Reflection and Diffraction-Aware Sound Source Localization

Inkyu An[1] Jung-Woo Choi[2] Dinesh Manocha[3] Sung-Eui Yoon[1]
School of Computing[1] and Dept. EE[2], KAIST, Sounth Korea
Dept. of CS & ECE, Univ. of Maryland at College Park, USA[3]

Abstract

We present a novel sound localization algorithm for a non-line-of-sight (NLOS) sound source in indoor environments. Our approach exploits indirect properties, the reflection and diffraction, of sound waves. We suggest a ray tracing based sound propagation algorithm for modeling the high order reflections. We then combine a ray tracing based sound propagation algorithm with a Uniform Theory of Diffraction (UTD) model, which simulate bending effects by placing a virtual sound source on a wedge in the environment. Our method identifies the convergence region of those generated acoustic rays as the estimated source position based on a particle filter. We have evaluated our algorithm in the scenario consisting of a dynamic NLOS sound source. In our tested case, modeling reflection sound using high order reflections improves the localization accuracy by 92% compared to only using direction sound. We also present a novel scheme to model NLOS sources by using an approximate diffraction formulation and observe the accuracy improvement of 125%.

1. Introduction

It is getting more imperative to understand environments in many fields such as computer vision and robotics. There have been many kinds of research efforts to perceive the environment by acquiring and using data from hardware sensors (e.g., GPS, CCD or depth cameras, and acoustics). One of the main research topics for understanding environments focuses on localizing objects and events occurring in such objects.

There have been many efforts to recognize and localize the object in the environment based on visual data [8, 3, 4]. They have focused on the sound accompanied by the event of the object. The majority of events and accompanied sounds are highly correlated and it can be a clue to solve the object localization problem. Based on the correlation between events and sounds, they tried to train the deep neural network from unlabeled data to localize the object generating sound.

Localization is an also fundamental capability required by an autonomous robot, as the current location is used to guide the future movement or actions [7, 10]. These methods reconstruct the environment to the virtual 3D space from RGB-D sensor data and, then, localize the current location of the robot from reconstructed data.

Recently, there are considerable works on using acoustic sensors for localization. The acoustic sensors use the properties of sound waves to compute the sound location. As the sound waves are emitted from a source, they transmit through the media and either reach the listener or microphone locations as direct paths, or after undergoing different wave effects including reflections, interference, diffraction, scattering, etc. Some of the earliest work on sound source localization (SSL) makes use of the time difference of arrival (TDOA) at the receiver [5]. These methods only exploit the direct sound and its direction at the receiver, and do not take into account of reflections or other wave effects. As a result, it does not provide sufficient accuracy for many applications.

In many scenarios, the sound source is not directly in the line of sight of the listener, i.e., NLOS (Non-Line-of-Sight), and is occluded by obstacles. In such cases, indirect sound propagation paths emitted at the source should be significant and prominent for the listener In particular, we focus on the reflection and diffraction propagation paths, and model these indirect sound effects based on using ray-based geometric propagation paths. Furthermore, a full version of this article was published to ICRA 2018 [2] and ICRA 2019 [1], and we provide a brief summary here.

2. Reflection and Diffraction-Aware SSL

We present a novel sound localization algorithm that takes into diffraction effects, especially from non-light-of-sight or occluded sources. Our approach is built on a ray tracing framework and models diffraction using the Uniform Theory of Diffraction (UTD) [6] along with the wedges. During the precomputation phase, we use SLAM and reconstruct a 3D triangular mesh for an indoor environment. The reconstructed 3D triangular mesh is used to generate the reflection and diffraction acoustic ray at run-

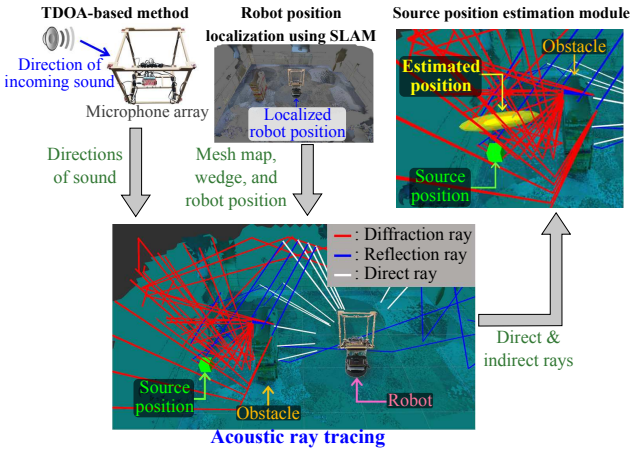


Figure 1: We show run-time computations using acoustic ray tracing with diffraction rays for sound source localization. The diffraction-aware acoustic ray tracing is highlighted in blue and our main contribution in this paper. This figure refers to Fig. 3 in [1].

time.

2.1. Reflection-Aware Ray Tracing

In this section, we explain how our acoustic ray tracing technique generates direct and reflection.

At runtime, we first collect the directions of the incoming sound signals from the TDOA algorithm. For each incoming direction, we generate a primary acoustic ray in the backward direction; as a result, we perform acoustic ray tracing in a backward manner. At this stage, we cannot determine whether the incoming signal is generated by one of the states: direct propagation, reflection, or diffraction. We can determine the actual states of these primary acoustic rays while performing backward acoustic ray tracing. Nonetheless, we denote this primary ray as the direct acoustic ray since the primary ray is a direct ray from the listener's perspective.

We represent a primary acoustic ray as r_n^0 for the n-th incoming sound direction. Its superscript denotes the order of the acoustic path, where the 0-th order denotes the direct path from the listener. We also generate a (backward) reflection ray once an acoustic ray intersects with the scene information under the assumption that the intersected material mainly consists of specular materials [2]. The main difference from the prior method [2] is that we use a mesh-based representation, while the prior method used a voxel-based octree representation for intersection tests. This mesh is computed during precomputation and we use the triangle normals to perform the reflections. As a result, for the n-th incoming sound direction, we recursively generate reflection rays with increasing orders, encoded by a ray path that is defined by $R_n = [r_n^0, r_n^1, ...]$. The order of rays increases

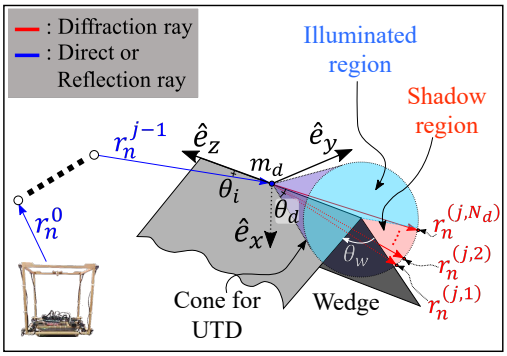


Figure 2: This figure illustrates our acoustic ray tracing method for handling the diffraction effect. Suppose that we have an acoustic ray r_n^{j-1} satisfying the diffraction condition, hitting or passing near the edge of a wedge. We then generate N_d diffraction rays converging the possible incoming direction (especially, in the shadow region) of rays that cause the diffraction. This figure refers to Fig. 4 in [1].

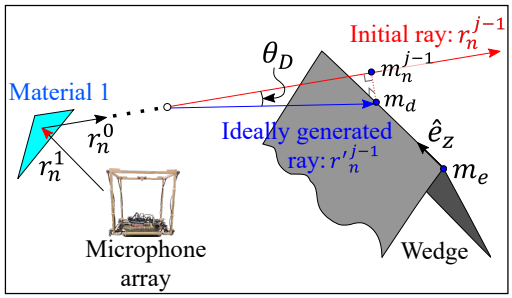


Figure 3: The diffraction condition. When a ray r_n^{j-1} passes close to an edge of a wedge, we consider the ray to be generated by the edge diffraction. We measure the angle θ_D between the ray and its ideal generated ray that hits the edge exactly, for checking our diffraction condition. This figure refers to Fig. 5 in [1].

as we perform more reflection and diffraction.

2.2. Diffraction-Aware Ray Tracing

We now explain our algorithm to model the diffraction effects efficiently within acoustic ray tracing to localize the sound source. Since our goal is to achieve fast performance in localizing the sound source, we use the formulation based on Uniform Theory of Diffraction (UTD) [6]. The incoming sounds collected by the microphone array consist of contributions from different effects in the environment, including reflections and diffractions.

Edge diffraction occurs when an acoustic wave hits the edge of a wedge. In the context of acoustic ray tracing, when an acoustic ray hits an edge of a wedge between two neighboring triangles, the diffracted signal propagates into all possible directions from that edge. The UTD model assumes that the point on the edge causing the diffrac-

tion effect is an imaginary source generating the spherical wave [6].

In order to solve the problem of localizing the sound source, we simulate the process of backward ray tracing. Suppose that an n-th incoming sound direction denoted by the ray r_n^{j-1} is generated by the diffraction effect at an edge. In an ideal case, the incoming ray will hit the edge of a wedge and generate the diffraction acoustic ray r_n^j, as shown in Fig. 2. It is important to note that there can be an infinite number of incident rays generating diffractions at the edge. Unfortunately, it is not easy to link the incident direction exactly to the edge generating the diffraction. Therefore, we generate a set of N_d different diffraction rays in a backward manner that covers the possible incident directions to the edge based on the UTD model. This set is generated based on an assumption that one of those generated rays might have the actual incident direction causing the diffraction. When there are sufficient acoustic rays, including the primary, reflection, and diffraction rays, it is highly likely that those rays will pass through or near to the sound source location.

This explanation begins with the ideal case, where the acoustic ray r_n^{j-1} hits the edge of a wedge. Because our algorithm works on the real environment containing various types of errors from sensor noises and resolution errors from the TDOA method, it is rare that an acoustic ray intersects an edge exactly.

In order to support various cases that arise in real environments, we propose using the notion of *diffraction-condition* between a ray and a wedge. The diffraction-condition simply measures how close the ray r_n^{j-1} passes to an edge of the wedge. Specifically, we define the *diffractability* v_d according to the angle θ_D between the acoustic ray and its ideally generated ray for the diffraction with the wedge: i.e. $v_d = \cos(\theta_D)$, where the cos function is used to normalize the angle θ_D (Fig. 3).

If the diffractabilty v_d is larger than a threshold value, e.g., 0.95 in our tests, our algorithm determines that the acoustic ray is generated from the diffraction at the wedge, and we thus generate the secondary, diffraction ray at the wedge in the backward manner.

We now present how to generate the diffraction rays when the acoustic ray satisfies the diffraction-condition. The diffraction rays are generated along the surface of the cone (Fig. 2) because the UTD model is based on the principle of Fermat [6]: the ray follows the shortest path from the source to the listener. Ideally, we have to generate every set of the shortest paths on the cone for UTD model. However, because of the limitation of the resources, we generate the N_d number of diffraction rays on the cone for UTD model.

In order to accelerate the process, we only generate the diffraction rays in the shadow region, which is defined by the wedge; the rest of the shadow region is called the illuminated region. We use this process because covering only the shadow region but not the illuminated region generates minor errors for a simulation of the sound propagation [9].

Given the new diffraction rays, we apply our algorithm recursively and generate another order of reflection and diffraction rays. Given the n-th incoming direction signal, we generate acoustic rays, including direct, reflection, and diffraction rays and maintain the ray paths R_n in a tree data structure. The root of this tree represents the direct acoustic ray, starting from the microphones. The depth of the tree denotes the order of its associated ray. Note that we generate one child and N_d children for handling reflection and diffraction effects, respectively.

2.3. Estimating the Source Position

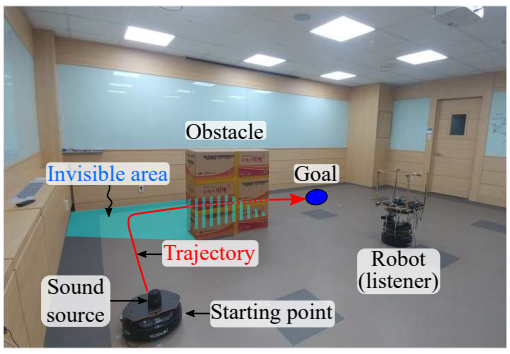
We explain our method used to localize the sound source position using acoustic rays. Our estimation is based on Monte-Carlo localization (MCL), also known as the particle filter [2]. Our estimation process assumes that there is a single sound source in the environment, which causes a high probability that all those acoustic ray paths pass near that source. In other words, the acoustic rays converge in a region located close to the source, and our estimation aims to identify such a convergence region out of all the generated rays.

The MCL approach generates initial particles in the space as an approximation to the source locations. It allocates higher weights to particles that are closer to acoustic rays and re-samples the particles to get more particles in regions with higher weights [2]. Specifically, we adopt the generalized variance, which is a one-dimensional measure for multi-dimensional scatter data, to see whether particles have converged. When the generalized variance is less than a threshold (e.g., $\sigma_c = 0.5$), we treat that a sound occurs and the mean position of those particles as the estimated sound source position.

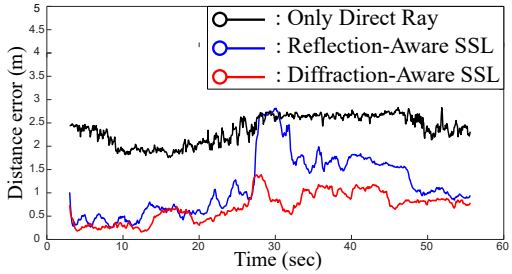
3. Result and Discussion

We describe our setup consisting of a robot with microphones and testing environments, and highlight the performance of our approach. The hardware platform is based on Turtlebot2 with a 2D laser scanner, Kinect, a computer with an Intel i7 process, and a microphone array, which is an embedded system for streaming multi-channel audios, consisting of eight microphones. For all the computations, we use a single core, and perform our estimation every 200ms, supporting five different estimations in one second.

We have evaluated our method in indoor environments containing a box-shaped object that blocks direct paths from the sound to the listener. As shown in Fig. 4a, the scenario contains a moving source and an obstacle. Where the source moves along the red trajectory, the obstacle cause the NLOS source while the source is located inside the invisible area.



(a) A NLOS moving source scene around an obstacle.



(b) Accuracy errors.

Figure 4: These figures show the testing environment (7m by 7m with 3m height) (a) and the accuracy error of our method with the dynamically moving sound source (b). This figure refers to Fig. 1 in [1].

This scenario is tested on the room that size is $7m \times 7m$ and 3m height.

During the precomputation phase, we perform SLAM and reconstruct a mesh of the testing environment. We ensure that the resulting mesh has no holes using the MeshLab package.

Moving sound source around an obstacle. We evaluate the accuracy by computing the L2 distance errors between the positions estimated by our method and the ground-truth positions. We measure the accuracy for the three cased (using only direct rays, adding reflection rays, and adding diffraction rays) to show the efficiency of the proposed method.

The accuracy graph of the scenario is presented in Fig. 4b; the average distance errors of three cases of only using direct rays, adding reflection rays, and adding diffraction rays are 1.83m, 0.95m, and 0.7m, respectively, indicating a 92% and 125% improvement in accuracy considering reflection and diffraction rays. The working videos and more details are available at (http://sgvr.kaist.ac.kr/~ikan/papers/DA-SSL) and (http://sgvr.kaist.ac.kr/~ikan/papers/RA-SSL).

Future work. Even if our work can localize the 3D location of the NLOS source well, it still needs to be supplemented because the accuracy is relatively low (0.7m). As we already mentioned in Sec. 1, there are considerable works that are visual-based localization methods [8, 3, 4]. As a part of future work, we would like to combine our acoustic-drive localization method and novel audio-visual localization methods.

Acknowledgment

This work was supported by the NRF grant funded by the MSIT (No. 2019R1A2C3002833) and the MOTIE under Industrial Technology Innovation Program (No.10067202).

References

[1] I. An, D. Lee, J.-w. Choi, D. Manocha, and S.-e. Yoon. Diffraction-aware sound localization for a non-line-of-sight source. In *2019 IEEE International Conference on Robotics and Automation (ICRA)*. 1, 2, 4

[2] I. An, M. Son, D. Manocha, and S.-e. Yoon. Reflection-aware sound source localization. In *ICRA*, 2018. 1, 2, 3

[3] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 1, 4

[4] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016. 1, 4

[5] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust., Speech, Signal Process.*, 24(4):320–327. 1

[6] R. G. Kouyoumjian and P. H. Pathak. A uniform geometrical theory of diffraction for an edge in a perfectly conducting surface. *November*, 88:1448–1461, 1974. 1, 2, 3

[7] R. Mur-Artal and J. D. Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 1

[8] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 1, 4

[9] N. Tsingos, T. Funkhouser, A. Ngan, and I. Carlbom. Modeling acoustics in virtual environments using the uniform theory of diffraction. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 545–552. ACM, 2001. 3

[10] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger. Elasticfusion: Real-time dense slam and light source estimation. *The International Journal of Robotics Research*, 35(14):1697–1716, 2016. 1