3D Localization of a Sound Source Using Mobile Microphone Arrays
Referenced by SLAM

Simon Michaud, Samuel Faucher, François Grondin, Jean-Samuel Lauzon, Mathieu Labbé,
Dominic Létourneau, François Ferland, François Michaud

Abstract—A microphone array can provide a mobile robot with the capability of localizing, tracking and separating distant sound sources in 2D, i.e., estimating their relative elevation and azimuth. To combine acoustic data with visual information in real world settings, spatial correlation must be established. The approach explored in this paper consists of having two robots, each equipped with a microphone array, localizing themselves in a shared reference map using SLAM. Based on their locations, data from the microphone arrays are used to triangulate in 3D the location of a sound source in relation to the same map. This strategy results in a novel cooperative sound mapping approach using mobile microphone arrays. Trials are conducted using two mobile robots localizing a static or a moving sound source to examine in which conditions this is possible. Results suggest that errors under 0.3 m are observed when the relative angle between the two robots are above 30° for a static sound source, while errors under 0.3 m for angles between 40° and 140° are observed with a moving sound source.

I. INTRODUCTION

Over the last 20 years, there has been a growing interest in developing real-time on-board artificial audition capabilities on robots, with libraries like FlowDesigner [1], HARK [2], ManyEars [3] and ODAS [4]. In the recent five years, products equipped with microphone arrays (MAs) (e.g., Amazon Echo, Apple HomePod, Google Home) opened a booming market, and development kits are now commonly available (e.g., ReSpeaker 6-MA, XMOS xCORE 7-MA, 8SoundsUSB 8-MA [3] and 16SoundsUSB 16-MA). Artificial audition technology aims at providing more natural interaction with connected devices such as mobile robots. Artificial audition on a mobile robot can enrich visual perception of the environment by helping to discover interesting elements in real world settings. For instance, a person talking or an object making a sound can be used to draw the robot’s attention to something worth looking at more closely, associating the perceived sound with the image at that same location. This information can then be used to make interesting multimodal associations [5]: face recognition can be used to identify the voice signature of a person, an image tagged to a ring may be designated as a telephone, etc.

Doing so requires associating visual data and audio events in relation to the same reference frame. Visual SLAM (Simultaneous Localization and Mapping) can be used to generate a map of the environment to provide such reference frame of robots equipped with a MA doing sound source localization (SSL). However, assuming that the sound sources are far from the robots compared to the MA aperture (a condition known as the far field effect), SSL only provides elevation and azimuth of sound sources [6]. Triangulating data from two or more MAs can be used to evaluate the 3D location of a sound source, as demonstrated in [7] when the locations of static MAs are known. Using mobile MAs would make it possible, considering that the MA’s positions are derived using SLAM and a shared reference map, to evaluate the 3D location of a sound source without having MAs placed in fixed positions.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Fonds de recherche du Québec - Nature et technologies (FRQNT) and ACERP-3IT Funds, Université de Sherbrooke.

of the results. The closest intersection point of $\lambda_1$ and $\lambda_2$ is used to estimate the 3D location $l_m$ of the sound source. The objective is identify the minimal conditions for 3D triangulation using mobile MAs is possible. The paper is organized as follows. Section II provides an overview of work to situate our approach in relation to combining SSL and SLAM. Section III presents the approach implemented. Section IV describes the experimental setup, followed by Section V with the observed results.

II. RELATED WORK

Rao-Blackwellized particle filter with Kalman filtering are commonly used for tracking sound sources. For instance, Lin et al. [9] estimate the relative poses of a team of mobile robots, each robot equipped with a pair of microphones and emitting a specially-designed sound in 2D. This acoustic data is combined with odometry and filtering is used to resolve the heading angle and the front-back ambiguities, implementing what is referred to as cooperative acoustic robot localization [10]. Teams of micro air vehicles (MAVs) equipped with 4-MAs use a similar concept with Extended Kalman Filtering (EKF) to position themselves in relation to a beacon MAV circling around a reference point in space while emitting continuous predefined acoustic chirps [11].

We identify three categories of approaches combining SSL with SLAM. Acoustic SLAM (aSLAM) makes it possible to localize the trajectory of a MA on a mobile robot whilst estimating the acoustic map of surrounding sound sources [12], [13], [14]. aSLAM basically exploits the movement of a MA to constructively triangulate over time the 3D cartesian location of sound sources from bearing-only 2D Direction-Of-Arrival (DoA) measurements, estimating the robot trajectory from the apparent displacement of sound sources observed from multiple positions. aSLAM performance is therefore affected by the trajectory followed by the MA in relation to the sound sources. Only validated in simulation, this approach is limited to a single robot with mapping referenced to the robot’s initial position, and requires at least two sound sources to work. Similar limitations apply to [15], [16] which adopt a similar approach using Kalman filtering. In the same category, Sasaki et al. [17], [18] derive 2D positions of multiple sound sources using a 32-MA and sound observations over the last 2 sec, and sound source categorization is used to remove undesirable cross points. In more recent work, Sasaki et al. [19] designed a hand-held unit equipped with a 3D lidar and Inertial Measurement Unit (IMU) for SLAM, and using HARK with a 32-MA for SSL. Particle filtering from data taken over time (from 7.5 to 15 sec) with the hand-held unit moving (rotation, displacement) provides 3D positions of two sound sources.

An extension to aSLAM is audio-visual SLAM (AVSLAM), exploiting acoustic and visual features [20] for human tracking. In [20], validation is done using one robot equipped with a 8-MA running HARK and a RGB camera moving in a straight line over 2.5 m in front of a stationary human sound source, providing only 2D localization. Bayram and Ince [21] present audio-visual multi-person 2D tracking by doing sensor fusion of a SSL module with a visual face recognition module. Results are presented using two Kinect cameras and a 7-MA running HARK.

Finally, the concept of audio-based SLAM [6] involves considering the SSL problem as a SLAM problem. For instance, the FastSLAM [22] algorithm is used to estimate the time offset and position of robots equipped with MAs and the position of sound sources [23]. Sekiguchi et al. [6] use FastSLAM with static MA-equipped robots to consider the MAs as one big array. Results using HARK and three static robots with 8-MAs in an anechoic chamber and two moving talkers are provided. Audio-based SLAM has also been used for online calibration of asynchronous MAs [24], [25] and for optimizing the relative positions of multiple mobile robots with MAs for cooperative sound source separation [26].

III. SOUND SOURCE MAPPING USING MOBILE MAs

Our approach aims at using two mobile MAs to provide instantaneous 3D location of a sound source. In relation to Section II, our approach differs by having mobile MAs, localized using SLAM based on $m$, triangulate sound source locations in 3D also in relation to $m$. The concept can be designated as cooperative sound mapping, illustrated by the architecture diagram presented by Fig. 1. Each mobile robot $i$ is equipped with a lidar, a RGB-D camera and onboard odometry to localize its location $l_i^m \in \mathbb{R}^3$ (3D position and a quaternion for rotation) in relation to a reference map $m$ using RTAB-Map. Each robot $i$ is also equipped with a 16-MA and uses ODAS to do sound source localization, tracking and separation. ODAS provides a 3D unit vector $\mathbf{x}_i \in S^2$ pointing in the direction of the sound source with respect to the robot. Using data from two mobile robots ($i = \{1, 2\}$), the Cooperative Sound Mapping module triangulates the position of the sound source in relation to the reference map.

A. RTAB-Map

RTAB-Map (Real-Time Appearance-Based Mapping) [27] is an open source library1 implementing graph-based SPLAM (Simultaneous Planning, Localization And Mapping) [28] i.e., the ability to simultaneously map an environment, localize itself in it and plan paths. RTAB-Map provides the robots’ positions and orientations, denoted as $l_i^m$ and $l_i^R$, respectively. RTAB-Map uses a combination of odometry, lidar and camera to robustly create a map and to localize in it. The lidar is used to create the 2D occupancy grid map for obstacle avoidance and path planning. Appearance-based loop closure detection and localization are done with visual features extracted from the RGB image of the RGB-D camera using a bag-of-words approach. By estimating the 3D positions of visual features using the depth image, a position and orientation against the map can be computed. The localization is then refined using the lidar to improve accuracy when environments are lacking visual features but has a lot of geometry (which is often the case indoor).

---

1http://introlab.github.io/rtabmap/
B. ODAS

ODAS [4] is an open source library² performing sound source localization, tracking and separation. ODAS generates the DoA for each of the two robots \( (i = \{1,2\}) \), denoted as \( \lambda_1 \) and \( \lambda_2 \). This library relies on a localization method called Steered Response Power with Phase Transform based on Hierarchical Search with Directivity model and Automatic calibration (SRP-PHAT-HSDA). The proposed approach decomposes the search space in coarse and fine grids, which speeds up the search in 3D for the DoA of one or many sound sources. Localization generates noisy potential sources, which are then filtered with a tracking method based on a modified 3D Kalman filter (M3K) that generates one or many tracked sources. Sound sources are then filtered and separated using directive geometric source separation (DGSS) to focus the robot’s attention only on the target sound source, and ignore ambient noise. This library also models microphones as sensors with a directive polar pattern, which improves sound source localization, tracking and separation when the direct path between microphones and the sound sources is obstructed by the robot’s body.

In this work, ODAS is configured to return the loudest sound source DoA per robot, denoted as \( \lambda_1 \) and \( \lambda_2 \) for robots \( i = \{1,2\} \). Time synchronization between \( \lambda_1 \) and \( \lambda_2 \) is facilitated using ODAS³ tracking module output because the DoAs are smoothed over time.

C. Cooperative Sound Mapping

The Cooperative Sound Mapping module combines the information provided by the two robots to determine the location of the sound source \( \mathbf{1}_m^S \in \mathbb{R}^3 \). It first rotates the DoA for each robot in relation to its orientation to derive the vectors \( \mathbf{1}_m \mathbf{1}_1 \) and \( \mathbf{1}_m \mathbf{1}_2 \). In 3D space, \( \mathbf{1}_m \mathbf{1}_1 \) and \( \mathbf{1}_m \mathbf{1}_2 \) rarely intersect each other. The estimation of \( \mathbf{1}_m^S \) is derived by finding the smallest distance between \( \mathbf{1}_m \mathbf{1}_1 \) and \( \mathbf{1}_m \mathbf{1}_2 \), as represented by the dotted line in Fig. 1. Using the Ray to Ray algorithm [29] as in [7], the sound source position is estimated using (1):

\[
\mathbf{1}_m^S = \frac{1}{2} (\mathbf{1}_m^1 + G_1 \mathbf{1}_m \lambda_1) + G_2 \mathbf{1}_m \lambda_2)
\]

(1)

where the expressions \( G_1 \) and \( G_2 \) are given as follows:

\[
G_1 = \frac{(\mathbf{1}_m \mathbf{1}_1 \cdot \mathbf{1}_m \lambda_1 \cdot (\mathbf{1}_m^1 - \mathbf{1}_m^2)) - \mathbf{1}_m \cdot (\mathbf{1}_m^1 - \mathbf{1}_m^2)}{1 - (\mathbf{1}_m \mathbf{1}_1 \cdot \mathbf{1}_m \lambda_1)^2}
\]

(2)

\[
G_2 = \frac{(\mathbf{1}_m \mathbf{1}_2 \cdot (\mathbf{1}_m^2 - \mathbf{1}_m^1)) - \mathbf{1}_m \cdot (\mathbf{1}_m^1 - \mathbf{1}_m^2)}{1 - (\mathbf{1}_m \mathbf{1}_2 \cdot \mathbf{1}_m \lambda_2)^2}
\]

(3)

and (\cdot) stands for the dot product.

IV. EXPERIMENTAL METHODOLOGY

Two SecurBot⁴ mobile robots are used: one equipped with a Jetson Nano core mounted on a TurtleBot2 base, and the other with a Jetson TX2 installed on a modified Pioneer2-DX base. Both robots are equipped with a 16SoundsUSB⁵ MA, an Intel Realsense D435 camera and a RP Lidar. Each MA is located 0.48 m above the ground and provides synchronous acquisition of microphone signals through USB to the robot’s computer. Figure 3 and Table I present the MA configuration of each robot.

The experiments are conducted in a 150 m² room filled with different objects to provide visual features for SLAM. The reference map is created using RTAB-Map with one of the robot. The room has a reverberation level of \( RT_{60} = 600 \) msec and no background noise. ODAS is configured with similar parameters as those in [4], except for the covariance

TABLE I: Position of the microphones on the MA (cm)

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>( \Delta x_1 )</th>
<th>( \Delta x_2 )</th>
<th>( \Delta y_1 )</th>
<th>( \Delta y_2 )</th>
<th>( \Delta z_1 )</th>
<th>( \Delta z_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pioneer2-DX</td>
<td>19.1</td>
<td>33.2</td>
<td>24.5</td>
<td>36.0</td>
<td>2.5</td>
<td>4.0</td>
</tr>
<tr>
<td>TurtleBot2</td>
<td>18.5</td>
<td>31.0</td>
<td>19.3</td>
<td>28.4</td>
<td>2.6</td>
<td>3.9</td>
</tr>
</tbody>
</table>

²http://odas.io

³https://github.com/introlab/securbot

⁴https://github.com/introlab/16SoundsUSB
component of the observation noise matrix in the Kalman filter of ODAS’ tracking module, which is increased to \( \sigma_{R}^2 = 0.01 \) for more sensitivity to sound source acceleration, which influences \( \lambda \).

![3D view](image)

\[ \Delta x_1 \]
\[ \Delta x_2 \]
\[ \Delta y_1 \]
\[ \Delta y_2 \]

**Fig. 3: MA configurations**

Trials are conducted with the sound source located at a static location at 1.12 m of height or by being manually moved horizontally and vertically, with its location being monitored using a Vicon motion system. The sound source is a loudspeaker generating white noise, with perceived amplitude ranging from 25 to 30 dB over 3 m. Robots move around the sound source by following preset trajectories defined in relation to the reference map. Figure 4 shows the trajectories followed by the robots using ROS’s navigation stack [30]. These trajectories have the robots move from 0 to 2 m/sec and are set to avoid collisions between the two robots and to cover a variety of DoA configurations in relation to the sound source. A remote laptop computer, also running ROS, monitors the trials and records the localization and audio data. RViz is used to display the position of the robots, the map, the DoAs, the estimated and the known locations of the sound source. RViz also displays Root Square Error (RSE) in meter between \( l_{m} \) and the actual sound source location with colored dots ranging from green to black from 0 to 0.5 m, and to red as it increases.

![Top view](image)

**Fig. 4: Experimental conditions with the two SecurBot robots**

**V. RESULTS**

RSE is examined in relation to \( d_{m} \), the distances between the robots and the sound source, and the angle \( \theta \) as defined by Fig. 5. \( \theta \) is the angle between the lines from the theoretical location of the sound source to \( l_{1} \) and \( l_{2} \) in 3D, as shown by Fig. 1.

![Side view](image)

**Fig. 5: Definition of \( d_{m} \) and \( \theta \)**

Figure 6 and Fig. 7 summarizes the observations made during a trial with a static sound source and a moving sound source. For the static sound source, data is recorded at 10 Hz. Fig. 6a to Fig. 6c illustrate that RSE remains lower than approximately 0.3 m with negligible variance for \( \theta \) higher than approximately 30°, regardless of the distances between the robots and the sound source, as illustrated by RViz in Fig. 8. These results are confirmed by Fig. 6c which represents, over intervals of 15°, the average RSE.

---

5http://www.ros.org
and its standard deviation. For smaller $\theta$ (which occur at the first and the last $\sim$18 sec of the static sound source trial), $\lambda_1$ and $\lambda_2$ are becoming parallel, and small changes in $\theta$ lead in larger errors. When parallel, the denominator of (1) results in a division by zero. So when $\theta$ is small, the closest intersection point found between $\lambda_1$ and $\lambda_2$ quickly changes as the denominator of (1) comes closer to zero. Figure 9 illustrates this situation. One alternative to limit such occurrences would be to use a third robot, to actively reposition the robots to have $\theta$ higher than $15^\circ$, or to use robots with MAs at different heights.

The static sound source trial limits the possible range for $\theta$ because the sound source is located higher than MAs. The moving sound source trial makes it possible to have $\theta$ change from 0 to $180^\circ$. Fig. 7a to Fig. 7b present data (recorded at 100 Hz to synchronize with the Vicon system). Between 0 to 30 sec, the robots are immobile and near each other, and the sound source is moving. Because $\theta$ is small ($9^\circ$), large errors are observed as explained with the static condition. From 30 to 61 sec, robots are moving away from each other, with one coming closer to the sound source being static: as $\theta$ increases, RSE decreases. The peak at 55 sec is caused by $\sigma^2_{R}$, overshooting $l_{S}$ based on the influence of the sound source acceleration as robots are near each other and the sound source starts moving. From 61 to 80 sec, large peaks occur as $\theta$ is small (i.e., $\lambda_1$ and $\lambda_2$ are almost parallel) and the robots’ motion makes $\theta$ change quickly. From 80 to 155 sec, $\theta$ is sufficient to have small RSE except when $\theta$ is near $180^\circ$ (111 sec) and at 120 sec when there was an obstacle between one robot and the sound source. For the remaining time of the trial, $\theta$ rapidly decreases toward 0, creating RSE peaks and larger errors. Figure 7c suggest that for $\theta$ between $\sim$40° to $\sim$140°, the RSE is lower than 0.3 m.

VI. CONCLUSION

This paper validates the concept of cooperative sound mapping by demonstrating, using RTAB-Map ad ODAS, that it is possible to derive the 3D location of a sound source.
Overall complexity for cooperative sound mapping, targeting of sound source tracking instead of SSL will simplify the measurements according to their positions in relation to the positioning of the mobile robots to provide reliable 3D location and reverberation conditions, and coordinate the positioning of the mobile robots to provide reliable 3D location measurements according to their positions in relation to the sound sources. We believe that directly using the output of sound source tracking instead of SSL will simplify the overall complexity for cooperative sound mapping, targeting onboard centralized or distributed processing.

REFERENCES