2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Robust Sound Source Localization considering Similarity of Back-Propagation Signals

Inkyu An[1], Byeongho Jo[2], Youngsun Kwon[1], Jung-woo Choi[2], and Sung-eui Yoon[1]
http://sgvr.kaist.ac.kr/~ikan/papers/SSL-BPS/

***Abstract*— We present a novel, robust sound source localization algorithm considering back-propagation signals. Sound propagation paths are estimated by generating direct and reflection acoustic rays based on ray tracing in a backward manner. We then compute the back-propagation signals by designing and using the impulse response of the backward sound propagation based on the acoustic ray paths. For identifying the 3D source position, we use a well-established Monte Carlo localization method. Candidates for a source position are determined by identifying convergence regions of acoustic ray paths. Those candidates are validated by measuring similarities between back-propagation signals, under the assumption that the back-propagation signals of different acoustic ray paths should be similar near the ground-truth sound source position. Thanks to considering similarities of back-propagation signals, our approach can localize a source position with an averaged error of 0.55 m in a room of 7 m by 7 m area with 3 m height in tested environments. We also place additional 67 dB and 77 dB white noise at the background, to test the robustness of our approach. Overall, we observe a 7 % to 100 % improvement in accuracy over the state-of-the-art method.**

I. INTRODUCTION

As robots become more widely available, it is getting more imperative for a robot to understand environments for safe and accurate operations. There have been many kinds of research efforts to perceive the environments by acquiring and using data from hardware sensors. One of the main research topics for understanding the environments focuses on identifying locations of a robot itself and other objects in environments from collected data by vision cameras and depth sensors. Departing from these approaches, an acoustic data measured by acoustic sensors has recently attracted attention as an important clue for localizing various objects.

The problem identifying the location of a sound source from collected acoustic data is widely known as the sound source localization (SSL). There has been a significant amount of efforts to localize a sound source by estimating direction of arrival (DOA) of sound waves. There have been fundamental methods for estimating DOA based on time difference of arrival (TDOA) of a microphone array [1], [2], [3], and the beamforming algorithms are widely used to enhance desired signals at the specific directions from the sound source [4], [5], [6].

Thanks to the advantage of the spherical configuration, many DOA estimation methods focus on using the spherical microphone arrays. Rafaely [7] presented a theoretical framework of spherical harmonic array processing, and the delay-and-sum beamformer is extended to process on the spherical harmonics domain. Many advanced beamforming techniques [8], [9], [10] were proposed by using the minimum variance distortionless response (MVDR) power spectra on the spherical harmonics domain. Li *et al.* [11] presented a MUSIC (Multiple Signal Classification) based DOA estimation algorithm, which uses orthogonality between a noise-only subspace and a signal-plus-noise subspace on the spherical harmonics domain.

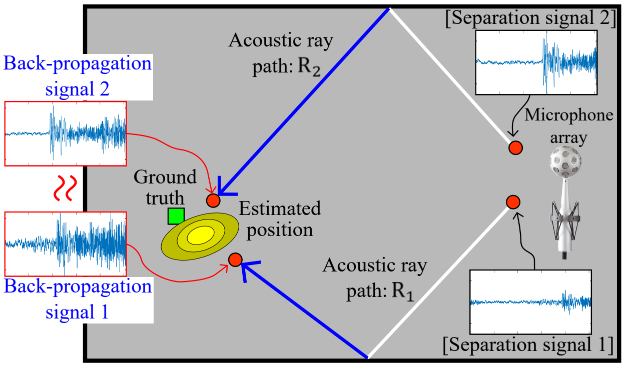


Fig. 1. Our approach generates direct and indirect acoustic ray paths and localizes the sound source while considering back-propagation signals on generated acoustic ray paths. The back-propagation signals are virtually computed signals that could be heard at particular locations and computed by using impulse responses. When two back-propagation signals of acoustic ray paths are highly correlated, we treat them to be originated from the same source.

Unfortunately, these methods were designed for detecting DOA, not the 3D location of a sound source in an arbitrary environment. Especially, when a sound source is occluded by an obstacle, most prior approaches cannot specify the location of the source generating the sound signal.

To address this issue, recent techniques were proposed to find a 3D source location even if the sound source is in the non-line-of-sight state [12], [13]. These techniques estimate sound propagation paths from the source to microphones as acoustic rays, generated by the ray tracing technique, and identify the 3D source location by using generated acoustic rays. However, the accuracy of these methods decreases in environments with background noise and imperfect reconstruction of the 3D environments. This low accuracy is caused mainly because several errors, like background noise and imperfect 3D reconstruction, are accumulated along each acoustic ray for estimating the source location.

[1] I. An, [1] Y. Kwon, and [1] S. Yoon (Corresponding author) are with the School of Computing, KAIST, Daejeon, South Korea; [3] B. Jo and [3] J. Choi is with the School of Electrical Engineering, KAIST; {inkyu.an, byeongho, youngsun.kwon, jwoo}@kaist.ac.kr, sungeui@kaist.edu

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

Main Contributions. To robustly identify the sound source location, we present a novel, sound source localization algorithm using back-propagation signals (Fig. 1). Using a beamforming algorithm, we first compute DOA of the sound wave and separation signals corresponding to those specific DOAs (Sec. II-A). We then estimate sound propagation paths by generating acoustic ray paths in the reverse direction to DOAs of the sound (Sec. II-B), and compute the back-propagation signals using the impulse response of the acoustic ray path from the separation signal (Sec. II-C). Intuitively speaking, back-propagation signals are virtually computed signals that could be heard at a particular location on acoustic paths from the measured signals at the microphone array.

Finally, we use the Monte Carlo localization algorithm estimating a location of the sound as a converging region of computed acoustic ray paths. In particular, we utilize the computed back-propagation signals of different acoustic ray paths for robust estimation of the sound location, under the intuitive assumption that acoustic paths coming from the same sound source should have similar back-propagation signals at the estimated location (Sec. II-D).

II. Sound source localization using back-propagated signals

Our work is built upon ray tracing-based sound source localization (SSL) [13]. In a real environment involving moving sound sources, obstacles, or noise, acoustic rays generated naively by the prior ray tracing based SSL may converge to a position other than the actual location of the sound source.

To solve this problem, we aim to generate and utilize back-propagation signals to a candidate 3D location along each acoustic ray. This back-propagation signals at a location can be computed by simulating the reverse process of sound propagation, i.e., by reversely performing ray tracing.

A. Beamforming

To generate acoustic rays, we estimate DOAs of the sound waves at the spherical microphone array using a EB-MVDR (Eigenbeam-minimum variance distortionless response) beamformer [8], [9], [10]. Note that our input signals are measured by almost uniformly sampled microphones on a rigid sphere (32 channel microphone positions), but each microphone signal is, in fact, a mixture of signals from different directions. We therefore aim to extract signals from different DOAs, and for this purpose, the EB-MVDR beamformer is utilized.

The array signal, $\mathbf{x} = [x_1(k), \cdots, x_Q(k)]^T$, measured by Q microphones of the spherical array consists of sound pressure signals $\mathbf{p} = [p_1(k), \cdots, p_Q(k)]^T$ and noise signals, $\mathbf{n} = [n_1(k), \cdots, n_Q(k)]^T$:

$$\mathbf{x} = \mathbf{p} + \mathbf{n}, \tag{1}$$

where $k = 2\pi f/c$ is the wavenumber determined by the frequency f and speed of sound c. Note that the measured sound signal $\mathbf{p}$ is the consequence of sound propagation and reflections through a direct or indirect propagation path. We apply the spherical Fourier transform (SFT) to the array signal $\mathbf{x}$ [7], which yields the spherical harmonic (SH) coefficients $\mathbf{x}_{\nu\mu}$ defined over different orders ν and degrees μ of spherical harmonics. For the SH coefficients measured up to the order $\nu = \nu'$, there are $(\nu' + 1)^2$ coefficients in total. Since the SFT is a linear operation, we also have the following relation:

$$\mathbf{x}_{\nu\mu} = \mathbf{p}_{\nu\mu} + \mathbf{n}_{\nu\mu}, \tag{2}$$

Our objective is to identify DOAs and extract the sound signal coming from each DOA. The beamformer does this by multiplying a beamformer weight vector $\mathbf{w}_{\nu\mu}(\Omega)$ defined for a specific pair of zenith and azimuth angle $\Omega = (\theta, \phi)$ to the measured signal $\mathbf{x}_{\nu\mu}$. The output of the beamformer S, therefore, can be written as the inner product of $\mathbf{w}_{\nu\mu}(\Omega)$ and $\mathbf{x}_{\nu\mu}$:

$$S(\Omega) = \mathbf{w}_{\nu\mu}(\Omega)^H \mathbf{x}_{\nu\mu}, \tag{3}$$

where $(\cdot)^H$ is the Hermitian transpose. Among many beamformers, we adopt the EB-MVDR that is known to provide a good spatial resolution and signal separation performance. With the EB-MVDR beamformer, DOAs are estimated from the beamforming power defined as:

$$\beta_{MV}(\Omega) = \frac{1}{\mathbf{v}_{\nu\mu}(\Omega)^H \mathbf{R}^{-1}_{\mathbf{x}_{\nu\mu}\mathbf{x}_{\nu\mu}} \mathbf{v}_{\nu\mu}(\Omega)}, \tag{4}$$

where $\mathbf{R}_{\mathbf{x}_{\nu\mu}\mathbf{x}_{\nu\mu}}$ is the covariance matrix of which elements are cross-spectral densities of measured signals $\mathbf{x}_{\nu\mu}$, and $\mathbf{v}_{\nu\mu}(\Omega)$ denotes a steering vector given by the wave propagation model. In this work, we use the plane wave model to define the steering vector $\mathbf{v}_{\nu\mu}$. Fig. 2 shows the beamforming power calculated for every direction Ω; all directions correspond to 10242 grids on the unit sphere that is based on the recursive subdivision of an icosahedron [14].

Local maxima of the beamforming power can represent the direct and indirect DOAs of the propagation paths. That is

$$[\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_N] = F_{max}\{\beta_{MV}(\Omega)\}, \tag{5}$$

where $\mathbf{d}_n = (\cos\phi_n \sin\theta_n, \sin\phi_n \sin\theta_n, \cos\theta_n)$ denotes a directional vector of the n-th local maximum of the beamforming power among N different local maxima in a frame, and $F_{max}\{\cdot\}$ is a function for finding local maxima of the beam energy function. In practice, we identify top-four local maxima on average in our tested experiments.

We then extract sound signals, called the separated signal S_n, coming from a specific direction Ω_n with the directional vector $\mathbf{d}_n$. The beamformer weight $\mathbf{w}_{\nu\mu}(\Omega_n)$ of the EB-MVDR beamformer is given by:

$$\mathbf{w}_{\nu\mu}(\Omega_n) = \frac{\mathbf{v}_{\nu\mu}(\Omega_n)^H \mathbf{R}^{-1}_{\mathbf{x}_{\nu\mu}\mathbf{x}_{\nu\mu}}}{\mathbf{v}_{\nu\mu}(\Omega_n)^H \mathbf{R}^{-1}_{\mathbf{x}_{\nu\mu}\mathbf{x}_{\nu\mu}} \mathbf{v}_{\nu\mu}(\Omega_n)}, \tag{6}$$

which minimizes the total beamforming power while satisfying the distortionless-response constraint to the looking direction Ω_n $\left(\mathbf{w}_{\nu\mu}(\Omega_n)^H \mathbf{v}_{\nu\mu}(\Omega_n) = 1\right)$. This beamformer weight is used in Eq. 3 for computing four separated signals S_n.

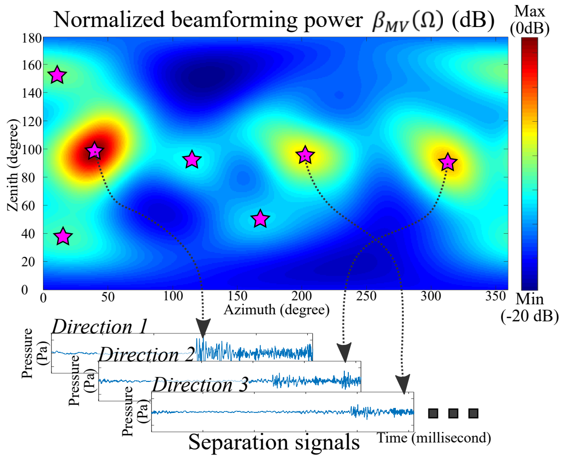


Fig. 2. A beamforming power is computed by a beamforming algorithm, where the horizontal axis is the azimuth angle and the vertical axis is the zenith angle of the unit sphere. Local maxima of the beamforming power are treated most significant directions of arrival (DOAs) of sound. The sound signal impinging from each DOA is extracted by applying the EB-MVDR beamformer to the signals measured by microphones.

The separated signals are then back-propagated to the directions $\mathbf{d}_n$ by reconstructing acoustic rays to the true source positions.

B. Acoustic ray tracing

We explain how to generate acoustic rays from estimated directions $[\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_N]$ that are the reverse directions of incoming sounds. We want to estimate propagated paths (e.g., direct and reflection paths) of the sound from its source location to the microphone array location using the acoustic rays. We generate such acoustic rays considering direct and reflection paths based on the RA-SSL algorithm [13].

Unlike the prior work of RA-SSL, we use a mesh representation of the surroundings captured from sensors. We construct the mesh that is robust to minor noise, and use it for acoustic interactions between the surroundings and generated acoustic rays. Starting from the point cloud collected by the depth sensor, i.e., Velodyne VLP-16, we apply the voxelization in order to reduce the sensor noise, and then reconstruct the environment in the form of a mesh map from the voxelized point cloud using the Poisson surface reconstruction algorithm [15].

For the n-th acoustic ray path, denoted by R_n, its primary acoustic ray, r_n^0, is created into the n-th direction vector $\mathbf{d}_n$, as shown in Fig. 3. If the acoustic ray collides with an obstacle, its secondary, reflection ray is generated by assuming the specular reflection, and is denoted by r_n^1, where the superscript represents the order of the acoustic ray path; refer to [13] for the detailed process on ray generation. When R_n is propagated until a $(D-1)$-th order, it indicates that the acoustic ray path R_n consists of D acoustic rays: i.e., $R_n = [r_n^0, r_n^1, \cdots, r_n^{D-1}]$.

C. Back-propagation signals

We introduce how to compute back-propagation signals based on the generated acoustic ray paths $[R_1, R_2, \cdots, R_N]$ and separated signals $[S_1, S_2, \cdots, S_N]$; there is a tuple of (R_n, S_n) for the reverse direction vector $\mathbf{d}_n$ of the n-th incoming sound. We want to compute the back-propagation signal P_n from the separated signal S_n by designing and using an impulse response of backward sound propagation based on the acoustic ray path R_n. The impulse response describes the reaction of any linear system as a function of time-independent variables; the input is the separated signal and the output is the back-propagation signal in our approach.

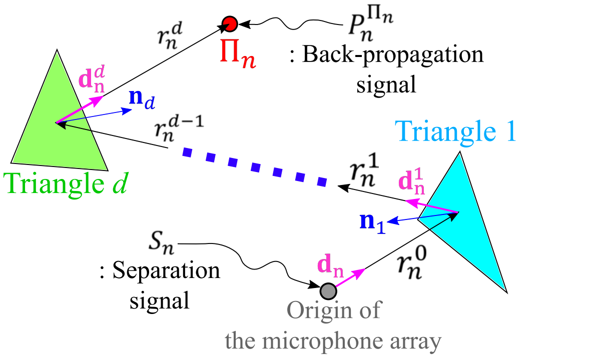


Fig. 3. An example of generating an acoustic ray path R_n and its back-propagation signal. The primary acoustic ray, r_n^0, of the n-th acoustic ray path R_n is generated to the direction vector $\mathbf{d}_n$ that is the reverse direction of the n-th incoming sound. When the acoustic ray r_n^0 hits an obstacle represented by Triangle 1, its reflection acoustic ray r_n^1 is generated according to the specular reflection based on the normal vector $\mathbf{n}_1$ of Triangle 1. The back-propagation signal P_n is computed by using the impulse response of R_n at a specific point, Π_n, on the path from the separated signal S_n.

In this work, we utilize the impulse response for the backward propagation to improve the accuracy of the sound source localization. In forward sound propagations [16], [17], [18], [19], the impulse response of an acoustic ray path is described by sound attenuations according to the travel distance of a ray path and reflection. For example, the travel distance attenuation represents the decrease of sound pressure inversely proportional to the travel distance of the ray path, because the sound is propagated according to the spherical wave in 3D environments; similar for the reflection attenuation.

On the other hand, for the backward propagation problem, the attenuation of travel distance and reflection becomes an amplification of the sound pressure. Suppose that we aim to compute the back-propagation signal from the starting point to a specific point, Π_n (Fig. 3), on an acoustic ray path using the backward impulse response, where there is the n-th tuple (R_n, S_n) and the acoustic ray path R_n consists of D acoustic rays $[r_n^0, \cdots, r_n^{D-1}]$; r_n^0 is a primary ray and r_n^d is the d-th reflection ray $(1 \le d \le D-1)$. In the frequency domain, the backward impulse response, $H_n^{\Pi_n}$, is described by amplifications because of the travel distance l and the reflection until the d-th order reflection ray r_n^d:

$$H_n^{\Pi_n}[k] = \exp\left(\frac{\mathbf{i}kl}{c}\right) \cdot A^T[l] \cdot A^R[R_n, d, k], \tag{7}$$

where the term inside the exponential function is for shifting the back-propagation signal to the time delay of the sound propagation at the specific point Π_n and $\mathbf{i}$ is the imaginary unit. A^T is a coefficient of the travel distance

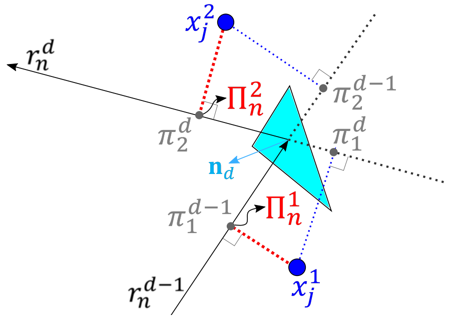


Fig. 4. Examples of determining the point of the acoustic ray path for computing the back-propagation signal. For the particle of x_j^2, the perpendicular foots π_2^d on all d-th order acoustic rays of the n-th acoustic ray path are computed. We then decide the representative perpendicular foot Π_n^2 satisfying the shortest distance from x_j^2 to R_n.

amplification, and is defined by a function of the travel distance l: $A^T[l] = 4\pi(1+l)$. Also, A^R is a coefficient of the reflection amplification, and is defined by considering specular reflections until the d-th order reflection ray:

$$A^R[R_n,d,k] = \prod_{\delta=1}^{d}\left[\frac{1}{\Gamma_\delta[k]}\right], \tag{8}$$

where Γ_δ denotes the reflectivity (reflection coefficient) of the triangle hit by the $(\delta-1)$-th order ray; the reflection coefficient is a function of wavenumber k and we refer to coefficient values reported by [20].

The back-propagation signal $P_n^{\Pi_n}$ at the specific point Π_n on the acoustic ray path R_n is finally computed by the product of the backward impulse response $H_n^{\Pi_n}$ and the separated signal S_n in the frequency domain:

$$P_n^{\Pi_n}[k] = S_n[k]\cdot H_n^{\Pi_n}[k]. \tag{9}$$

D. Estimating a source position

Our estimation process of localizing the sound source is based on the Monte Carlo (MC) localization method. The MC sound source localization identifying the convergence region of acoustic ray paths was suggested in the prior work (RA-SSL) [13]; the convergence region means the area where acoustic ray paths gather.

However, the accuracy of MC localization can decrease in real environments. When there are background noises of sound or complex scene configurations causing uncertainty of the reconstructed environment, they can trigger to generate many arbitrary or incoherent acoustic ray paths. By considering back-propagation signals, we aim to identify those arbitrary and incoherent acoustic ray paths and cull away acoustic ray paths with different back-propagation signals indicating that they are from different sound sources. Intuitively speaking, if there are two acoustic ray paths caused by the same source, their back-propagation signals should be similar near the location of their sound source. In other words, when back-propagation signals of two acoustic ray paths are different at a location, the location is unlikely to be a candidate for a converging region of the sound source.

The MC localization consists of three parts: sampling, computing a weight of particles, and resampling. The main differentiation of our approach over the prior technique is that our method improves the localization accuracy based on a novel module for computing weights of particles based on our back-propagation signals.

Suppose there are i-th particles, x_j^i, representing hypothetical locations of the sound source at a j frame. We compute how close the particle is to acoustic ray paths. For this, we define a specific point Π_n^i, which is decided to be the point satisfying the shortest distance between x_j^i and any point on the n-th acoustic ray path; i.e., $\Pi_n^i = \mathrm{argmin}_{\pi_i^d}||x_j^i - \pi_i^d||$, where π_i^d is the perpendicular foot on the d-th order acoustic ray from the x_j^i position (Fig. 4). We then compute our back-propagation signal according to Eq. 7 at the shortest point Π_n^i on the n-th acoustic ray path from the particle x_j^i.

From the back-propagation signal $P_n^{\Pi_n^i}[k]$ in the frequency domain, we compute the back-propagation signal $p_n^{\Pi_n^i}[t]$ in the time domain signal. We then calculate a particle weight, w_j^i, representing the probability of being a convergence region of the sound source, based on two factors: a distance weight, w_d, representing how away the particle is from the n-th acoustic ray path and a similarity weight, w_s, indicating how similar between $p_n^{\Pi_n^i}[t]$ and other signals given acoustic ray paths:

$$w_j^i = P(O_j|x_j^i) = \frac{1}{n_c}\sum_{n=1}^{N_j}[w_d(x_j^i,R_n)\cdot w_s(x_j^i,R_n)], \tag{10}$$

where N_j is the number of acoustic ray paths at the j frame, O_j is the observation containing $[P_1^{\Pi_1^i},\cdots,P_{N_j}^{\Pi_{N_j}^i}]$ and $[R_1,\cdots,R_{N_j}]$, and n_c is a normalizing constant.

The distance weight w_d is calculated by using the Euclidean distance between the particle location x_j^i and the point Π_n^i:

$$w_d(x_j^i,R_n) = G(||x_j^i - \Pi_n^i||\,|0,\sigma_w), \tag{11}$$

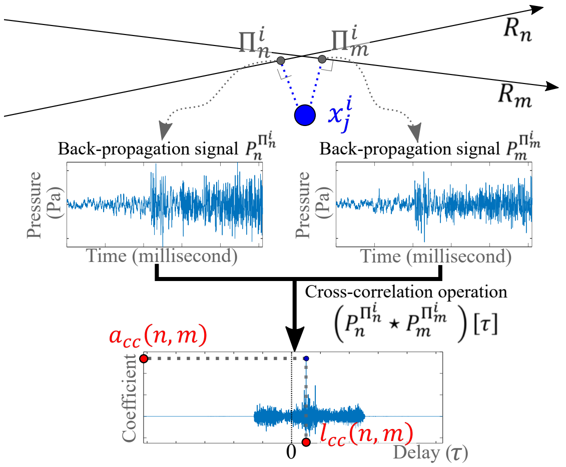
where G is the Gaussian distribution function with the zero mean and a standard deviation σ_w. w_d is maximized when the particle x_j^i is on the perpendicular foot Π_n^i. The similarity weight $w_s(x_j^i,R_n)$ measures the similarity between the back-propagation signal $p_n^{\Pi_n^i}$ from the n-th acoustic ray path and ones of other acoustic ray paths:

$$\frac{1}{n_s}\sum_{m=1,m\neq n}^{N_j}\begin{cases}\frac{L-(1-\alpha)\cdot l_{cc}(n,m)}{L}, & \text{if } a_{cc}(n,m) > a_{th}\\ 0, & \text{otherwise,}\end{cases} \tag{12}$$

where n_s is the normalizing constant, L is the length of the back-propagation signal, $a_{cc}(\cdot)$ is the peak coefficient in a normalized range of -1 to 1, $l_{cc}(\cdot)$ is the peak coefficient delay, α denotes a parameter for adjusting the similarity weight, and a_{th} denotes the threshold value of $a_{cc}(\cdot)$. Both variables of $a_{cc}(\cdot)$ and $l_{cc}(\cdot)$ are computed by applying the cross-correlation operation between n-th and m-th signals:

$$\begin{aligned} a_{cc}(n,m) &= \max\{(p_n^{\Pi_n^i}\star p_m^{\Pi_m^i})[\tau]\},\\ l_{cc}(n,m) &= \mathrm{argmax}_\tau\{(p_n^{\Pi_n^i}\star p_m^{\Pi_m^i})[\tau]\},\end{aligned} \tag{13}$$

where $\star$ is the cross-correlation operator.



Π_n^i
Π_m^i
R_n
R_m
x_j^i
Back-propagation signal $P_n^{\Pi_n^i}$
Back-propagation signal $P_m^{\Pi_m^i}$
Pressure (Pa)
Time (millisecond)
Pressure (Pa)
Time (millisecond)
Cross-correlation operation
$\left(P_n^{\Pi_n^i} \star P_m^{\Pi_m^i}\right)[\tau]$
$a_{cc}(n,m)$
Coefficient
$l_{cc}(n,m)$
0
Delay (τ)

Fig. 5. An example of computing the peak coefficient a_{cc} and the peak coefficient delay l_{cc} by using the cross-correlation operation. Given two back-propagation signals, $p_n^{\Pi_n^i}$ and $p_m^{\Pi_m^i}$ at Π_n^i and Π_m^i, respectively, we perform the cross-correlation operation between two signals. The maximum coefficient becomes the peak coefficient a_{cc} and the time delay from the time origin, 0, to the time realizing the maximum coefficient becomes the peak coefficient delay l_{cc}.

As shown in Fig. 5, $a_{cc}(\cdot)$ represents how much both back-propagation signals are correlated, and l_{cc} shows the time difference of occurrence between both back-propagation signals. As both back-propagation singles are from the same sound source, ideally a_{cc} and l_{cc} become one and zero, respectively.

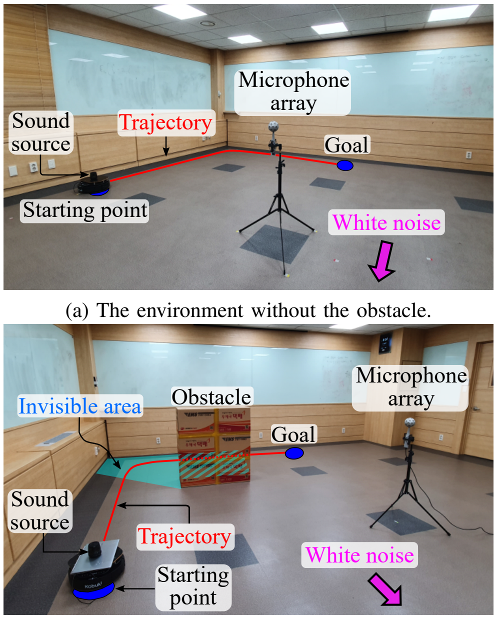
Getting back to Eq. 12, we treat that two back-propagation signals are similar, when their peak coefficient is bigger than the threshold, i.e., $a_{cc} > a_{th}$. In this case, we assign a higher weight according to the relative time delay of the length of the signal, $(\frac{L-(1-\alpha)\cdot l_{cc}}{L})$ that becomes a value in a range of α to 1; i.e., we give the highest weight when two signals are matched without any delay, under the assumption that those two signals are originated from the same sound source. If there is no back-propagation signal satisfying the condition, $a_{cc} > a_{th}$, the signal similarity weight w_s has a constant value α that is the smallest value of $(\frac{L-(1-\alpha)\cdot l_{cc}}{L})$.

III. Results and Discussion

The yellow disk in Fig. 1 represents a 95% confidence area for the sound source location estimated by our method. We also compare distance errors of our approach to the prior work (RA-SSL) that does not use the similarity of back-propagation signals, to demonstrate the effectiveness of considering the back-propagation signals.

The hardware platform consists of Eigenmike, the 32-channel microphone array of the mh acoustics, and the i7 CPU computer. As mentioned in Sec. II-B, we use Velodyne VLP-16 and build a mesh map as the reconstruction of tested indoor environments. The reflection coefficients are appropriately assigned to the triangles by referring to the reported values in [20].

We report values of parameters used for our algorithm: α for controlling the influence of each weight is 0.5, the standard deviation σ_w of the Gaussian distribution function used for computing the distance weight is 0.5 that is determined by the consideration of the size of the indoor environment (about one-tenth of the room width 7m), and the threshold value a_{th} for checking the correlation between back-propagation signals is 0.15.



(a) The environment without the obstacle.

(b) The environment with the obstacle.

Fig. 6. The test environments w/ and w/o an obstacle that can make the sound source non-line-of-sight one. We use the clapping sound in the sound source. We put an additional noise (67 dB and 77 dB white noises) as the distractor in the the back of the test environments.

We use 1024 samples for the separation signal, where the sampling frequency is 12 kHz; 1024 audio samples (85 ms) are a sufficient length for covering direct and first-bounce reflection signals as indicated in [21]. We set our algorithm to estimate the source position every 256 ms in order to respond appropriately to the movement of the source. Specifically, beamforming and generating acoustic rays take 50ms and 0.54ms respectively on average, which are less than the audio length (85ms), and estimating the source position based on the particle filter takes 200 ms on an average that is less than the iteration period 256 ms.

A. Benchmarks

Different experiments were conducted in two scenes: the moving sound without and with an obstacle. In both environments (Fig. 6a and Fig. 6b), a robot equipped with an omni-directional speaker moved along the red trajectory, and the 32-channel microphone array recorded the audio signals, and these data are used for various tests with the ground truth information on the sound source locations. In Fig. 6b, we put an obstacle made by paper boxes, to cause the robot invisible along the robot's trajectory for the microphone array; at the invisible area, the sound source becomes the non-line-of-sight (NLOS) source.

Handling the NLOS source was reported a quite difficult problem in prior methods [13], because direct sound propagation paths are blocked by the obstacle and we have to rely

TABLE I

THE AVERAGE DISTANCE ERRORS W/ DIFFERENT NOISE LEVELS. NUMBERS IN THE PARENTHESES SHOW THE IMPROVEMENT.

An moving source w/o an obstacle	w/o white noise	67 dB white noise	77 dB white noise
SNR	20.64 dB	15.74 dB	9.35 dB
Our approach	0.57m (7%)	0.58m (18%)	0.56m (38%)
RA-SSL	0.61m	0.69m	0.78m
An moving source w/ an obstacle	w/o white noise	67 dB white noise	77 dB white noise
SNR	20.83 dB	17.33 dB	9.65 dB
Our approach	0.51m (64%)	0.54m (75%)	0.53m (100%)
RA-SSL	0.84m	0.95m	1.08m

on indirect sound paths that are incoherent and sensitive to noise. Furthermore, the number of indirect acoustic ray paths passing near the ground truth is usually small, and thus the accuracy of the localization algorithm tends to deteriorate.

Additionally, these scenes are not free from noise (e.g., various noise from outside the room and moving sound of the tested sound source), naturally occurring in a typical environment where the signal-to-noise ratios (SNRs) of both scenes containing the moving sound without and with an obstacle are 20.64 dB and 20.83 dB. To further test the robustness of the proposed method, we expose these scenes additional white noise, whose average sound pressure levels are 67 dB and 77 dB. These noises can cause to trigger many incoherent acoustic ray paths, hindering them to converge in a single location.

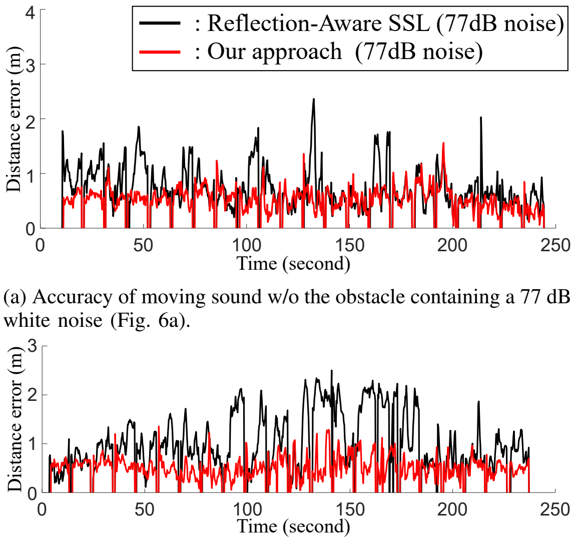
B. A moving sound source

We first show how our approach has the advantage compared to the prior method in a simple scene with a moving sound. In Table I, the accuracy of RA-SSL in the moving source scene gradually deteriorates, as the power of noises increase, where the SNRs containing 67 dB and 77 dB noises are 15.74 dB and 9.35 dB, respectively. On the other hand, the accuracy of our work is rather robust with different power of noise. This shows that our method is robust even in noisy environments, thanks to considering the back-propagation signals on estimated source locations; the similarity weight improves the robustness of the source localization algorithm. To show the positive effect of back-propagation signals on the 3D sound source localization, we append a description on coherence among back-propagation signals compared to separation signals on the video submission.

Fig. 7a shows the distance errors of RA-SSL and our approach, where there is 77 dB white noise. The average distance errors are 0.7839 m for RA-SSL and 0.5678 m for our approach; the accuracy of the sound source localization is improved about 38% based on our approach.

C. A moving sound around an obstacle

We now show results with the more challenging environment including an obstacle between the source trajectory and the microphone array (Fig. 6b). Fig. 7b shows graphs of the distance errors of RA-SSL and our approach with the 77 dB white noise; SNR in this scene is 9.65 dB. The average distance errors of RA-SSL and our approach are 1.083 m and 0.5364 m, respectively. Especially, where the sound source is in the NLOS state from 90 to 180 seconds, the accuracy of RA-SSL decreases drastically, because blocking the direct sound propagation paths makes the convergence of acoustic rays weak near the ground truth. On the other hand, even in this challenging case, we get a stable result, 100% improvement compared to RA-SSL, by considering the similarity between back-propagation signals of indirect acoustic paths.



(a) Accuracy of moving sound w/o the obstacle containing a 77 dB white noise (Fig. 6a).

(b) Accuracy of moving sound w/ the obstacle containing a 77 dB white noise (Fig. 6b).

Fig. 7. The distance errors between the ground truth and the estimated source positions. In this scene, there is the additional 77 dB white noise, on top of natural occurring noise.

As we have stronger white noise (Table I), SNRs in the moving source scene w/ the obstacle decrease, which are 20.83 dB, 17.33 dB, and 9.65 dB, and the accuracy of RA-SSL then dramatically deteriorates. However, the accuracy of our approach is stable even with different noise energy, demonstrating the robustness and usefulness of our approach.

IV. LIMITATIONS AND FUTURE DIRECTIONS

While we have demonstrated benefits of our approach, it has several limitations and opens up many interesting future directions. When the white noise is larger than 87 dB (sound level like a truck noise [22]), we found that our approach did not work properly because of the relatively weak energy of the sound source (77.34 dB). Other sound propagation phenomena such as scattering and diffraction that are frequently observed at low frequencies are not handled yet. The acoustic material properties such as reflection coefficients of triangles of objects are not automatically assigned, and recent deep learning approaches showing promising results can be employed to solve this problem [20].

ACKNOWLEDGMENT

This research was supported by the SW StartLab program (IITP-2015-0-00199), NRF grant funded by MSIT (No. 2019R1A2C3002833), and MOTIE (No.10067202).

References

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay", *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327.

[2] J-M Valin, François Michaud, Jean Rouat, and Dominic Létourneau, "Robust sound source localization using a microphone array on a mobile robot", in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*. IEEE, 2003, vol. 2, pp. 1228–1233.

[3] Charles Blandin, Alexey Ozerov, and Emmanuel Vincent, "Multi-source tdoa estimation in reverberant audio using angular spectra and clustering", *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[4] J-M 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 *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*. IEEE, 2004, vol. 1, pp. 1033–1038.

[5] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering", *Robot. Auton. Syst.*, vol. 55, no. 3.

[6] Cha Zhang, Dinei Florêncio, Demba E Ba, and Zhengyou Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings", *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[7] Boaz Rafaely, *Fundamentals of spherical array processing*, vol. 8, Springer, 2015.

[8] Haohai Sun, Edwin Mabande, Konrad Kowalczyk, and Walter Kellermann, "Joint doa and tdoa estimation for 3d localization of reflective surfaces using eigenbeam mvdr and spherical microphone arrays", in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 113–116.

[9] Daniel P Jarrett, Emanuël AP Habets, and Patrick A Naylor, "Spherical harmonic domain noise reduction using an mvdr beamformer and doa-based second-order statistics estimation", in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 654–658.

[10] Shefeng Yan, Haohai Sun, U Peter Svensson, Xiaochuan Ma, and Jens M Hovem, "Optimal modal beamforming for spherical microphone arrays", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 361–371, 2011.

[11] Xuan Li, Shefeng Yan, Xiaochuan Ma, and Chaohuan Hou, "Spherical harmonics music versus conventional music", *Applied Acoustics*, vol. 72, no. 9, pp. 646–652, 2011.

[12] I. An, D. Lee, J. Choi, D. Manocha, and S. Yoon, "Diffraction-aware sound localization for a non-line-of-sight source", in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 4061–4067.

[13] I. An, M. Son, D. Manocha, and S. Yoon, "Reflection-aware sound source localization", in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 66–73.

[14] F. Gyorgy, "Rendering and managing spherical data with sphere quadtrees", in *Proceedings of the First IEEE Conference on Visualization: Visualization '90*, Oct 1990, pp. 176–186.

[15] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe, "Poisson surface reconstruction", in *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006, vol. 7.

[16] Chunxiao Cao, Zhong Ren, Carl Schissler, Dinesh Manocha, and Kun Zhou, "Interactive sound propagation with bidirectional path tracing", *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 180, 2016.

[17] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng, "Scene-aware audio for 360 videos", *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 111, 2018.

[18] Hengchin Yeh, Ravish Mehra, Zhimin Ren, Lakulish Antani, Dinesh Manocha, and Ming Lin, "Wave-ray coupling for interactive sound propagation in large complex scenes", *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 165, 2013.

[19] Carl Schissler, Ravish Mehra, and Dinesh Manocha, "High-order diffraction and diffuse reflections for interactive sound propagation in large environments", *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 39, 2014.

[20] Carl Schissler, Christian Loftin, and Dinesh Manocha, "Acoustic classification and optimization for multi-modal rendering of real-world scenes", *IEEE transactions on visualization and computer graphics*, vol. 24, no. 3, pp. 1246–1259, 2018.

[21] Jingdong Chen and Jacob Benesty, "A time-domain widely linear mvdr filter for binaural noise reduction", in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*. IEEE, 2011, pp. 105–108.

[22] Yang-Hann Kim and Jung-Woo Choi, *Sound visualization and manipulation*, John Wiley & Sons, 2013.