Robust Sound Source Mapping using Three-layered Selective Audio Rays for Mobile Robots

Daobilige Su[1,2], Keisuke Nakamura[2], Kazuhiro Nakadai[2] and Jaime Valls Miro[1]

***Abstract*— This paper investigates sound source mapping in a real environment using a mobile robot. Our approach is based on audio ray tracing which integrates occupancy grids and sound source localization using a laser range finder and a microphone array. Previous audio ray tracing approaches rely on all observed rays and grids. As such observation errors caused by sound reflection, sound occlusion, wall occlusion, sounds at misdetected grids, etc. can significantly degrade the ability to locate sound sources in a map. A three-layered selective audio ray tracing mechanism is proposed in this work. The first layer conducts frame-based unreliable ray rejection (sensory rejection) considering sound reflection and wall occlusion. The second layer introduces triangulation and audio tracing to detect falsely detected sound sources, rejecting audio rays associated to these misdetected sounds sources (short-term rejection). A third layer is tasked with rejecting rays using the whole history (long-term rejection) to disambiguate sound occlusion. Experimental results under various situations are presented, which proves the effectiveness of our method.**

I. Motivation and Background

The ability of a robot to build a map of its surroundings is a fundamental characteristic required for autonomous navigation in unknown spaces. Most *Simultaneous Localization And Mapping (SLAM)* systems which are implemented for indoor environments are vision or LIDAR based [1]. Despite substantial developments with these sensing modalities, audio-based mapping is still in its primitive phase, and remains an open subject of research given the particularly challenging conditions associated to environmental acoustic noise and reflections. Because of its importance, e.g. for Human-Robot Interaction, sound source mapping has recently become a main challenge in the field of robot audition, and several methods have been reported.The existing methods can be mainly categorized into two approaches.

The first approach combines *Sound Source Localization (SSL)* and robot's odometry within localization strategies such as triangulation [2], particle filters [3], FastSLAM [4], Evidence Grids [5] or PSFS [6]. The approach is relatively easy to implement as relies only on a microphone array mounted on the robot. Its performance is relatively unaffected by external factors such as room shape etc. although it is contrained by two critical factors:

1) The robot needs to see all sound sources from different angles so as to locate the sound sources precisely.
2) Sound reflection is not taken in consideration, resulting in performance degradation in reverberant environments.

Issue 1 is particularly apparent when the robot drives directly towards a sound source; the sound source is observed from only one angle and the methods will eventually fail. This situation is more likely to happen when the robot is driving in a narrow corridor and there is one sound source at the end of the corridor. Issue 2 becomes especially critical when applying the methods in indoor environments.

The second approach relies on occupancy grids in addition to SSL and odometry to develop a ray tracing approach to detect sound sources [7]. Thanks to the fusion of SSL with the distance scan data, locations of sound sources can be obtained by one single position of the robot. Therefore, sensing sound sources from different angles is no longer needed, which solves the issue 1). This approach has mainly the following assumptions:

3) An audio ray hit a sufficiently narrow area of occupied grids so that the sound location is uniquely determined.
4) All sound sources are on occupied grids.
5) Sounds do not pass through occupied grids.

However, these assumptions are not always satisfied in real-world applications. For instance, the assumption 3) is problematic when there is wall occlusion. Especially when localizing a single isolated sound source, the wall behind the true sound source will get higher probability and be mistakenly detected as a sound source [7]. The assumption 4) is not met especially when using a planar laser range finder. If there is any sound source that cannot be scanned by the laser, this method will trace the location of the sound source to the obstacle behind it and fail. The assumption 5) is also not satisfied if there are acoustically transparent materials or low walls which accept diffraction, which induces the wall in front of the true sound source will mistakenly get higher probability.

In this paper a mechanism for sound source mapping suitable for real environments able to tackle the above issues is investigated. Following on the ground work of our previous approach [8], we use audio rays combined with occupancy grids to solve issue 1). To solve other issues, this paper proposes a three-layered selective audio ray tracing inspired by the multi-store model. The first layer conducts frame-based unreliable ray rejection (*sensory rejection*) considering sound reflection and wall occlusion to solve issues 2) and 3). To solve issue 4), the second layer introduces triangulation [2] using all observed audio rays to detect sounds at misdetected

[1]Daobilige Su and Jaime Valls Miro are with Centre for Autonomous System (CAS), University of Technology, Sydney (UTS), Australia daobilige.su@student.uts.edu.au, jaime.vallsmiro@uts.edu.au

[2]Keisuke Nakamura and Kazuhiro Nakadai are with Honda Research Institute Japan Co., Ltd., Wako, Saitama, 351-0114, Japan {keisuke,nakadai}@jp.honda-ri.com

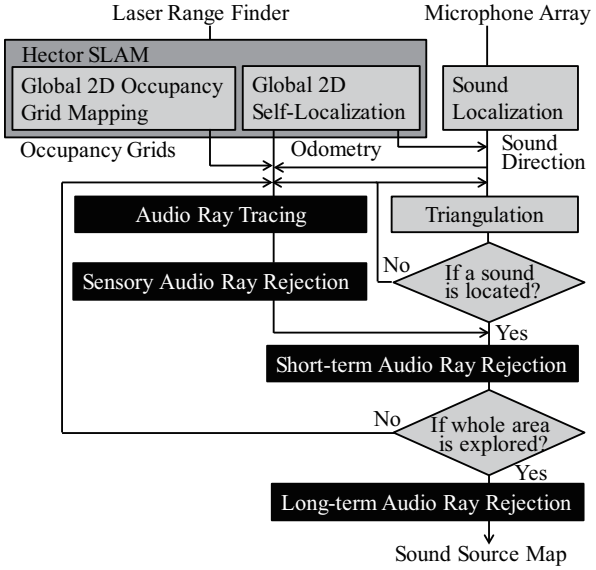


Fig. 1. Block Diagram of Sound Source Mapping

grids and reject audio rays related to the misdetected sounds source (*short-term rejection*). A third layer then rejects rays using the whole history (*long-term rejection*) to disambiguate the sound occlusion to solve issue 5).

II. Proposed Method

Fig. 1 shows the process flow of the proposed sound source mapping. Same as our previous approach [8], we use a laser range finder and a microphone array to compute robot odometry, occupancy grids, and direction of sound sources. The audio ray tracing block is an extension of the conventional ray tracing [7] described in Section II-A. Other three black blocks conduct the three-layered audio ray selection. The sensory audio ray rejection block is conducted every frame (Section II-B.1). When a new sound source gets localized, the short-term audio ray rejection is conducted (Section II-B.2). Finally, the long-term audio ray rejection is conducted after the robot finishes exploration (Section II-B.3). Below, we explain the algorithm of each block briefly.

A. Audio Ray Tracing Using SSL and Occupancy Grids

In our previous approach [8], we used MUltiple SIgnal Classification (MUSIC [9]) for SSL and Hector SLAM [10] for occupancy grid mapping.

Since this paper focuses on 2D sound source mapping, MUSIC is used to estimate azimuth of sound sources in the robot coordinate, denoted as ψ^r, where the superscript r represents that the status is in the robot coordinate C^r. Same as our previous work [8], the spatial spectrum is computed frame-by-frame, denoted as $P(\psi^r, f)$, where f is the frame index. This paper defines the estimated direction as a set of ψ^r having local maxima of $P(\psi^r, f)$. To select reliable local maxima, we use two different thresholds depends on if there are a lot of occupied grids close to the robot or not. We observed that if a robot is close to large obstacles, such as a corner of walls, reverberation gets stronger, resulting in higher spatial spectrum. Thus, we choose a higher threshold T_h for the situation in which the robot is close to a large

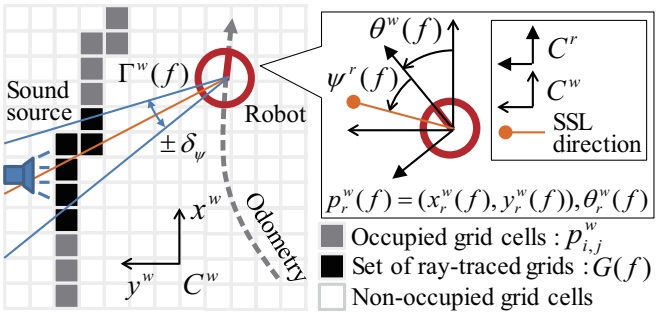


Fig. 2. Model of Audio Ray Tracing

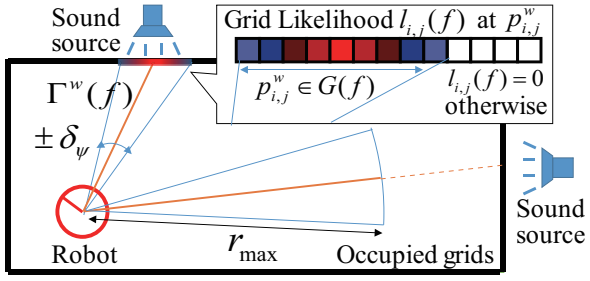


Fig. 3. Example of Audio Rays and Their Likelihoods

obstacle and a lower threshold T_l otherwise. Hereinafter, the estimated direction at the f-th frame is represented as $\psi^r(f)$.

Same as our previous work [8], we use Hector SLAM to estimate both 2D planar robot odometry and occupancy grids in the f-th frame, namely 2D robot location $\boldsymbol{p}_r^w(f) = [x_r^w(f), y_r^w(f)]^T$, the robot orientation $\theta_r^w(f)$, and the 2D occupied grid location $\boldsymbol{p}_{i,j}^w = [x_{i,j}^w, y_{i,j}^w]^T$, where the superscript w represents that the status is in the world coordinate C^w, and i, j is the horizontal and vertical index of grids in the occupancy grid mapping.

Fig. 2 shows the model of audio ray tracing. In the f-th frame, SSL estimates $\psi^r(f)$ in the robot coordinate C^r. Then, an audio ray is casted from the robot in the direction of $\psi^w(f) = \psi^r(f) + \theta_r^w(f)$ in C^w with the maximum range $r_{\max}$, which is hereinafter described as $\Gamma^w(f)$. The conventional ray tracing [7] uses particle filtering to estimate the robot location, so a set of audio rays casted from all weighted particles is used to compute the probability of the sound existence. This paper, on the other hand, estimates one robot pose using Hector SLAM and can use only one audio ray. Therefore, the sound probability is computed under assumption that the probability has a normal distribution with a mean in the direction of $\psi^w(f)$ and a standard deviation of δ_ψ as shown in Fig. 3. The likelihood of traced occupied grids is computed as follows:

$$l_{i,j}(f) = \mathcal{N}\left(\frac{\psi^w(f) - \arg(\boldsymbol{p}_{i,j}^w - \boldsymbol{p}_r^w(f))}{\delta_\psi} \middle| 0, 1 \right) , \quad (1)$$

where $\mathcal{N}(x|0,1)$ is a 0 mean 1 sigma Gaussian distribution. Here, Eq. 1 is computed only for a set of grids, satisfying

$$G(f) = \left\{ \boldsymbol{p} \in \boldsymbol{p}_{i,j}^w \middle| \, |\psi^w(f) - \arg(\boldsymbol{p} - \boldsymbol{p}_r^w(f))| \leq \delta_\psi \right\} , \quad (2)$$

meaning the grids are within $\psi^w(f) \pm \delta_\psi$. If $\boldsymbol{p}_{i,j}^w \notin G(f)$, $l_{i,j}(f) = 0$. As ray tracing process continues, the sound

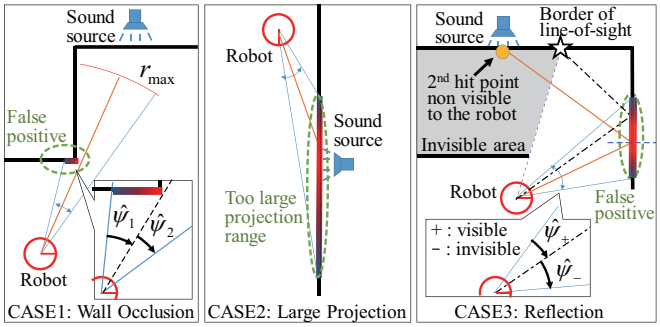


Fig. 4. Cases for Sensory Audio Ray Rejection

source likelihood value of each grid (i, j) will be simply summed up as follows:

$$\mathcal{L}_{i,j}(f) = \mathcal{L}_{i,j}(f-1) + l_{i,j}(f) \ . \tag{3}$$

Finally, to define the probability of each grid being a sound source, these accumulated likelihood rescaled between 0 and 1 as below:

$$\bar{\mathcal{L}}_{i,j}(f) = 1 - \exp(-\alpha_s \mathcal{L}_{i,j}(f)), \tag{4}$$

where $\bar{\mathcal{L}}_{i,j}(f)$ is the probability of the (i, j) grid. α_s is the parameter to tune the exponential curve, which controls the sensitivity of sound source mapping.

B. Three-layered Audio Ray Selection

1) Sensory Audio Ray Rejection: The first-layer selects audio rays frame-by-frame considering the issues 2) and 3). Fig. 4 shows examples of false positives in detecting sound sources induced by the issues.

CASE1 is the wall occlusion, where audio rays are projected more than one wall even if there is one sound source. To identify this case, $\boldsymbol{p}^w_{i,j} \in G(f)$ is checked if the grids are linked together. If not fully linked, we compute the azimuth occupancy rate of unlinked sets of grids as shown in CASE1 in Fig. 4, denoted as $\hat{\psi}_k(f)$. Finally, the f-th audio ray $\Gamma^w(f)$ is rejected if satisfying the following condition:

$$\forall k : \frac{|\hat{\psi}_k(f)|}{2\delta_\psi} < \varepsilon_1 \ . \tag{5}$$

In **CASE2**, the audio ray is projected in a large area of the occupancy grids since the robot is close to the wall, which is not desirable to locate the sound source. Thus, $\Gamma^w(f)$ is rejected if satisfying the following condition:

$$\sqrt{\mathrm{var}(\boldsymbol{X}^w_{i,j}) + \mathrm{var}(\boldsymbol{Y}^w_{i,j})} > \varepsilon_2 \ , \tag{6}$$

where $\boldsymbol{X}^w_{i,j}$ and $\boldsymbol{Y}^w_{i,j}$ are sets $x^w_{i,j}$ and $y^w_{i,j}$ when $\boldsymbol{p}^w_{i,j} \in G(f)$.

CASE3 shows an example of false positive due to reflection. This situation can be eliminated by checking the observability of the reflected rays. Same as our previous work [8], we assume the reflection follows the image model [12]. Thus, we first compute the border line between visible and invisible areas shown as a dotted line in CASE3 in Fig. 4 and obtain the cross point of the border and the wall, described as a star in the figure. The audio ray hitting the star, denoted as a chained line in the figure, is the border line of rays if the reflection is visible or not. If the most of reflected area is not directly observable, it is difficult to estimate whether the sound is direct or reflected. Therefore, we reject rays based on the azimuth occupancy rate divided by the chained line as follows:

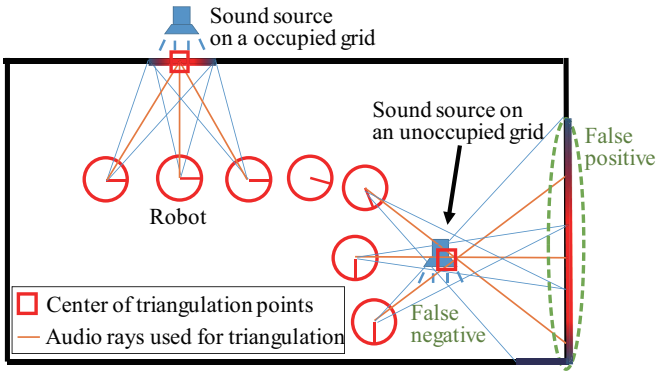


Fig. 5. Cases for Short-term Audio Ray Rejection

$$\frac{|\hat{\psi}_+|}{2\delta_\psi} < \varepsilon_3 \ . \tag{7}$$

2) Short-term Audio Ray Rejection: The second layer detects sound sources which are not detected by the laser range finder and rejects rays related to these sounds. As mentioned in Section I, the audio ray tracing is suitable for localizing sound sources attached to obstacles that can be scanned by the laser range finder, which is hereinafter called *on-wall* sound sources. However, the method will fail if there are sound sources that cannot be scanned by the laser, which is hereinafter called off-wall sound sources. The triangulation [2] does not use the laser range finder and can detect off-wall sound sources, but it needs the robot to navigate and observe all sound sources from different angles. The second layer introduces the triangulation [2] and integrate it with the audio ray tracing to solve the issues.

As shown in Fig. 5, the audio ray tracing generates false positives behind the off-wall sound source. In order to reject these rays, firstly clusters of triangulation points are classified into on-wall and off-wall sources by simply thresholding the distance from each cluster center (red boxes in Fig. 5) to the closest occupied grid, with a thresholding parameter ε_r. As it is observed from empirical results that audio ray tracing has better localization accuracy than triangulation for an on-wall sound source, the localization result from audio ray tracing is adopted. For off-wall sound sources, localization results from triangulation are adopted, and falsely projected rays need to be removed. Every time when a off-wall cluster is detected, the rays having triangulation points of the detected off-wall sound source are rejected.

3) Long-term Audio Ray Rejection: After the sound source mapping, the third layer disambiguate false positives in detecting sound sources in consideration of all sound source location and the occupancy grids. As shown in Fig. 6, low height walls (or obstacles) or acoustically transparent walls generate false positives. Here, we assume that N sound sources have been detected and located at $\boldsymbol{p}^w_{sn}$, where $1 \leq n \leq N$. This type of false rays is eliminated after the

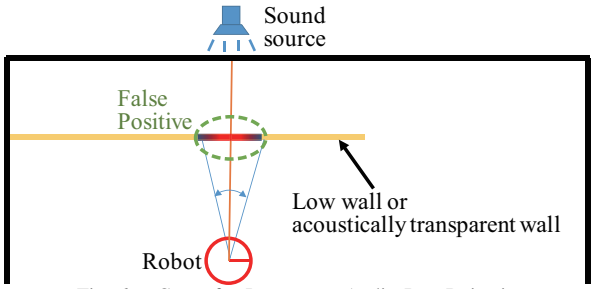


Fig. 6. Cases for Long-term Audio Ray Rejection

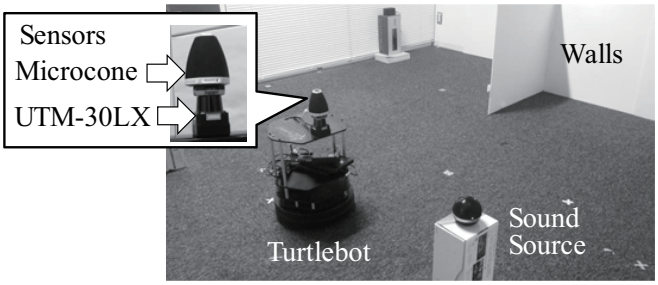


Fig. 7. Experimental Equipment

robot navigation terminates as follows:

$$\forall \Gamma^w(f), \boldsymbol{p}^w_{sn}: \\ |\arg(\Gamma^w(f)) - \arg(\boldsymbol{p}^w_{sn} - \boldsymbol{p}^w_r(f))| < \varepsilon_{\mathrm{L1}} \ \& \\ \|\boldsymbol{p}^w_{sn} - \boldsymbol{p}^w_r(f)\| > \|\boldsymbol{p}^w_\Gamma(f) - \boldsymbol{p}^w_r(f)\| + \varepsilon_{\mathrm{L2}} \ ,$$

where $\boldsymbol{p}^w_\Gamma(f)$ is the 2D location of the traced grid by $\Gamma^w(f)$. If there is a occluded sound source behind where $\Gamma^w(f)$ hits the obstacle grids (at least further away by $\varepsilon_{\mathrm{L1}}$) and the direction from robot pose to the occluded sound source is sufficiently close (less than $\varepsilon_{\mathrm{L2}}$), the rays are rejected. After this rejection, grids in sound probability map will be re-clustered and detected sound sources will be updated.

When the robot navigation terminates, k-means [13] algorithm is used to cluster the grids of sound probability occupancy map with the maximum number of clusters $N_{\max}$, and the minimum distance between each cluster $\Delta_{\min}$. Sound source probability $\bar{\mathcal{L}}_{i,j}(f)$ associated to each cell (i,j) is treated as the weight during clustering. After clustering, those clusters whose maximum probability of contained grids is higher than a predefined threshold will be determined as a valid sound source.

C. Parameters Selection and Learning

The parameters in the proposed method can be learned/tuned as described below. δ_ψ and α_s (described in section II-A) can be learned by observing a sound source with the practical hardware. Specifically, δ_ψ represents a microphone array's SSL observation noise and can be estimated by observing a static sound source with the actual microphone array and computing the standard deviation from the SSL observations.

α_s can be obtained by operating a mobile robot around a sound source with a wide angle observation base line (90-180 degree) in a less reverberant environment. In this case, the robot should fully observe the sound source with high certainty of that coming from a given sound source. Therefore, α_s can be obtained by setting the maximum normalized accumulated likelihood $\bar{\mathcal{L}}_{i,j}(f)$ in Eq. 4 to a large confidence value (0.95 0.99).

ε_1, ε_2 and ε_3 (section II-B.1) determine how strictly we want the proposed three layered audio ray rejection to be. High values of ε_1, ε_3 and low value ε_2 mean we will reject most audio ray estimates. This however implies that sometimes we might reject all audio rays. On the other hand, being too flexible means we will accept most of the audio rays and the proposed method becomes similar to the work proposed in [7]. This will lead to an increase in false positive detection, particularly in reverberant environments.

ε_r (section II-B.2) represents the minimum distance of the closest off-wall sound source to the wall. Empirical observations indicate that in practical settings it can be safely set to 0.2m-0.5m.

$\varepsilon_{\mathrm{L1}}$ (section II-B.3) is also related to microphone array's observation noise, and therefore it can be set equal to δ_ψ.

Lastly, $\varepsilon_{\mathrm{L2}}$ (section II-B.3) depends on the width of the wall and hence can be set to 0.2m-0.4m for most cases.

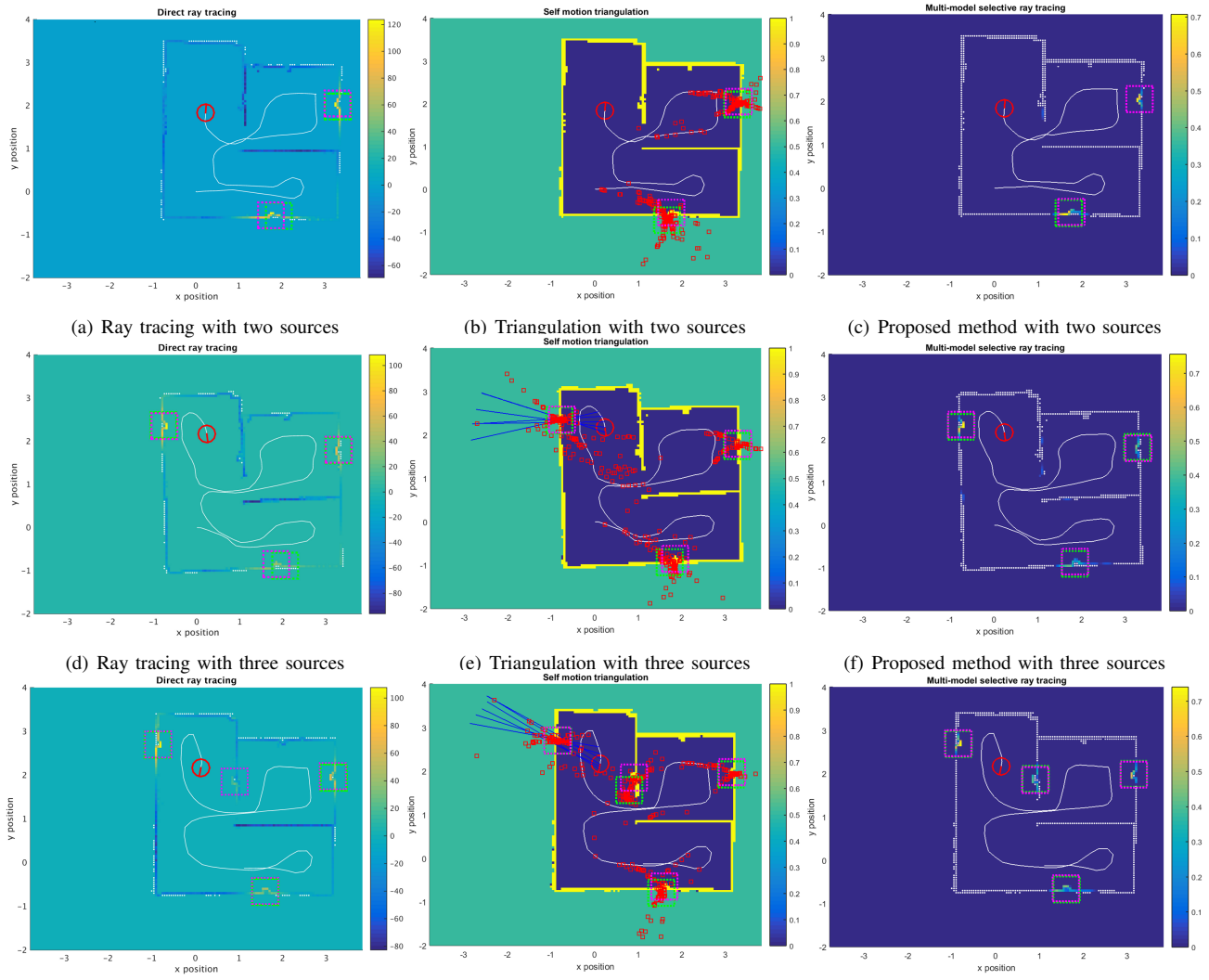
III. Experimental Validation

In this section, validation results of our method is presented. The proposed system explained above was implemented with a Turtlebot (see Fig. 7), located in a normal room whose reverberation time was 0.2 seconds. The room size is 7.0m $\times$ 4.0 m. We utilized Hokuyo UTM-30LX for Hector SLAM.

For SSL, we utilized a Microcone manufactured by DevAudio which has an 6-ch circular microphone array and 1-ch microphone on the top. All sensors were mounted on the robot as shown in Fig. 7. We computed the transfer functions of the Microcone using a wave propagation model, whose resolution was 5°. The acoustic signal was sampled with 16 kHz and 16 bits. The window and shift length for frequency analysis were set to 512 and 160 samples, respectively. For SSL, we utilized MUSIC[9] implemented in the robot audition software, HARK[14]. White noise is used for sound source mapping.

For validation, sound source localization accuracy using the proposed method, ray tracing as proposed by [7] and triangulation as per [2] are compared under various different situations are presented. As all three methods have been implemented by the authors for this exercise, the key parameters employed are collected in TABLE I. Throughout the comparison given by Figs. 8, Figs. 10, 11, and 12, we use the same notation. The green boxes represent locations of detected sound sources using kmeans algorithm and the magenta boxes represent actual ground truth locations of the sound sources. The white dots in ray tracing represent the obstacles grids that are not hit by rays. Red markers in the triangulation method denote triangulated points.

Fig. 8 shows sound source mapping results using three different methods with on-wall sound sources. Figs. 8(a)-(c) shows the results with two on-wall sources. As can be seen from the figure, both the ray tracing and the triangulation

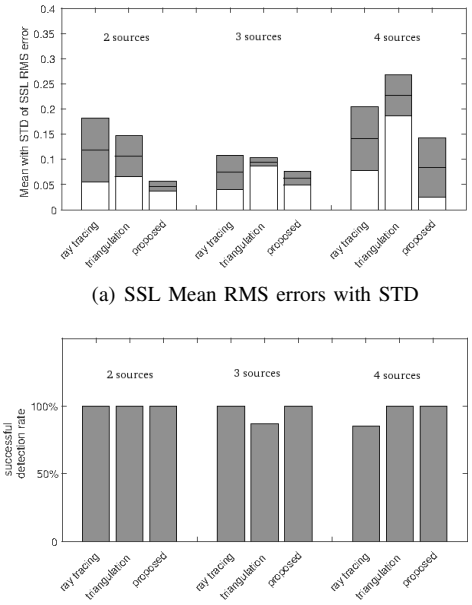


(a) Ray tracing with two sources (b) Triangulation with two sources (c) Proposed method with two sources

(d) Ray tracing with three sources (e) Triangulation with three sources (f) Proposed method with three sources

(g) Ray tracing with four sources (h) Triangulation with four sources (i) Proposed method with four sources

Fig. 8. Sound Source Mapping with only on-wall sources



(a) SSL Mean RMS errors with STD

(b) sound source successful detection rate

Fig. 9. Mean RMS error with STD and successful detection rate after 5 times Monte Carlo runs for each situation

TABLE I

KEY PARAMETERS SETTINGS IN THE EXPERIMENTS

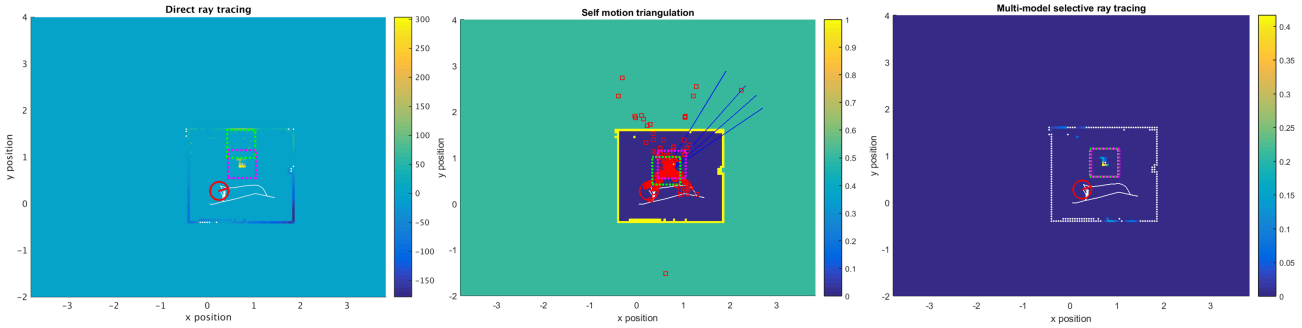
General parameters	Values
$r_{\max}$	3m
$N_{\max}$	5
$\Delta_{\min}$	0.5m

Ray tracing [7]	Values
SSL	SRP-PHAT
L_{min}	0.01
L_{max}	0.99
α	0.9
$\mathcal{L}_{\max}$	400
ε_{max}	31.0
ε_{range}	1.0
ε_{norm}	31.5

Proposed method	Values
SSL	MUSIC
δ_ψ	10 degree
α_s	1.0
ε_1	0.7
ε_2	0.15m
ε_3	0.6
ε_r	0.5m
$\varepsilon_{\mathrm{L1}}$	10 degree
$\varepsilon_{\mathrm{L2}}$	0.3m

Triangulation [2]	Values
# of obs.	10
Outlier rejection	RANSAC

detected two sources (Fig. 8(a) and Fig. 8(b)), but the accuracy was not high due to the large projection of audio rays. The proposed method in Fig. 8(c) localized two sources with higher accuracy than conventional methods of ray tracing and triangulation regarding the large projection in the sensory audio ray rejection. Similar results are observed in 3 and 4 sound sources situations. In the case of four sound sources,



(a) Ray tracing with one source (b) Triangulation with one source (c) Proposed method with one source

(d) Ray tracing with one source (e) Triangulation with one source (f) Proposed method with one source

Fig. 10. Sound Source Mapping with one isolated sound source and one sound source in a highly sound reflective corridor

the ray tracing missed one sound source. By reducing α in the ray tracing, the fourth sound source can be detected, but we keep the value since the low α degrade the accuracy of SSL.

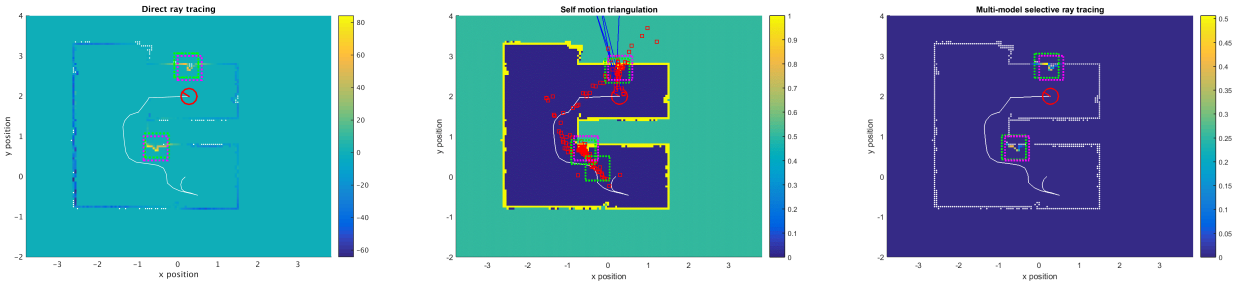
Then, in order to quantitatively evaluate our SSL accuracy, we run 5 Monte Carlo runs for each experiment shown above. Mean RMS errors and standard deviation (STD) of SSL in sound source mapping along with successful sound source detection rates after 5 times Monte Carlo runs are shown in bar graph in Fig. 9. As can be seen from two bar graphs, the proposed method can detect all sound sources successfully in all runs and localization accuracies are better than conventional ray tracing and triangulation methods for all three cases. As the number of sound sources increase, reverberation condition becomes more evidence and SSL accuracy will reduce. Therefore mean RMS errors of four sound source is larger than those of two and three sound sources.

Next, we studied two challenging cases of SSL, which are localization of an isolated sound source and SSL in a highly reflective environment. Figs. 10(a)-(c) shows the results for an isolated sound source which is a challenging case as stated in [7]. As seen in Fig. 8(a), the ray tracing assigned positive sound source likelihood value not only to the sound source in the middle, but also to the wall behind it as explained in [7]. This results in the centroid of the cluster stayed at the middle and not accurately localized the source. As seen in Fig. 8(b), the triangulation detected one cluster corresponds to the true sound source. However, the accuracy is not high due to the variability of the triangulation points. As seen in Fig. 8(c), our method could detect one source with better accuracy, which confirms that the three-layered audio ray selection worked for an isolated sources considering wall occlusion and isolated source classification based on triangulation. Figs. 10(d)-(f) shows results for SSL in a highly reflective corridor. The ray tracing method in 10(d) and The triangulation method in 10(e) all detected a false positive sound source due to reflection from the wall at the bottom. Our proposed method, thanks to the **CASE3** of the proposed three-layered Audio Ray Selection II-B, most of false positive audio rays reflected from the bottom wall are successfully rejected.

The comparison with low height walls is shown in Fig. 11 when having one source (Figs. 11(a)-(c)) and two sources (Figs. 11(d)-(f)). Two horizontal walls in the middle right side of space are low height walls. The ray tracing in Fig. 11(a) detected two clusters. One cluster in the middle were due to the false rays hitting to the low hight walls, caused by the issue 5) in Section I. The triangulation in Fig. 11(b) got a lot of false triangulation points in the beginning since the sound propagated through the low walls. As a result, it mistakenly detected a sound source in the bottom. The proposed method detected one sound source close to the ground truth. Although many rays hit the wall in the middle at the beginning, most of these rays were eliminated by the sensory audio ray rejection, resulting in the RMS error of 0.0412m.

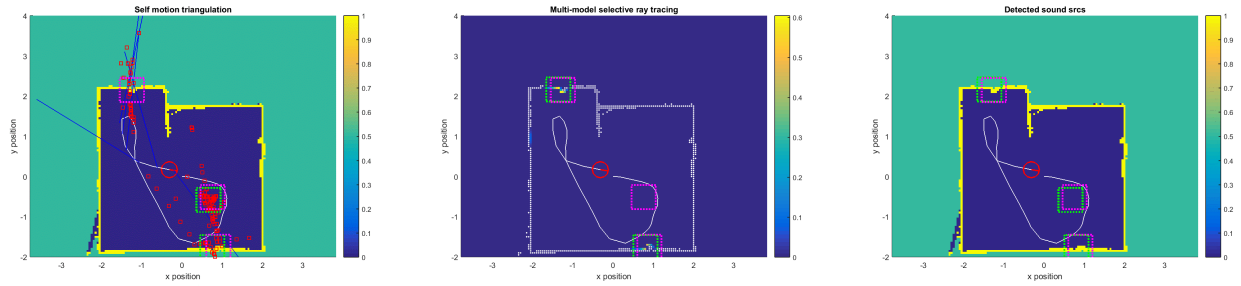
Figs. 11(d)-(f) show the mapping result with two sound sources. Here, we put the second source where the ray tracing got false positives in Fig. 11(a) so as to challenge the ambiguous situations. All three methods detected the two sound sources except the triangulation detected one more sound source in the beginning of robot trajectory due to many SSL when the robot stayed at the start point. The

(a) Ray tracing with one source (b) Triangulation with one source (c) Proposed method with one source



(d) Ray tracing with two sources (e) Triangulation with two sources (f) Proposed method with two sources

Fig. 11. Sound Source Mapping (with low height walls in the middle)



(a) Triangulation (b) Proposed method after sensory rejection (c) Proposed method after all layers

Fig. 12. Sound Source Mapping with both on- and off-wall sources

proposed method in Fig. 11(f) successfully detected two sources, which means that the long-term audio ray rejection worked for disambiguating the auditory occlusion. Finally, the proposed method estimated both sources with the RMS error of 0.1042m.

Fig. 12 shows the performance of the proposed method when both on-wall and off-wall sound sources exist. As seen in Fig. 12(a), triangulation detects two sound sources in the middle and bottom. However the sound source on top is not detected since the robot observes this sound source from very narrow angle and performance of triangulation becomes very poor in this case. The proposed method after the sensory audio ray rejection is shown in Fig. 12(b). It can be seen that two on-wall sources are successfully detected. The final localization result after the short-term and long-term audio ray rejection is shown in Fig. 12(c). As seen in the figure, most of false positives due to the off-wall sound source in the middle are removed successfully, and all 3 sound sources are successfully detected, whose RMS error was 0.1138m.

IV. Conclusions

In this paper sound source mapping for mobile robots in real environments is investigated. Our approach is based on ray tracing, and proposed three-layered audio ray selection to robustify ray tracing toward finding sound sources in real environments, considering the effect of anomalies such as sound reflections, wall occlusions, etc. We evaluate the mapping performance in practical environments and compare it with conventional methods to confirme considerable improvements in all tested scenarios. We did not consider the influence (accuracy, time) of robot motion planning during the sound source mapping process, which is left for future work.

References

[1] S. Thrun, B. Wolfram, and F. Dieter, *Probabilistic robotics*, vol. 1, Cambridge: MIT press, 2005.

[2] Y. Sasaki *et al.*, "Online Short-Term Multiple Sound Source Mapping for a Mobile Robot by Robust Motion Triangulation", *Advanced Robotics*, vol. 23, no. 1–2, pp. 145–164, 2009.

[3] S. Kagami *et al.*, "2D sound source mapping from mobile robot using beamforming and particle filtering", in *IEEE ICASSP*, pp.3689–3692, 2009.
[4] H. Jwu-Sheng *et al.*, "Simultaneous localization of mobile robot and multiple sound sources using microphone array", in *IEEE ICRA*, pp. 29–34, 2009.
[5] E. Martinson *et al.*, "Discovery of sound sources by an autonomous mobile robot", *Autonomous Robots*, vol. 27, no. 3, pp. 221–237, 2009.
[6] C.-C. Wang *et al.*, "Probabilistic structure from sound", *Advanced Robotics*, vol. 23, no. 12–13, pp. 1687–1702, 2009.
[7] N. Kallakuri *et al.*, "Probabilistic approach for building auditory maps with a mobile microphone array", in *IEEE ICRA*, pp. 2270–2275, 2013.
[8] G. Narang *et al.*, "Auditory-aware navigation for mobile robots based on reflection-robust sound source localization and visual SLAM", in *IEEE SMC*, pp. 4021–4026, 2014.
[9] R. Schmidt, "Multiple emitter location and signal parameter estimation", *IEEE Trans. Ant. Prop.*, vol. 34, no. 3, pp. 276–280, 1986.
[10] S. Kohlbrecher *et al.*, "A Flexible and Scalable SLAM System with Full 3D Motion Estimation", in *IEEE SSRR*, pp. 155–160, 2011.
[11] M. S. Brandstein *et al.*, "A robust method for speech signal time-delay estimation in reverberant rooms", in *IEEE ICASSP*, vol. 1, pp. 375–378, 1997.
[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics", *J. Acoust. Soc. Am.* vol. 65, no. 4, 943 (1979).
[13] D. Arthur *et al.*, "k-means++: The advantages of careful seeding", in *18th annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, 2007.
[14] K. Nakadai *et al.*, "Design and Implementation of Robot Audition System HARK", *Advanced Robotics*, vol. 24, pp. 739–761, 2009.