.
   (b) If the time since the landmark was initialized as a potential landmark is above a user defined threshold, i.e. $(k - t_j) > t_{max}$, and the criterion in (a) has not been met, the landmark is permanently removed from the list of potential landmarks.

Landmark Quality Check
The landmark quality is estimated using the probability density function (PDF) of the observations associated with each landmark. The quality Q_j of a landmark can be taken as the ratio of the PDF of all the observations associated with a landmark and the maximum PDF value that would be achieved if all observations coincided with their expected values [Maksarov and Durrant-Whyte, 1995]. Thus

$$Q_j = \frac{\Sigma_{i=1}^{l} \frac{1}{2\pi} \det(\boldsymbol{S}_i(k))^{-\frac{1}{2}} \exp(-\frac{1}{2}\boldsymbol{\nu}_{ij}(k)\boldsymbol{S}_i^{-1}(k)\boldsymbol{\nu}_{ij}^T(k))}{\Sigma_{i=1}^{l} \frac{1}{2\pi} \det(\boldsymbol{S}_i(k))^{-\frac{1}{2}}} \tag{15}$$

where l is the number of observations that have been associated with the jth landmark.

At reasonable intervals, the quality of each landmark is compared to a user defined threshold, Q_{min}, and if $Q_j < Q_{min}$, then the landmark is permanently removed from the state vector.

Table 1: Parameters for SSLT experiments.

Parameter	Symbol	Scenario I	Scenario II
– EKF parameters:			
Quality threshold	Q_{min}	0.6	0.6
Potential landmark counter	c_{min}	300	400
Potential landmark timer	t_{max}	2000	3000
Azimuth standard deviation	σ_θ	5°	4°
Elevation standard deviation	σ_φ	5°	4°
Inverse depth standard deviation	σ_ρ	0.5m^{-1}	0.5m^{-1}
Landmark initialization depth	ρ_0	0.5m	0.5m
Mahalanobis Distance threshold	γ	5.991	5.991
– MUSIC parameters:			
MUSIC Resolution	R	5°	5°
MUSIC Observation Frequency	f	100Hz	100Hz

5 Experimental Results

This section describes two practical demonstrations of the proposed SSLT framework. The purpose of the experiments are twofold. Firstly, the ability of the system to manage spurious measurements caused by sound reverberations and sound signal processing inaccuracies is observed. This determines the ability of the algorithm to correctly initialize the actual sound source landmarks in the environment. Secondly, the system's ability to track and localize these confirmed sound sources is examined. The methods are implemented using Robot Operating System (ROS) [Quigley *et al.*, 2009] and results are processed using MATLAB.

Two experiments are carried out in a room of size $7 \times 4\,\text{m}^2$. The landmarks in the environment are omnidirectional sound speakers which emit consistent white noise. Scenario I comprises 3 speakers placed in stationary positions throughout the room. This experiment tries to replicate a scenario consisting of people sitting or standing in a room and communicating with each other and/or the robot. Scenario II includes one stationary and one moving speaker. This scenario replicates a situation in which there is a person moving throughout the room and communicating with another stationary person or object. For instance, this situation can occur in everyday home activities in which the robot needs to be aware of its surrounding for effective interactions. However, analyzing the content of acoustic signals for decision making is beyond the scope of this paper and is an interesting future research direction. In both cases, the robot is manually controlled as it navigates through the environment. Table 1 displays the parameters used for both experimental scenarios.

5.1 Hardware Specifications

The robot used in the following experiments is a Turtlebot 2 capable of nonholonomic motion along flat surfaces. The Turtlebot is equipped with a Microsoft Kinect v2 RGB-D camera and a Hokuyo UTM-30LX laser scanner. Hereinafter, this robot will be referred to as the test robot.

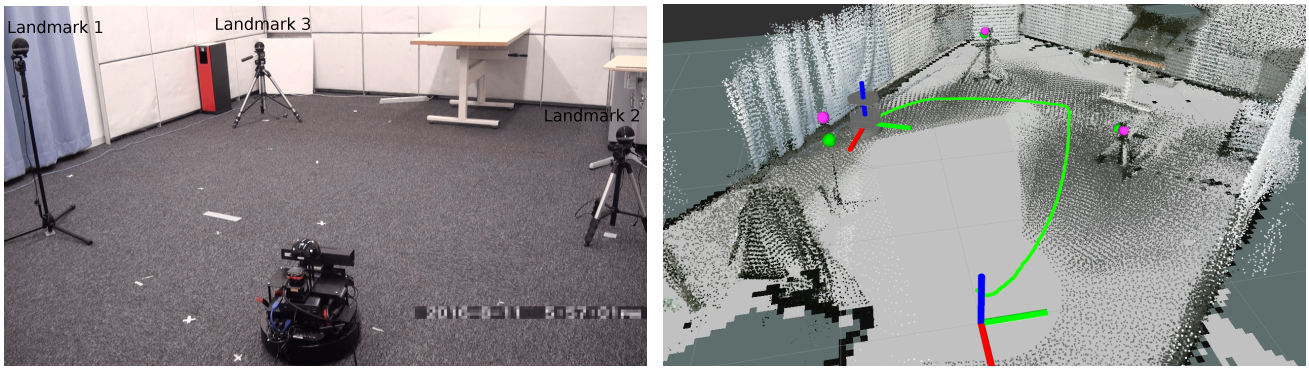

Figure 2: Left figure shows the experimental setup of scenario I consisting of three stationary sound sources and the robot navigating through the environment. Right figure shows the constructed 3D scene of the environment built during the experiment. Groundtruth and estimated locations of the landmarks are represented by pink and RGB markers, respectively. The RGB axis in the foreground represents the map coordinate system, and the green path indicates the robot trajectory.

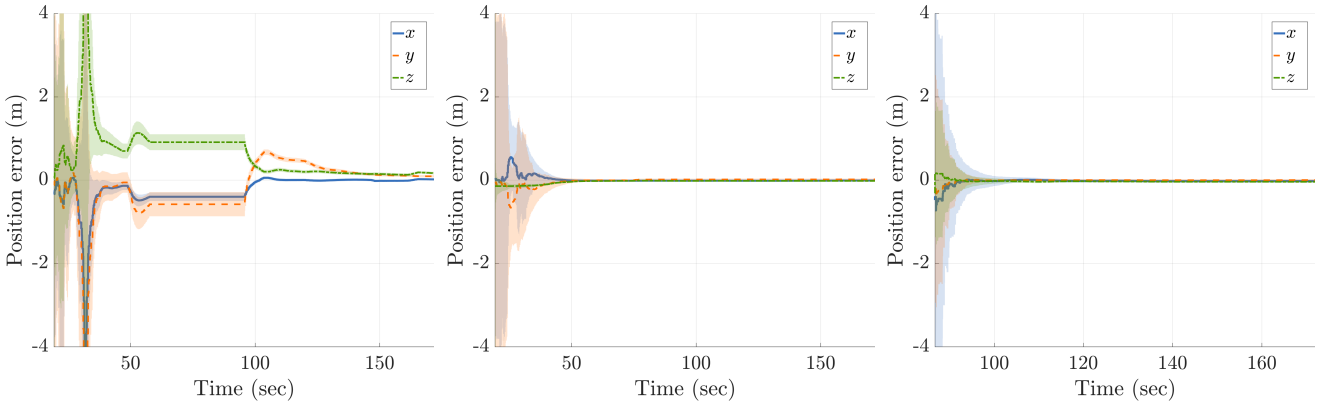

Figure 3: Position error of landmarks 1, 2, and 3 in x, y, and z directions in Scenario I. Shaded region depicts 95% confidence interval.

A UM7 attitude and heading reference system is used to stabilize incoming laser scans and camera images. The circular microphone array employed in the experiments has 8 channels, a sampling rate of 16kHz and a bit depth of 24 bits. A Zotac Magnus Mini-PC is fitted to the Turtlebot for onboard computations. The onboard computer streams the sound source bearing data, robot pose and 3D environment map over Wi-Fi to an external computer.

5.2 Scenario I: Stationary Sound Sources

The quality threshold, potential landmark counter and potential landmark timer values used in this scenario are specific to the microphone array and signal processing software used and thus are determined empirically. The azimuth and elevation standard deviation values of the MUSIC and the microphone array used have not been precisely characterized, so these values are chosen based on the resolution of MUSIC and then tuned experimentally. It is shown by experimental validation that the values of the landmark initialization depth and standard deviation are relatively unimportant as long as they included infinity in the 2σ confidence interval [Civera *et al.*, 2008]. The Mahalanobis distance threshold is chosen based on 95% confidence of correct association.

The test robot starts at the origin of the map and makes a single pass by each of the three stationary sound sources at a speed of approximately 0.05m/s. Figure 2 shows a photo of the environment as well as the 3D occupancy grid built using the RGB-D SLAM. Figure 3 shows the landmarks' position estimation errors. The errors are computed by the difference between the EKF and the groundtruth values in x, y, and z directions for each sound source landmark. Landmarks 1 and 2 are observed from the initial vehicle location, while the third landmark is observed after about 70 seconds. The shaded areas depict the 95% confidence intervals of the corresponding landmark errors. The landmarks are initialized with large covariance values to represent the high

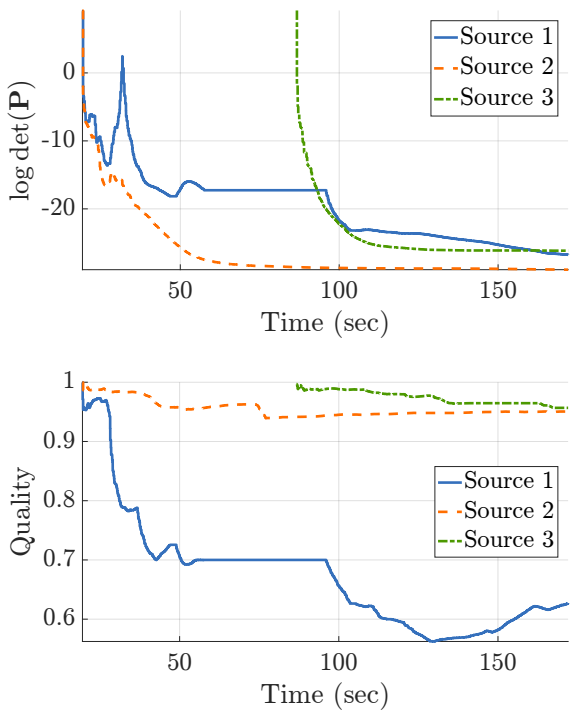

Figure 4: Scenario I; top plot shows the evolution of the determinant of the covariance matrices of each sound source landmark. The graph also shows the efficiency of the estimator in the D-optimal sense. Bottom plot shows the quality of all confirmed landmarks in the environment. Quality ranges from 0 (lowest quality) to 1 (highest quality).

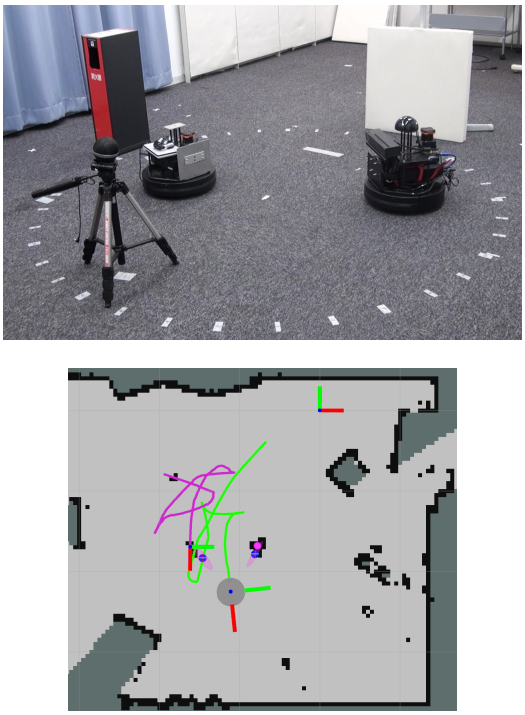

Figure 5: Top figure shows the experimental setup of scenario II consisting of one stationary sound source, one moving sound source, and the robot navigating through the environment. Bottom figure shows the constructed scene of the environment built during the experiment. Groundtruth and estimated locations of the landmarks are represented by pink and RGB markers, respectively. The RGB axis in the top right represents the map coordinate system. The green and purple paths indicate the test robot and the moving sound source's trajectories, respectively.

degree of uncertainty associated with bearing-only measurements. Figure 4 shows the evolution of the logarithm of the determinant of the landmarks' covariance matrices, and the quality of all confirmed landmarks against time.

The power of the sound signal processed by the microphone is inversely proportional to the square of the distance. Consequently, when a low powered sound signal is processed by MUSIC, there will be more frequency bands which are dominated by noise, thus reducing the precision of the final bearing estimate. This effect can be seen in Figure 3(1). During the first half of the run, the landmark error remains high as the test robot moves past the source on the opposite side of the room, at a distance of about 2.5 m. As the test robot approaches the landmark at around the time 100 sec, the filter converges.

5.3 Scenario II: Moving Sound Sources

The environment in this scenario consists of one static and one moving sound source. The moving sound source is placed on a Turtlebot equipped with a 2D laser scanner, which moves throughout the environment. Both the test robot and moving sound source travel at speeds between $0.03 - 0.05$ m / sec.

Figure 6 shows the landmarks position estimation error while the shaded areas depict 95% confidence intervals. Figure 7 shows the evolution of the logarithm of the determinant of the landmarks' covariance matrices, and the quality of confirmed landmarks. In this scenario, both landmarks are observed at the start of the experiment and are initialized as confirmed sound sources after about 14 sec. The moving landmark is initially stationary as the test robot moves past it which allows the EKF to form a reasonable estimate of its location.

As observed in Scenario I, the accuracy of the sensor measurements provided by MUSIC degrade with distance to the source. In order to control the quality of these measurements, the movement of both the test robot and the moving sound source are limited to a $2 \times 2.5\,\mathrm{m}^2$ area. Accordingly, the standard deviation of both the azimuth and elevation bearing values are set to values smaller than those used in Scenario I.

5.4 Discussion

To mitigate noise reverberations that are present in indoor environments, we use delayed landmark initializa-

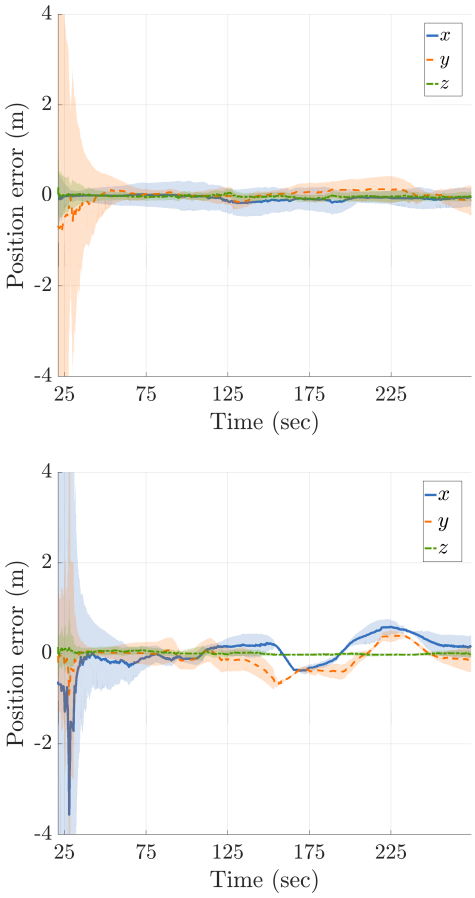

Figure 6: Position error of the stationary (top) and moving (bottom) landmarks from Scenario II in x, y, and z directions. Shaded area depicts 95% confidence interval.

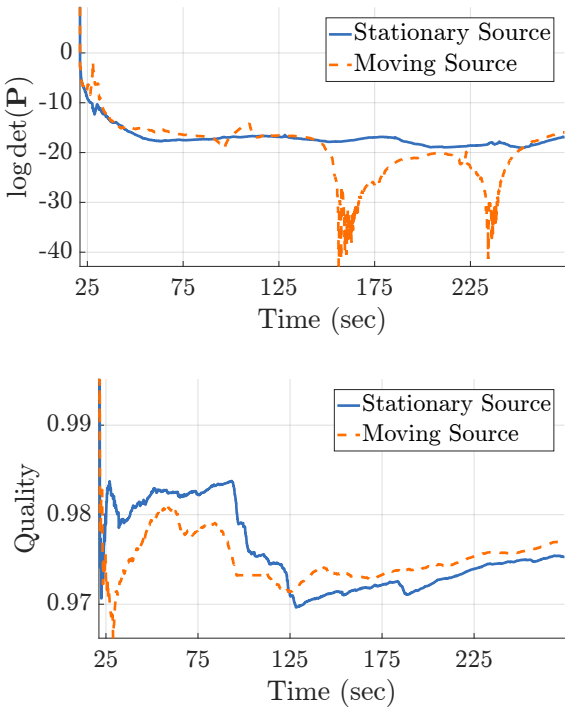

Figure 7: Scenario II; top plot shows the evolution of the determinant of the covariance matrices of the landmarks. Bottom plot shows the quality of all confirmed landmarks in the environment during the experiment.

tion. This process, outlined in Section 4.5, ensures that any sources of sound that are only briefly observed are discounted as spurious. It can, however, track semi-continuous sounds such as human conversation as the number of associations in this case will accumulate over time. Sound reverberations in the test environment are concentrated on the roof of the test room as the roof, unlike the walls, is not sound proofed. Due to the motion of the robot, these reflections do not produce consistent sound bearings and are, therefore, filtered out by the map management techniques (counter threshold).

While this approach vastly improves the algorithm's ability to correctly initialize and associate observations with sound sources, using delayed initialization has associated shortcomings. Currently, once a sound source is initialized as a potential landmark, it is not updated by incoming observations until it is confirmed and initialized in the state vector. Consequently, when a new observation is received, the data association step is performed between the observation and the location that the potential landmark was initialized. If the robot or landmark has moved considerably since the source was initialized, then an observation that should be associated with the landmark may not be associated correctly. Another issue with delayed initialization is that if the potential landmark is confirmed, then all of the measurements of that landmark that are observed leading up to the time of confirmation are lost. This can be particularly damaging if the source is observed for a brief period of time. As the potential landmark counter threshold, c_{min}, increases, the effects of both these issues are exacerbated.

In future work, the observations mentioned above may be stored and incorporated into the state vector and covariance matrix at the time of landmark confirmation, similar to the approach used in [Lemaire *et al.*, 2005]. Alternatively, a separate EKF could be run in parallel to update potential landmarks until they are confirmed.

Sound reverberations also cause issues with the signal processing algorithm. MUSIC encounters problems when there are multiple sound sources with similar acoustic properties located close together. This scenario can be particularly problematic when there is a pseudo sound source created by sound reflections from an actual source. In these situations, MUSIC often cannot differentiate between the two sources as the acoustic properties of the source and its reflection are almost identical. As a result, MUSIC publishes only one bearing value for

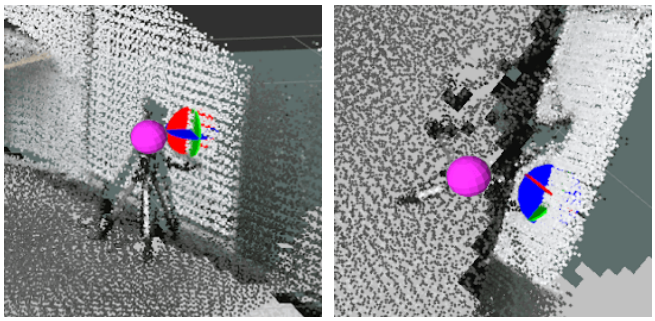

Figure 8: Example of localization of false landmark caused by reflections of sound source reverberating off a nearby wall. Ground truth and estimated locations of the landmark is represented by the pink and RGB markers, respectively. Left image shows a perspective view and the right image shows a top down view of the sound source.

one of the sources. This phenomenon can be observed in Figure 8 in which the omnidirectional sound source indicated by the pink marker is producing reflections off the window behind. In this case, the only bearing measurements provided by MUSIC are of the sound reflections instead of the actual source.

6 Conclusions

In this paper, we propose a unified system that can identify, localize, and track sound sources in indoor environments. The system performs online without any prior knowledge from the environment and auditory conditions of the place. We demonstrate the proposed algorithms through experimental results in medium-sized rooms and the results are promising for applications such HRI. However, the nature of the problem studied in this paper is multi-modal as the reverberations and received observations due to the non-line of sight and reflections present outliers. The map management steps proposed in this work can alleviate the problem, but cannot solve the problem entirely. As such, in combination with the developed tracking filter, a non-linear filtering approach such as particle filters can be applied for more robust outlier detection and rejection. We leave the latter as our future work.

Appendix I: Sensor Model Jacobian

We can calculate the Jacobian of the sensor model, $\boldsymbol{H}(k)$, in low dimensional space for landmark j by

$$ {}^{low}\boldsymbol{H}_j = \frac{\partial \boldsymbol{h}_j}{\partial \boldsymbol{l}_j} = \begin{bmatrix} \frac{\partial \boldsymbol{\theta}_j^-}{\partial x} & \frac{\partial \boldsymbol{\theta}_j^-}{\partial y} & \cdots & \frac{\partial \boldsymbol{\theta}_j^-}{\partial \rho} \\ \frac{\partial \boldsymbol{\varphi}_j^-}{\partial x} & \frac{\partial \boldsymbol{\varphi}_j^-}{\partial y} & \cdots & \frac{\partial \boldsymbol{\varphi}_j^-}{\partial \rho} \end{bmatrix} $$

$$ \begin{aligned} r_x &= (1/\rho_j)\cos(\varphi_j)\cos(\theta_j) + x_j - x_r(k) \\ r_y &= (1/\rho_j)\cos(\varphi_j)\sin(\theta_j) + y_j - y_r(k) \\ r_z &= (1/\rho_j)\sin(\varphi_j) + z_j - z_r(k) \end{aligned} \tag{16} $$

where r_x, r_y and r_z are the Cartesian coordinates of the landmark locations in the world frame.

Let $a = r_x^2 + r_y^2$ and $b = r_x^2 + r_y^2 + r_z^2$. Let $c_\theta = \cos(\theta)$, $s_\theta = \sin(\theta)$, $c_\varphi = \cos(\varphi)$, and $s_\varphi = \sin(\varphi)$. The partial derivatives in Equation (16) can be calculated as follows.

$$ \begin{aligned} &\frac{\partial \boldsymbol{\theta}_j^-}{\partial x} = \frac{-r_y}{a}, \frac{\partial \boldsymbol{\theta}_j^-}{\partial y} = \frac{r_x}{a}, \frac{\partial \boldsymbol{\theta}_j^-}{\partial z} = 0, \frac{\partial \boldsymbol{\theta}_j^-}{\partial \theta} = \frac{c_\varphi(r_x c_\theta + r_y s_\theta)}{\rho a} \\ &\frac{\partial \boldsymbol{\theta}_j^-}{\partial \varphi} = \frac{s_\varphi(r_y c_\theta - r_x s_\theta)}{\rho a}, \frac{\partial \boldsymbol{\theta}_j^-}{\partial \rho} = \frac{c_\varphi(r_y c_\theta + r_x s_\theta)}{\rho^2 a} \end{aligned} \tag{17} $$

$$ \begin{aligned} &\frac{\partial \boldsymbol{\varphi}_j^-}{\partial x} = \frac{-r_z r_x}{b\sqrt{a}}, \frac{\partial \boldsymbol{\varphi}_j^-}{\partial y} = \frac{-r_y r_z}{b\sqrt{a}}, \frac{\partial \boldsymbol{\varphi}_j^-}{\partial z} = \frac{\sqrt{a}}{b} \\ &\frac{\partial \boldsymbol{\varphi}_j^-}{\partial \theta} = \frac{-r_z c_\varphi(r_y c_\theta - r_x s_\theta)}{\rho b\sqrt{a}}, \frac{\partial \boldsymbol{\varphi}_j^-}{\partial \varphi} = \frac{c_\varphi a + r_z s_\varphi(r_x c_\theta + r_y s_\theta)}{\rho b\sqrt{a}} \\ &\frac{\partial \boldsymbol{\varphi}_j^-}{\partial \rho} = \frac{-s_\varphi a + r_z c_\varphi(r_x c_\theta + r_y s_\theta)}{\rho^2 b\sqrt{a}} \end{aligned} \tag{18} $$

References

[Bar-Shalom, 1987] Yaakov Bar-Shalom. *Tracking and data association*. Academic Press Professional, Inc., 1987.

[Bosse and Zlot, 2009] Michael Bosse and Robert Zlot. Continuous 3d scan-matching with a spinning 2d laser. In *Proc. IEEE Int. Conf. Robot Automat.*, pages 4312–4319. IEEE, 2009.

[Civera *et al.*, 2008] Javier Civera, Andrew J Davison, and JM Martinez Montiel. Inverse depth parametrization for monocular SLAM. *IEEE Trans. Robot.*, 24(5):932–945, 2008.

[Cooper, 2005] Aron Jace Cooper. A comparison of data association techniques for simultaneous localization and mapping. Master's thesis, Massachusetts Institute of Technology, 2005.

[Cummins and Newman, 2008] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *The Int. J. Robot. Res.*, 27(6):647–665, 2008.

[Dellaert and Kaess, 2006] Frank Dellaert and Michael Kaess. Square root sam: Simultaneous localization and mapping via square root information smoothing. *The Int. J. Robot. Res.*, 25(12):1181–1203, 2006.

[Dissanayake *et al.*, 2001] M. W. M.Gamini Dissanayake, Paul Newman, Steven Clark, Hugh F. Durrant-Whyte, and M. Csorba. A solution to the simultaneous localization and map building (slam) problem. *IEEE transactions on robotics and automation*, 17(3):229–241, 2001.

[Even *et al.*, 2014] Jani Even, Yoichi Morales, Nagasrikanth Kallakuri, Jonas Furrer, Carlos Toshinori Ishi, and Norihiro Hagita. Mapping sound emitting structures in 3d. In *Proc. IEEE Int. Conf. Robot Automat.*, pages 677–682. IEEE, 2014.

[Grimson, 1990] W. E. L. Grimson. *Object Recognition by Computer: The Role of Geometric Constraints*. Artificial intelligence (Cambridge, Mass.). MIT Press, 1990.

[Hu *et al.*, 2011] Jwu-Sheng Hu, Chen-Yu Chan, Cheng-Kang Wang, Ming-Tang Lee, and Ching-Yi Kuo. Simultaneous localization of a mobile robot and multiple sound sources using a microphone array. *Advanced Robotics*, 25(1-2):135–152, 2011.

[Huang *et al.*, 1999] Jie Huang, Tadawute Supaongprapa, Ikutaka Terakura, Fuming Wang, Noboru Ohnishi, and Noboru Sugie. A model-based sound localization system and its application to robot navigation. *Robot. Auton. Syst.*, 27(4):199–209, 1999.

[Kagami *et al.*, 2009] Satoshi Kagami, Simon Thompson, Yoko Sasaki, Hiroshi Mizoguchi, and Tadashi Enomoto. 2d sound source mapping from mobile robot using beamforming and particle filtering. pages 3689–3692. IEEE, 2009.

[Labbe and Michaud, 2014] M. Labbe and F. Michaud. Online Global Loop Closure Detection for Large-Scale Multi-Session Graph-Based SLAM. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 2661–2666, Sept 2014.

[Lemaire *et al.*, 2005] Thomas Lemaire, Simon Lacroix, and Joan Sol. A practical 3d bearing-only slam algorithm. *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 2757–2762, 2005.

[Maksarov and Durrant-Whyte, 1995] D. Maksarov and H. Durrant-Whyte. Mobile vehicle navigation in unknown environments: a multiple hypothesis approach. *IEE Proc.-Control Theory Appl.*, 142:385–400, 1995.

[Martinson and Fransen, 2011] Eric Martinson and B Fransen. Dynamically reconfigurable microphone arrays. In *Proc. IEEE Int. Conf. Robot Automat.*, pages 5636–5641. IEEE, 2011.

[Matsumoto *et al.*, 2003] Mitsuo Matsumoto, Mikio Tohyama, and Hirofumi Yanagawa. A method of interpolating binaural impulse responses for moving sound images. *Acoustical Science and Technology*, 24(5):284–292, 2003.

[Nakadai *et al.*, 2000] Kazuhiro Nakadai, Tino Lourens, Hiroshi G Okuno, and Hiroaki Kitano. Active audition for humanoid. In *AAAI/IAAI*, pages 832–839, 2000.

[Nakadai *et al.*, 2010] Kazuhiro Nakadai, Hirofumi Nakajima, Gökhan Ince, and Yuji Hasegawa. Sound source separation and automatic speech recognition for moving sources. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 976–981. IEEE, 2010.

[Nakajima *et al.*, 2010] Hirofumi Nakajima, Kazuhiro Nakadai, Yuji Hasegawa, and Hiroshi Tsujino. Blind source separation with parameter-free adaptive step-size method for robot audition. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 18(6):1476–1485, 2010.

[Nakamura *et al.*, 2011] Keisuke Nakamura, Kazuhiro Nakadai, Futoshi Asano, and Gökhan Ince. Intelligent sound source localization and its application to multimodal human tracking. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 143–148. IEEE, 2011.

[Nakamura *et al.*, 2013] Keisuke Nakamura, Kazuhiro Nakadai, and Hiroshi G Okuno. A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition. *Advanced Robotics*, 27(12):933–945, 2013.

[Olson, 2009] Edwin B Olson. Real-time correlative scan matching. In *Proc. IEEE Int. Conf. Robot Automat.*, pages 4387–4393. IEEE, 2009.

[Otsuka *et al.*, 2014] Takuma Otsuka, Katsuhiko Ishiguro, Hiroshi Sawada, and Hiroshi G Okuno. Bayesian nonparametrics for microphone array processing. *IEEE/ACM Trans. on Audio, Speech and Lang. Processing (TASLP)*, 22(2):493–504, 2014.

[Quigley *et al.*, 2009] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*, volume 3, page 5, 2009.

[Sasaki *et al.*, 2009] Yoko Sasaki, Satoshi Kagami, and Hiroshi Mizoguchi. Online short-term multiple sound source mapping for a mobile robot by robust motion triangulation. *Advanced Robotics*, 23(1-2):145–164, 2009.

[Schmidt, 1986] Ralph O Schmidt. Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions on*, 34(3):276–280, 1986.

[Tard, 2001] Juan D Tard. Joint compatibility test. *IEEE Trans. Robot.*, 17(6):890–897, 2001.

[Teh *et al.*, 2006] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American statistical association*, 101(476), 2006.

[Thrun and Montemerlo, 2006] Sebastian Thrun and Michael Montemerlo. The graph SLAM algorithm with applications to large-scale mapping of urban structures. *The Int. J. Robot. Res.*, 25(5-6):403–429, 2006.

[Thrun, 2005] Sebastian Thrun. Affine structure from sound. In *Advances in Neural Information Processing Systems*, pages 1353–1360, 2005.

[Valin *et al.*, 2003] J-M Valin, François Michaud, Jean Rouat, and Dominic Létourneau. Robust sound source localization using a microphone array on a mobile robot. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, volume 2, pages 1228–1233. IEEE, 2003.

[Valin *et al.*, 2004] Jean-Marc Valin, François Michaud, Brahim Hadjou, and Jean Rouat. Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach. In *Proc. IEEE Int. Conf. Robot Automat.*, volume 1, pages 1033–1038. IEEE, 2004.

[Wang *et al.*, 2004] Qing Hua Wang, Teodor Ivanov, and Parham Aarabi. Acoustic robot navigation using distributed microphone arrays. *Information Fusion*, 5(2):131–140, 2004.

[Watanabe *et al.*, 2003] Kanji Watanabe, Shouichi Takane, and Yoiti Suzuki. Interpolation of head-related transfer functions based on the common-acoustical-pole and residue model. *Acoustical science and technology*, 24(5):335–337, 2003.

[Yamamoto *et al.*, 2006] Shun'ichi Yamamoto, Kazuhiro Nakadai, Mikio Nakano, Hiroshi Tsujino, Jean-Marc Valin, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Real-time robot audition system that recognizes simultaneous speech in the real world. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 5333–5338. IEEE, 2006.