Moving Sound Source Tracking in Wide Space by Multiple Microphone Arrays

Toru Takahashi* and Takuma Ekawa† and Masato Nakayama‡
* Osaka Sangyo University, Graduate School of Engineering, Osaka, Japan
Email: takahashi@ise.osaka-sandai.ac.jp Tel: +81-752-3001
† Osaka Sangyo University, Graduate School of Engineering, Osaka, Japan
Email: s20mh01@ge.osaka-sandai.ac.jp Tel: +81-752-3001
‡ Osaka Sangyo University, Graduate School of Engineering, Osaka, Japan
Email: nakayama@ise.osaka-sandai.ac.jp Tel: +81-752-3001

***Abstract*—This paper describes about a method for tracking a moving sound source based on multiple microphone arrays. A moving sound source tracking over a few meters, e.g. over ten meters, is investigated. A key idea of tracking is using multiple microphone arrays. When a sound source moving over five meters is tracked by a single microphone array, a sound source localization accuracy decreases as a distance between a microphone array and a sound source increases. To solve this problem, we propose to distribute multiple microphone arrays at intervals of several meters and selectively use two microphone arrays. A proposed method localizes a moving sound by using microphone arrays, of which positions are the closest and the second closest to the sound source. As a sound source moves and changes position, we have to re-select two microphone arrays. We also develop an automatic switching method. We evaluate tracking accuracy of a proposed method in a simulation and in a real environment. A sound source tracking experiment using three microphone arrays of which spatial intervals are four meters is conducted in a real environment. It is demonstrated that a proposed method shows high accuracy.**

I. Introduction

Speech information processing often handles distant-talking. For example, the system must handle distant-talking in order to capture the audio of someone talking while walking down the hallway. Lecouteus et al. [1], handled distant-talking in thirty square meters for a smart home application. However, they only handled the sounds coming from less than a few meters, such as less than three meters. Ono et al. [2] proposed a method to synchronize the signals observed by distributed microphones. The distributed microphone technique can also be applied to expand the sound collection space. However, their method cannot be applied to moving sound source. Miura et al. [3] proposed a method for tracking moving sound sources and simultaneously resolving synchronization between microphone channels by surrounding the sound source area with a large number of microphones. However, unlike our method of sparse microphone placement, their method requires covering an area with a very large number of microphones. In contrast to the above work, we deal with source tracking that can be observed up to eight meters away using multiple sparsely distributed microphone arrays.

Our research topic is a sound source tracking method that moves over a wide range that can only be tracked by multiple microphone arrays. A conventional research using multiple microphone arrays does not consider a moving sound sources [4], [5], and there is a limited amount of research dealing with moving sound in speech information processing [6], [7], [8], [9], [10], [11]. One of the reason is that the difficulty of generating a moving sound source with high reproducibility. It is hard to ensure reproducibility unless the playback samples of the sound source waveform at a certain time and the sound source coordinates are recorded in time synchronization. In some studies, the sound source position is not recorded in strict time synchronization, and the objective evaluation of sound source tracking accuracy is insufficient. Some studies use annotated ground truth source trajectories but someone annotates subjectively based on 30 fps CCD images [9], [10], [11]. It is too coarse to evaluate moving sound source localization. Furthermore, there is a problem that a driving sound, that moves a target sound source, interferes with it. To avoid these problem in this paper, a loud speaker is moved by hand. This moving method does not make noise. In addition, the loud speaker position is accurately tracked by LiDAR (Intel RealSense L515). This LiDAR is possible to measure distance to the speaker at 30 fps. Our preliminary test result shows that average error is 0.03 m.

The purpose of this research is to realize the tracking of sound sources that move in wide range space, which is very long moving sound source, with multiple microphone arrays. This paper is organized as follows. First a method of synthesizing a moving sound source is described in Section II. In Section III, we show that GCC-PHAT is an optimal method for localizing a moving sound souce. In Section IV, performance is evaluated in a computational simulation and we demonstrate that the proposed method works in a real environment. Our work is concluded in Section V.

II. Synthesizing moving sound source

In this section, we describe about a method of synthesizing a moving sound source P for our computinal simulation. We assume that a sound source moves in 2D space, but the same thing holds true in 3D space. A sound source located on $\tilde{\boldsymbol{x}}(t) = [\tilde{x}_1(t), \tilde{x}_2(t)]^T$ at time t is shown as $P(\tilde{\boldsymbol{x}}(t))$. ^{T} denotes the transpose of the vector. A sound source position

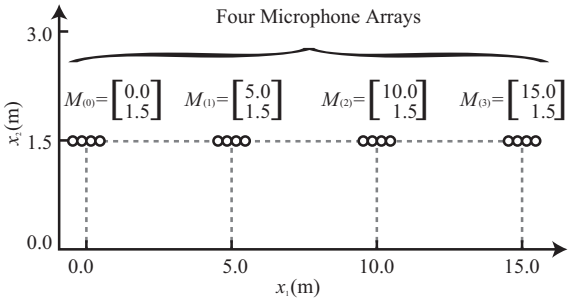

Fig. 1. Position of micrhophone arrays

is non-continuously updated since continuous updating causes hard to synthesize a moving sound source. A discrete version of a moving sound source position is defined as follows:

$$\boldsymbol{x}(n) = \tilde{\boldsymbol{x}}(\lfloor t/\delta + 0.5 \rfloor), \tag{1}$$

where n is time index and an update interval is δ. $\lfloor \cdot \rfloor$ shows the rounding down to the nearest integer. This representation enables us to handle moving sound sources by overlap-and-add method[12]. We can synthesize a moving sound source at every δ steps by convoluting dry source and a room impulse response.

Figure 1 shows four rmicrophone arrays which are lined up at 5 m intervals. $M_{(m)}(\boldsymbol{x})$ shows a m-th microhpone array $(m = 0, 1, 2, \ldots, M-1)$. M is a number of micrphone arrays and $\boldsymbol{x} = [x_1, x_2]^T$ is the center position of the microphone array. The room impulse responses between a sound source and microphones change with a movement of the sound source. To generate a moving sound source, there is a problem of how to convolute the variable impulse responses into a dry sound source. To Solve the problem, we convolve the time variable impulse response using a technique similar to the source filter model, i.e., overlap-and-add method[12] which are well known in the speech synthesis field. Each sample of a moving sound source can be generated by convolution processing for each segment of the dry source.

For accurately synthesizing a moving sound signal, we have to decide a period of sampling sound source position δ and a period of a frame over-lap-add attentively. In analyzing with a frame length N, a frame shift length $N/2$ and synthesizeing with a frame over-lap-add at $N/2$, the moving sound source can be synthesized under the condition of $\delta = N/2$. This means that we can change a room impulse response every δ. When sampling rate is $f_s = 16000$ Hz and $N = 512$, δ is 16 ms. Under this condition, we can reconstruct a time variant impulse response at 60 fps. This frame rate is enough to follow changing impulse responses caused by walking. The shorter δ, the more precisely the time variation of the impulse response can be expressed. However, due to the fact that N is shortened in proportion to δ, the problem arises that the impulse response length that can be convoluved is also shortened.

We assume that the change of impulse response corresponding to the movement is enough to be small and smooth. If a speed of moving sound source is less than 1.4 m/s (about walking speed), the moving distance of the sound source is up to 0.024 m in $\delta = 0.016$ s.

From the above, a moving sound source can be generated from given impulse responses. In this paper, we use a following impulse response model which is often used by other researchers (e.g. [4]). $h_{(n,m)}$ shows an impulse response between a sound source and m-th microphne array. $H_{(n,m)}(\omega)$ shows a frequency domain representation of $h_{(n,m)}$.

$$H_{(n,m)}(\omega) = g \exp\{-j\omega\tau_{(n,m)}\}, \tag{2}$$

$$\tau_{(n,m)} = l_{(n,m)}/c, \tag{3}$$

$$l_{(n,m)} = |M_{(m)} - P_{(n)}|^2, \tag{4}$$

$$g = 10^{\frac{K \cdot l_{(n,m)}}{20}}, \tag{5}$$

where $l_{(n,m)}$ is the distance between $M_{(m)}$ and $P_{(n)}$, $\tau_{(n,m)}$ is the travel time between $P_{(n)}$ and $M_{(m)}$, c is the speed of sound, and K is the decay rate of sound level.

III. Localization of moving sound source

A lower latency method is better suited for localizing a moving sound source. Furthermore a method that requires a fewer number of microphones is better for constructing multiple microphone arrays. Although there are many sound source localization is proposed such as GCC-PHAT[13], SRP-PHAT[14] and MUSIC[15], we use GCC-PHAT because it can estimate a direction of arrival in a short time frame of several tens of milliseconds. SRP-PHAT which is robust for noise and reverberant is also suitable for localizing a moving sound source but it requires more than two microphones. As it usually uses four microhones, we do not use SRP-PHAT. MUSIC method is used for noise robust sound source localization mainly using a microphone array of four channels or more. It estimats direction of arrival from a long time frame, such as several hundreds of milliseconds. The source had better not to move because covariance matrix is disturbed by changing direction of arrival while calculating the eignvalues. Although Ono, et al. [16] proposed a method for canceling a movement of sound source to avoid disturbing the covariance matrix, it is impossible to apply two channel microhpone array.

A. Baseline method

This subsection describes two types of baseline methods to be compared with a proposed method. The simplest method is to localize a wide range of a sound source with a single microphone array (Baseline-1). The second method estimates by averaging the four microphone arrays and integrating the four estimates (baseline-4).

The Baseline-1 method uses one mirophone array. The signal is obaseved by $M_{(0)}$. The direction θ_0 is calculated by GCC-PHAT. We can estimate source position of the moving sound source since we assumes the moving sound source moves on a known stright line, i.e. it is on a line $x_2 = 0$. The estimated point $P_{Baseline-1}$ is defined as the intersection

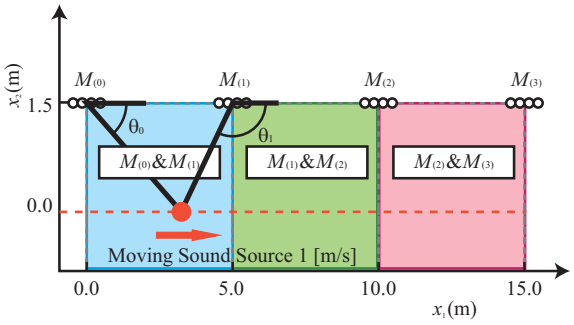

Fig. 2. Selection of micrhophone arrays and localization

of the lines $x_2 = 0$ and $x_2 = -x_1 \tan\theta_0 + 1.5$. Therefore,

$$P_{Baseline-1} = \left[\frac{1.5}{\tan\theta_0}, 0\right]^T \tag{6}$$

The Baseline-4 method uses all microphone array. The signal is observed by $M_{(0)}, M_{(1)}, M_{(2)}, M_{(3)}$. The direction $\theta_0, \theta_1, \theta_2$, and θ_3 are calculated by GCC-PHAT, respectively. Once source direction is given, we can calculate the source points $P_{(0)}, P_{(1)}, P_{(2)}, P_{(3)}$ from following equations.

$$P_{Baseline-4} = \frac{1}{4}\Big(P_{(0)} + P_{(1)} + P_{(2)} + P_{(3)}\Big), \tag{7}$$

$$P_{(m)} = \frac{1.5}{\tan\theta_m} + 5m, \tag{8}$$

where $m = 0, 1, 2, 3$.

B. Proposed method

A proposed method uses two microphone arrays that are the closest and the second closest to a moving sound source. The two microphone arrays are selected among four microphone arrays with depending on the sound source position. The moving sound source and two microhpone arryas are shown in Figure 2. An orange circle shows a moving sound source. It moves at 1 m/s along with the x_1-axis. Since currently it is in the blue area ($M_{(0)}$&$M_{(1)}$ area), $M_{(0)}$ and $M_{(1)}$ are selected. When it is in the green area, $M_{(1)}$ and $M_{(2)}$ are selected. In this way, the microphone arrays to be selected can be determined for each area. Since microphone arrays arranged in a straight line is used, the selection of two microphone arrays ($M_{(m)}$&$M_{(m+1)}$) close to the sound source can be obtained by

$$m = \arg\min_m \left\{ \left(\frac{\pi}{2} - \theta_m\right) + \left(\theta_{m+1} - \frac{\pi}{2}\right) \right\}. \tag{9}$$

After the selection, two sound source directions θ_m and θ_{m+1} are given. The position of the moving sound source is derived from θ_m and θ_{m+1} by triagulation. A result of triangulation do not always point on the line $x_2 = 0$. To compare with the baselines, the results is projected onto the line $x_2 = 0$.

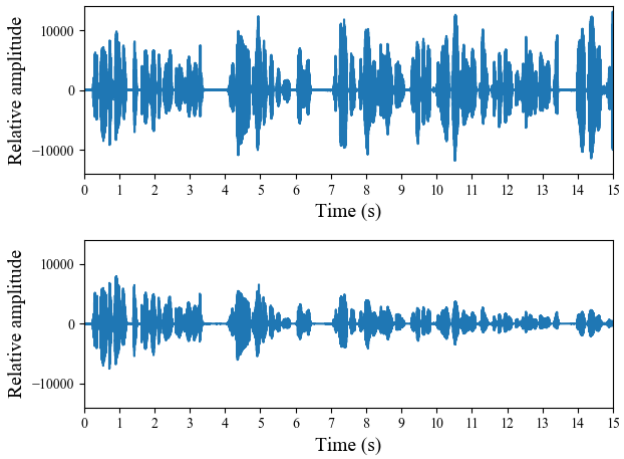

Fig. 3. Synthesized moving sound source (Upper panel: dry sound source, Lower panel: moving sound source).

IV. Experiment

In this section, the accuracy of localization is compared. The proposed method is compared with Baseline-1 and Baseline-4. First, we show computational simulations. Then we demonstrate that it is possible to track a moving sound source in a real environment.

A. Condition of simulation

A moving sound source is generated from the Japanese speech copus, JNAS[17]. The moving sound source is 15 s long signal which is concateneted ten phonetic balanced sentences utterd by a female (speaker ID: f001). The waveform of the moving sound source is shown in Figure 3. In the upper pannel, the dry sound source is shown. The moving sound source in the lower pannel is synthesized as a signal obserbed by one microphone in $M_{(0)}$ located at $[-0.06, 1.5]^T$ in a computer simulation. It starts from $[0, 0]^T$ and ends at $[15, 0]^T$. It's speed is 1 m/s. This figure shows the amplitude of the waveform decreases when time goes, since the moving sound source aparts from an observed point ($M_{(0)}$).

An experimental configuration is shown in Figure 2. Four microphone arrays are lined up intervals of 5 m. A sound source moves parallel to the stright line formed by four microphone arrays. A microhpone array consists of four elements which are lined up intervals of 0.04 m. The sound source keeps a distance of 1.5 m from the straight line.

The signals observed by the elements on both ends of the microphone array are processed by GCC-PHAT. A frame length and a frame shift length are 32 ms and 10 ms at 16 kHz sampling.

B. Evaluation by simulation

Figure 4 shows the result of localization. The horizontal and vertical axes represent time and the estimated position in x_1-axis. It notes that there is no error in x_2-axis under our assumption. The sound source positions estimated by three of

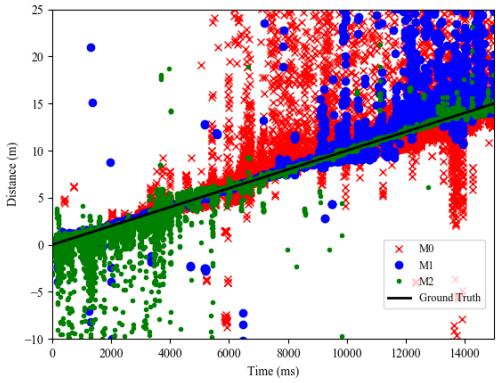

Fig. 4. Sound source trajectroy

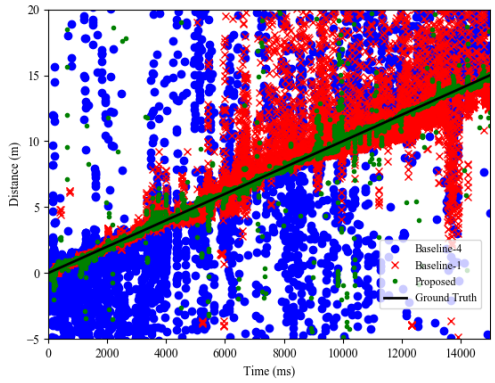

Fig. 5. Improvement of localization based on two nearest microphone arrays

the four microphone arrays $M_{(0)}, M_{(1)}, M_{(2)}$ are pointed as red cross($\times$), blue circle($\bullet$), and green dot($\cdot$), respectively. The solid black line represents the position of the moving sound source on the x_1-axis (Gound truth). The closer the estimated position is to the solid black line, the higher the estimation accuracy.

At $t = 0$, the sound source is directly below $M_{(0)}$ and gradually moves to the lower right and far away. At $t = 5000$, it passes under $M_{(1)}$, and at $t = 10000$, it reaches under $M_{(2)}$. As the sound source monotonically moves away from the $M_{(0)}$ position over time, the estimation results are distributed around the black line up to 5000 ms. However, the estimation results are widely distributed around the black line after 5000 ms. This is the effect of being too far from the sound source. It is suggested that it is hard to track sound sources away from over 5 m with one microphone array.

Figure 5 shows the estimation results of the proposed method (Proposed) and the conventional method (Baseline-1, Baseline-4). The horizontal and vertical axes represent time and the estimated position in x_1-axis. The estimated sound source positions of Baseline-1, Baseline-4 and Proposed methods are pointed as red cross ($\times$), blue circle ($\bullet$), and green dot ($\cdot$), respectively. The solid black line represents the position of the moving sound source on the x_1-axis (Gound truth).

Figure 5 indicats that Proposed is the highest accuracy. The estimated positions are distributed near the ground truth over all time from 0 to 15 s. On the otherhand, the positions estimated by Baseline-4 are widely distributed around the ground truth over all time from 0 to 15 s. It has also been shown that Baseline-1 is less accurate. The estimated positions are distributed near the ground truth from 0 to 5 s when the sound source is close to the microphone array. However, these are widely distributed around the ground truth after 5 s.

Figure 6 shows the error histgram to display error distributions. The horizontal and vertical axes represents absolute error and frequency. Blue, orange and green bars shows Baseline-1, Baseline-4 and Proposed mthods, respectively. A number of frequency bins is 15 and the bin width is 1 m, where the first

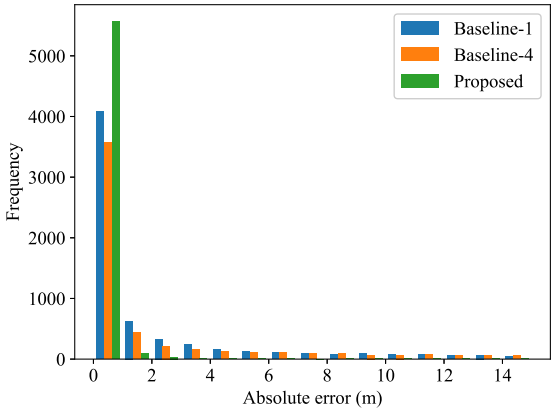

Fig. 6. Error distribution

bin width is 0.5 m (0 to 0.5).

All methods have the peak of the frequency at the first bin. It is shown that the errors are concentrated below 0.5 m. The frequency of the proposed method for all bins except for the first bin is the least among the methods. Therefore the proposed method achieves the least error to track the moving sound source than other methods.

In this way, the reason why a large estimation error occurs rarely is that the energy of the voice signal fluctuates locally. Even a distant sound source can be localized with high accuracy if the energy is large. Even if the sound source is close, if the energy is small, the localization accuracy decrease. In order to track the sound source robustly against such fluctuations, it is necessary to improve the sound source localization accuracy of one microphone array or to arrange multiple microphone arrays densely.

C. Evaluation in a real enviroment

For evaluating a tracking accuracy in a real environment, we conduct an experiment of tracking a moving sound source by

using two microphone arrays. A moving sound is realized by moving the loudspeaker by hand. The ground truth of moving sound source position is given by LiDAR (Light Detection and ranging). The LiDAR (Intel RealSense L515) had errors of ± 0.03 against a non-moving sound source when a position of a loudspeaker is less than 3.5 m from LiDAR.

We played white noise from the moving loudspeaker (Genelec 8020 DMP). The loudspeaker was moved between (0,1) to (0,5). The Output level was calibrated to LAeq 60 dB. The reference signal was played at 0.1 m away from the loudspeaker. A background noise wat about LAeq 42 dB.

Two microphone arrays $M_{(0)}$ and $M_{(1)}$ were placed at [0.5, 0], [-0.5, 0]. We recorded the moving sound source and simultaneously measured the position of the moving sound source by LiDAR placed at [0,0].

Figure 7 shows trajectory of a moving sound source. The horizontal and vertical axes represent time and the estimated position. A blue dotted line corresponds to the ground truth which is measured by LiDAR. A red solid line corresponds to the estimation of the proposed method based on GCC-PHAT.

We can see that the proposed method measures the position accurately, except for more than 3.5 m. These errors are due to the number of microphone elements that consist of the microphone array. It is thought that the effect of the estimation error of the source direction estimation method became dominant due to the decrease of the angular change of the source direction against the movement of the sound source. This is a problem that can be avoided by increasing the number of microphone elements.

The moving sound source starts from (0,5) and goes to (0,1). It stays there in a while. Then it go back to the start point. After that, the same movement is repeat once. The trajecotry estimated by proposed method based on GCC-PHAT is near to the LiDAR's trajectory under 3 m. This trend is the same as one of a computational simulation. These results demonstrate that the proposed method can localize practically when the nearest microphone arrays from a moving sound source is less than 3 m from the moving sound source.

In this experiment, white noise with constant energy was used as a sound source, and the same result cannot be expected when a sound source with temporal energy variation such as speech is used. The energy reduction in the analysis frame leads to a decrease in the accuracy of the source direction estimation, which leads to a wrong selection of the microphone array and a larger error in the source location estimation. In the future, we will need to evaluate the case where the moving source is speech sound. The local energy reduction causes a lack of source localization, and the source position does not show continuous values in time and space, but trajectories with scattered values are obtained. In order to deal with this problem, it is necessary to integrate tracking techniques based on state-space models such as Kalman filter and particle filter. However, the faster the moving source moves, the lower the tracking accuracy is expected to be. Therefore, in this research, we consider it equally important to improve the localization accuracy of frame-by-frame processing under low

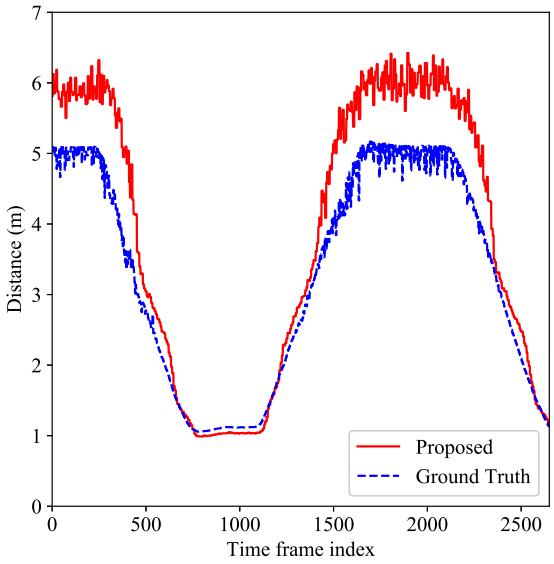

Fig. 7. Trajectories of a moving sound source (Red Solid Line: Proposed Method, Blue Dotted Line: ground truth measured by LiDAR)

latency conditions and to improve the tracking accuracy when dealing with moving sources. This is also the reason why the source direction estimation algorithm is based on GCC-PHAT. The time constant inherent in other algorithms such as MUSIC is long, and it is not suitable for localization of moving sources. There is also a problem that the accuracy drops drastically when the delay is forcibly reduced.Furthermore, while the MUSIC method discretizes the estimated direction, GCC-PHAT has the advantage that the estimated direction can be obtained as a continuous value without algorithmic constraints on the angular resolution of the estimated direction.

V. Conclusions

We propose a method for tracking a sound source that moves over a wide range using multiple microphone arrays. To evaluate the tracking accuracy, we use a method of estimating the direction under the assumption the moving sound source is moving on a known straight line in the simulation. And we demonstrated the proposed method practically is possible to track a moving sound source in a real environment.

The proposed method uses multiple microphone arrays that are close to the sound source, and selects the optimum arrays depending on the sound source position. It is shown that the proposed method can track with higher accuracy than the conventional method of tracking with one microphone and the method of tracking with all microphone arrays in the simulation. This simulation suggests that sound source localization accuracy can be improved by selecting several microphone arrays from multiple microphone arrays.

We evaluated accuracy of localization in 1D in this study, but our method can apply to 2D and 3D space. The reason of

evaluating in 1D is that it is hard to get ground truth in 2D/3D in a real environment. Another reason is that a range of sensing by LiDAR is limited. We would like to improve ability of the sensing by using multiple LiDAR. After that, we try to extend our method for applicable to the wider space.

As mentioned in the prveious section, there are a lot of future works. It is necessary to check the localization accuracy of the proposed method for a wide variety of sound sources, including speech sounds. And we have to evaluate the acuracy of the microphone array selection algorithm and the accuracy of the source localization including the selection error in a real environment. It might lead us to a new problem after extending our method into a two- or three-axis domain because the number of possible locations for the microphone array increases. We will solve these problems one by one.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number JP19K12056, JP21H03488.

REFERENCES

[1] B. Lecouteux, M. Vacher, and F. Portet, "Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions," *Proceedings Interspeech 2011 – International Speech Communication Association*, pp. 2273-2276, Aug. 2011.

[2] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind Alignment of Asynchronously Recorded Signals for Distributed Microphone Array," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 161-164, Oct. 2009.

[3] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "SLAM-based online calibration of asynchronous microphone array for Robot Audition," *Proceedings IROS 2010 – IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 524–529, Sept. 2011.

[4] K. Sekiguchi, Y. Bando, K. Itoyama, and K. Yoshii, "Layout optimization of cooperative distributed microphone arrays based on estimation of source separation performance," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 83–93, Feb. 2017.

[5] J. Mariani, S. Rosset, and L. Devillers(Eds), *Natural Interaction with Robots, Knowbots and Smartphones*. Springer, 2013.

[6] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Multiple sound source mapping of a mobile robot by self-motion triangulation," *Proceedings IROS 2006 – IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 380–385, Nov. 2006.

[7] K. Kwak, "Sound source tracking of moving speaker using multi-channel microphones in robot environments," *Proceedings ICRB 2011 – IEEE International Conference on Robotics and Biomimetics*, pp. 3017–3020, Dec. 2011.

[8] T. Kimura, K. Kakehi, K. Takeda, and F. Itakura, "Spatial coding based on the extraction of moving sound sources in wavefield synthesis," *Proceedings ICASSP 2005 – IEEE International Conference on Acoustic Speech and Signal Processing*, vol. III, pp. 293–296, March 2005.

[9] J. Nikunen, A. Diment, and T Virtanen, "Separation of Moving Sound Sources Using Multichannel NMF and Acoustic Tracking", *IEEE/ACM Transactions on Audio, Speech adn Language Processing*, vol. 26, no. 2, pp. 281-295, Feb. 2018.

[10] Q. Bao, F. Luan and J. Yang, "Improving the Accuracy of Beamforming Method for Moving Acoustic Souce Localization in Far-field", *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pp. 1-6, Oct. 2017.

[11] X. Guo, K. Yang, Q. Yang, S. Zhu, R. Cao and Y. Ma, "Tracking-position of sound speed profiles and moving acoustic source in shallow water", *IEEE Techno-Ocean 2016*, pp. 1-4, Oct. 2016.

[12] L. Rabiner, and B. Gold, *Theory and application of digital signal processing*. Prentice-Hall, 1975.

[13] C. H. Knapp, and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[14] J. H. DiBiase, "A hight-accuracy, low-latency technique for talker localization in reverberant environments using microhpone arrys," *Doctoral dissertation, Brown University*, May 2000.

[15] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, March 1986.

[16] Y. Wakabayashi, K. Yamaoka, and N. Ono, "Rotation-robust beamforming based on sound field interpolation with requlary circular microphone array," *Proceedings ICASSP 2021 – IEEE International Conference on Acoustic Speech and Signal Processing*, pp. 771–775, June 2021.

[17] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of Acoustical Society of Japan*, vol. 20, no. 3, pp. 199-206, 1999.