2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Daejeon Convention Center
October 9-14, 2016, Daejeon, Korea

Online Simultaneous Localization and Mapping of Multiple Sound Sources and Asynchronous Microphone Arrays

Kouhei Sekiguchi, Yoshiaki Bando, Keisuke Nakamura, Kazuhiro Nakadai,
Katsutoshi Itoyama, and Kazuyoshi Yoshii

***Abstract*— This paper presents an online method of simultaneous localization and mapping (SLAM) for estimating the positions of multiple moving sound sources and stationary robots and synchronizing microphone arrays attached to those robots. Since each robot with a microphone array can solely estimate the directions of sound sources, the two-dimensional source positions can be estimated from the source directions estimated by multiple robots using a triangulation method. In addition, sound mixtures can be separated accurately by regarding distributed microphone arrays as one big array. To perform these tasks, some methods have been proposed for localizing and synchronizing microphone arrays. These methods, however, can be used only if a single sound source exists because the time differences of arrival (TDOAs) between microphones are assumed to be directly observed. To overcome this limitation, we propose a unified state-space model that encodes the source and robot positions and the time offsets between microphone arrays in a latent space. Given the TDOAs and directions of arrival (DOAs) estimated by separating observed mixture sounds into source sounds, the latent variables are estimated jointly in an online manner using a FastSLAM2.0 algorithm that can deal with an unknown time-varying number of moving sound sources.**

I. Introduction

Computational auditory scene analysis has extensively been studied for understanding auditory events in a surrounding environment by conducting sound source localization or separation [1]. A single robot having a microphone array can solely estimate the directions of sound sources, although it is generally difficult to estimate the distance between the robot and a sound source when the distance is large compared to the size of the microphone array. Then, the two-dimensional positions of sound sources can be estimated at one time using multiple robots with a triangulation method [2], [3]. Moreover, multiple robots can conduct cooperative sound source separation by regarding distributed microphone arrays as one bigarray [4].

To perform sound source localization and separation based on distributed microphone arrays (robots), those arrays should be localized and synchronized in advance. The phase information of recorded multi-channel audio signals, which plays a central role in microphone array processing, is affected by the time offsets and relative positions of microphone arrays. To solve this problem, several studies have been conducted for synchronizing distributed microphones by using the time differences of arrivals (TDOAs) of sound sources between the microphones [5]–[7]. Since these methods assume that only one sound source is active at each time, a real environment where many people talk cannot be dealt with.

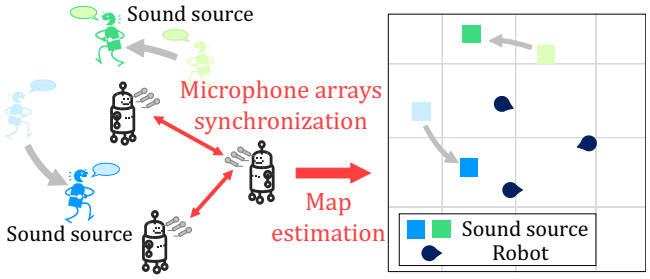


Fig. 1. Overview of the proposed audio-based SLAM method. The time offsets between microphone arrays and the positions of sound sources and robots are jointly estimated in an online manner using a state-space model.

In this paper we propose a statistical method that jointly estimates the time offset between each pair of microphone arrays, the positions of moving sound sources, and those of stationary robots in an online manner (Fig. 1). To estimate not only the TDOAs but also the direction of arrival (DOA) for each source, mixture signals recorded by microphone arrays are separated into individual source signals. Regarding both the TDOAs and DOAs of multiple sources as observed data, we formulate a state-space model that encodes the time offsets between microphone arrays and the positions of sources and robots in a latent space. Those latent variables are jointly estimated and updated over time using a FastSLAM2.0 algorithm [8]. The main contribution of this study is to propose a unified framework of audio-based simultaneous localization and mapping (SLAM) with microphone-array synchronization under a condition that multiple sound sources exist.

II. Related Work

This section reviews several studies on calibration of microphone arrays that aim to estimate the positions and time offsets of microphones. Such calibration is often needed under a realistic condition that neither multi-channel A/D converters nor geometrical information about microphones are available. One standard approach is to use loudspeakers that emit reference signals for estimating the time differences of arrival (TDOAs) between microphones. [9]–[12]. Peng *et*

K. Sekiguchi, Y. Bando, K. Itoyama and K. Yoshii are with the Graduate School of Informatics, Kyoto University, Sakyo, Kyoto, 606-8501, Japan. {sekiguch, yoshiaki, itoyama, yoshii}@kuis.kyoto-u.ac.jp

K. Nakamura and K. Nakadai are with Honda Research Institute Japan Co./ Ltd./ 8-1 Honcho, Wako, Saitama 351-0014, Japan. {keisuke, nakadai}@jp.honda-ri.com

al. [9], for example, estimated the positions of two asynchronous microphones by emitting specially designed sound signals from the loudspeakers near microphones. Pertila *et al.* [10] developed a method for estimating the positions and directions of the devices that each have a microphone and a loudspeaker without using special sound signals.

Another approach is to use only multi-channel audio signals asynchronously recorded by multiple microphones [5]–[7]. Hasegawa *et al.* [5], for example, proposed an offline method that estimates the positions and time offsets of microphones such that the mean square errors between observed and predicted TDOAs are minimized. In a standard setting of SLAM, mobile robots (microphone arrays) are used for localizing themselves and multiple stationary objects (sources). If multiple moving sound sources are observed by a single stationary robot, SLAM techniques can be used by reversing the roles of sources and robots. Su *et al.* [7] proposed an offline method that estimates the clock differences and time offsets between microphones, the position of a sound source, and those of the microphones. using a graph-based SLAM method. Miura *et al.* [6] proposed an online method that uses extended Kalman filter-based SLAM (EKF-SLAM) and delay-and-sum beamforming (DSBF) for judging the convergence of calibration by comparing the sound source positions estimated by EKF-SLAM and DSBF.

III. Proposed Method

This section describes an online method that estimates the time offset and position of each robot (microphone array) and the positions of sound sources when multiple sound sources exist. First, the time differences of arrival (TDOAs) and directions of arrival (DOAs) of sound sources are estimated. Those TDOAs and DOAs are used as observed data for a state-space model that encodes the time offset and position of each robot and the positions of sound sources as latent variables, which are estimated jointly in an online manner using a FastSLAM2.0 algorithm.

A. Problem Specification

We specify a problem of audio-based online SLAM for multiple robots and sound sources. Let M be the number of microphones on a single robot, I the number of robots, N the number of total sound sources, and F the number of frequency bins. In this paper, we assume that the robots and sound sources are on a two-dimensional plane. The estimation problem is defined as follows:

- **Input**: $I \times M$ channel input audio spectrogram $\boldsymbol{x}_t = [\boldsymbol{x}_{t1}, \cdots, \boldsymbol{x}_{t,I}]$.
- **Output**: (1) The two-dimensional positions and directions $\boldsymbol{r}_i$ of microphone array i ($i = 1, \ldots, I$).
  (2) The two-dimensional positions $\boldsymbol{s}_{k,n}$ of sound source n at the k-th measurement ($n = 1, \ldots, N$).
  (3) The time offset τ_{1j} between microphone array 1 and j.
- **Assumptions**:
  (1) At least one sound source is moving.
  (2) The robots are stationary.
  (3) Multiple microphone arrays are roughly synchronized (within about 10 ms). This is achieved by using a wireless connection of the robots and without using a special sound capturing system.

Here, $\boldsymbol{x}_{ti} = [\boldsymbol{x}_{ti1}, \ldots, \boldsymbol{x}_{tiM}]^T \in \mathbb{C}^{M \times F}$ denotes the spectrogram recorded by the microphone array i at the t-th time frame. $\boldsymbol{r}_i = [r_i^x, r_i^y, r_i^\theta]$ is a vector of two-dimensional position and direction of microphone array i. $s_{k,n}^x$ and $s_{k,n}^y$ denote the two-dimensional position of sound source n at the k-th observation.

B. Feature Extraction

The robot and sound source positions are estimated by using DOAs and TDOAs. If only DOAs are used to estimate positions, just the positions which are similar to the actual positions are estimated. DOAs and TDOAs enable robots to estimate the time offsets and the two-dimensional positions with the origin located at one of the robots.

1) DOA Estimation: The DOA of a sound source from each robot can be estimated by a microphone array processing method called multiple signal classification (MUSIC) [13]. To use MUSIC, which requires synchronized microphones, each robot is equipped with a synchronized microphone array. MUSIC can estimate DOAs even if the observed signals are mixtures of multiple sound sources, although we need to specify the number of sound sources beforehand. DOA estimation does not necessarily fail if the actual number of sound sources is not what we expected, but its accuracy may deteriorate.

2) TDOA Estimation: TDOAs are estimated only when sound sources are detected at the DOA estimation. If there is only one sound source, TDOA is estimated as follows. The cross-correlation coefficients are calculated by using a generalized cross-correlation with phase transform (GCC-PHAT) [14]. The coefficient G_{PHAT} of the GCC-PHAT between the microphone m_1 and m_2 is calculated as follows:

$$G_{PHAT}(f) = \frac{X_{m_1}(f) X_{m_2}^*(f)}{|X_{m_1}(f) X_{m_2}^*(f)|}, \tag{1}$$

where $X_m(f)$ is the Fourier transform of the signal recorded by the microphone m. To estimate the TDOA between the robot i_1 and i_2, the coefficients of GCC-PHAT between the first microphone of the robot i_1 and the first microphone of the robot i_2 is calculated. This coefficient is transformed into the time domain signals, and the peaks of this time-domain signals correspond to the TDOA; therefore, TDOA ξ is calculated as follows:

$$\xi = \mathrm{argmax}_\xi \int G_{PHAT}(f) e^{j2\pi f \xi} df. \tag{2}$$

When there are multiple sound sources, the TDOA of each must be estimated. This can not be done using above method, because even if the cross correlation coefficients have the same number of peaks as the sound sources, it is impossible to estimate which peaks correspond to which sound sources.

This problem is solved by using sound source separation. Fig. 2 shows the outline of the TDOA estimation from the

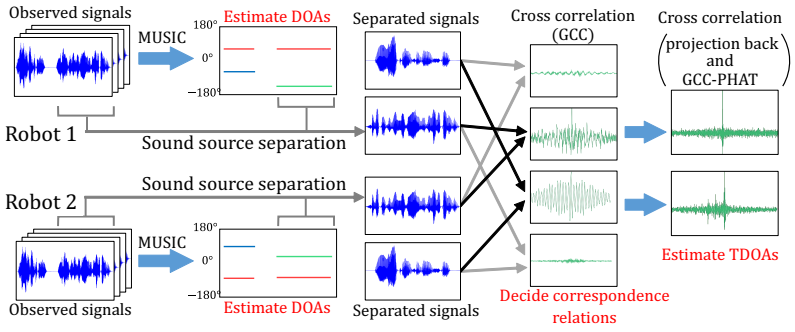


Fig. 2. How to estimate DOAs and TDOAs when there are multiple sound sources.

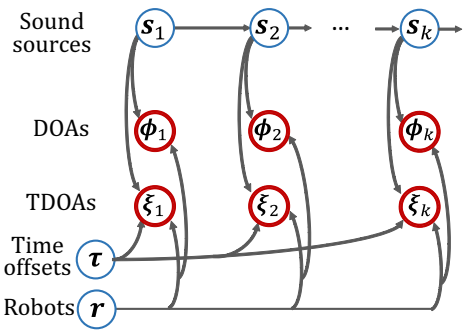


Fig. 3. Graphical representation of the state-space model.

mixture signals. The observed signals are separated into each sound source signal. We use geometric-constrained high-order decorrelation-based source separation (GHDSS) as a source separation method. Since the computational cost is low and the separation performance of the method is high, it is suitable for robot audition in which real-time processing is needed. To estimate TDOA of each sound source, we need correspondences of separated signals among robots. In addition, each robot outputs different number of separated signals. Moreover, some of the separated signals may not be used due to the failure of the sound source separation. To decide the correspondence relations, the cross correlation coefficients are calculated for all combinations of separated signal of each robot by using a generalized cross correlation (GCC) [14]. Let $Y_{i_1 l_1}$ be the separated signal of the l_1-th sound source detected by robot i_1. The coefficient G of GCC between the separated signals $Y_{i_1 l_1}$ and $Y_{i_2 l_2}$ is calculated as follows:

$$G(f) = Y_{i_1 l_1}(f) Y^*_{i_2 l_2}(f). \tag{3}$$

If the maximum value of this coefficient is larger than a threshold, the separated signal $Y_{i_1 l_1}$ and $Y_{i_2 l_2}$ are regarded as signals from the same source. As can be seen from Eqs. (1) and (3), GCC is different from GCC-PHAT in that GCC-PHAT focuses on only the difference in phase, while GCC focuses on the differences in phase and power.

To calculate TDOAs of the separated signals, there is a problem that sound source separation makes the phase of the separated signal different from that of the observed signal. To eliminate the difference in phase, the phase of the l-th separated signal is shifted by multiplying the l-th column of the inverse matrix of the separation matrix. This process is called *projection back* [15] and was originally used to solve a scaling problem of blind source separation.

C. State-Space Model

To estimate the robot and sound source positions (states) and time offsets, our method uses a state-space model. As shown in Fig. 3, the sound source positions are defined as time-dependent latent variables, and the robot positions and the time offsets are defined as time independent latent variables. To estimate the n-th sound source position $\boldsymbol{s}_{k,n}$ at the k-th measurements from that at the $(k-1)$-th measurements, movement speed $s^v_{k,n}$ and movement direction $s^\theta_{k,n}$ of a sound source are added to the latent variables. Therefore, the latent variable at the k-th measurement $\boldsymbol{z}_k$ is defined as follows:

$$\boldsymbol{z}_k = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_I, \boldsymbol{s}_{1,k}, \cdots, \boldsymbol{s}_{N,k}, \boldsymbol{\tau}] \tag{4}$$

where the i-th robot state $\boldsymbol{r}_i$, the n-th sound state $\boldsymbol{s}_{n,k}$, and the time offsets $\boldsymbol{\tau}$ are given by

$$\boldsymbol{r}_i = [r^x_i, r^y_i, r^\theta_i] \tag{5}$$

$$\boldsymbol{s}_{k,n} = [s^x_{k,n}, s^y_{k,n}, s^v_{k,n}, s^\theta_{k,n}] \tag{6}$$

$$\boldsymbol{\tau} = [\tau_{12}, \cdots, \tau_{1I}]. \tag{7}$$

The states of sound sources, robots, and time offsets are estimated by using a FastSLAM2.0 algorithm, which is originally an algorithm for solving a SLAM problem.

1) State Update Model: Since robots are stopping, the state update is conducted only for sound source states. Assuming that the sound source states follow the Gaussian distribution, the state update model of the sound source n is represented as follows:

$$\begin{bmatrix} s^x_{k+1,n} \\ s^y_{k+1,n} \\ s^v_{k+1,n} \\ s^\theta_{k+1,n} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} s^x_{k,n} + s^v_{k,n}\cos(s^\theta_{k,n})\Delta t \\ s^y_{k,n} + s^v_{k,n}\sin(s^\theta_{k,n})\Delta t \\ s^v_{k,n} \\ s^\theta_{k,n} \end{bmatrix}, Q \right), \tag{8}$$

where $Q \in \mathbb{R}^{4\times 4}$ is the covariance matrix of the state update noise, Δt is the elapsed time since the last observation, and the initial source state $\boldsymbol{s}_{0,n}$ follows a uniform distribution.

2) Measurement Model: The measurements to estimate latent variables are DOAs and TDOAs. DOAs are calculated with regard to each robot, and TDOAs are calculated using robot 1 as a standard. Let $\phi_{k,i,n}$ be the direction from the i-th robot to the n-th sound source at time k, and let $\xi_{k,j,n}$ be the TDOA of sound source n between robot 1 and j at time k. Since all measurements are independent, the measurement model $p(\boldsymbol{\phi}_k, \boldsymbol{\xi}_k | \boldsymbol{s}_k, \boldsymbol{r}, \boldsymbol{\tau})$ is calculated as follows:

$$p(\boldsymbol{\phi}_k, \boldsymbol{\xi}_k | \boldsymbol{s}_k, \boldsymbol{r}, \boldsymbol{\tau}) = \prod_{n=1}^{N} \left(\prod_{i=1}^{I} p(\phi_{k,i,n} | \boldsymbol{s}_k, \boldsymbol{r}) \prod_{j=2}^{I} p(\xi_{k,j,n} | \boldsymbol{s}_k, \boldsymbol{r}, \boldsymbol{\tau}) \right). \tag{9}$$

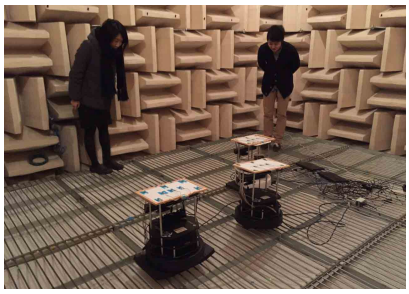
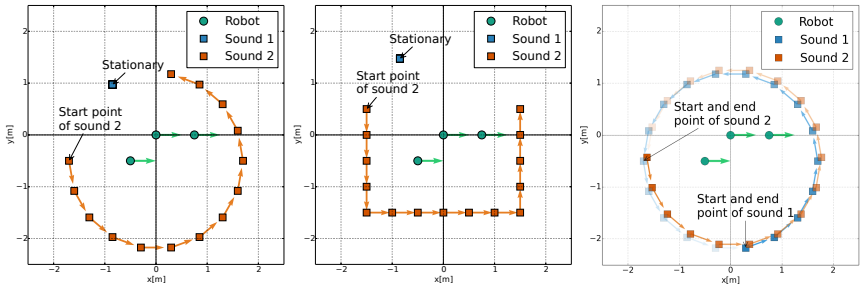


Fig. 4. Experimental condition in an anechoic chamber. There are three robots with 8-ch microphone arrays, and two sound sources (people).



(a) Pattern 1 (b) Pattern 2 (c) Pattern 3

Fig. 5. Configuration of sound source movements and robot positions. Circles and squares indicate the robot and sound source positions, respectively. Arrows on circles indicate the directions of robots, and arrows on squares indicate the movement directions of sound sources.

Assuming that the distribution of measurements to be Gaussian, $p(\phi_{k,i,n}|\boldsymbol{s}_k, \boldsymbol{r})$ and $p(\xi_{k,j,n}|\boldsymbol{s}_k, \boldsymbol{r}, \boldsymbol{\tau})$ are expressed as follows:

$$p(\phi_{k,i,n}|\boldsymbol{s}_k, \boldsymbol{r}) = \mathcal{N}\left(\arctan\left(\frac{s^y_{k,n} - r^y_{k,i}}{s^x_{k,n} - r^x_{k,i}}\right) - r^\theta_{k,i}, \sigma^2_\phi\right), \tag{10}$$

$$p(\xi_{k,j,n}|\boldsymbol{s}_k, \boldsymbol{r}, \boldsymbol{\tau}) = \mathcal{N}\left((l_{k,j,n} - l_{k,1,n})/C + \tau_{1j}, \sigma^2_\xi\right), \tag{11}$$

where σ^2_ϕ and σ^2_ξ are variance parameters, $l_{k,j,n}$ is the distance between the robot j and the sound source n at time k, and C is the speed of sound.

D. State Estimation Algorithm

A FastSLAM2.0 algorithm [8] is used for estimating the robot and sound source positions and the time offsets. Robot positions in a general SLAM problem correspond to the sound positions in our problem, and landmarks correspond to the robot positions and the time offsets. We select the FastSLAM2.0 algorithm because it can be used when the number of sound sources is unknown and because it can deal with the unknown-data-association problem described later. In this section, we give a brief summary of the FastSLAM2.0 algorithm.

This algorithm approximates the posterior distribution $p(\boldsymbol{s}_k, \boldsymbol{r}, \boldsymbol{\tau}|\boldsymbol{\phi}_{1:k}, \boldsymbol{\xi}_{1:k})$ by a set of samples. If multiple sound sources are observed at the same time, each measurement is processed sequentially, regarding the elapsed time Δt of the second and following measurements as zero. With regard to a sample m at time k, we first need to determine the data association $c^{[m]}_k$, which indicates which already detected sound sources the measurement arises from. The data association is determined by calculating a likelihood $p(\boldsymbol{\phi}_k, \boldsymbol{\xi}_k|\hat{\boldsymbol{s}}^{[m]}_k, \boldsymbol{r}^{[m]}_{k-1}, \boldsymbol{\tau}^{[m]}_{k-1}, c^{[m]}_k)$, where $\hat{\boldsymbol{s}}^{[m]}_k$ is sampled from the proposal distribution $p(\boldsymbol{s}^{[m]}_k|\boldsymbol{\phi}_k, \boldsymbol{\xi}_k, \boldsymbol{s}^{[m]}_{k-1}, \boldsymbol{r}^{[m]}_{k-1}, \boldsymbol{\tau}^{[m]}_{k-1}, c^{[m]}_k)$ calculated by using the extended Kalman filter (EKF). If the maximum likelihood is smaller than a threshold, the measurement is considered to be generated from a new sound source. In this case the robot positions and the time offsets are not updated, and the position of the new sound source is calculated by a triangulation method. Due to the noise of DOAs and the uncertainty of the robot positions, the number of the intersection points of the triangulation is up to ${}_IC_2$. The position of the new sound source $[s^{[m]x}_{k,\mathrm{new}}, s^{[m]y}_{k,\mathrm{new}}]$ is calculated as the mean of these intersection points as follows:

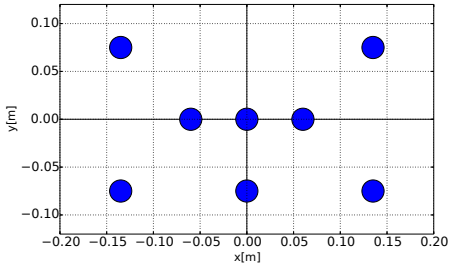


Fig. 6. The layout of an 8-ch microphone array on each mobile robot.

$$\begin{bmatrix} s^{[m]x}_{k,\mathrm{new}} \\ s^{[m]y}_{k,\mathrm{new}} \end{bmatrix} = \frac{1}{{}_IC_2} \sum_{r_1}^{I} \sum_{r_2 \neq r_1}^{I} \begin{bmatrix} \frac{\alpha^{[m]}_{k,r_1} - \alpha^{[m]}_{k,r_2}}{\tan\psi^{[m]}_{k,r_2} - \tan\psi^{[m]}_{k,r_1}} \\ \frac{\alpha^{[m]}_{k,r_1}\tan\psi^{[m]}_{k,r_2} - \alpha^{[m]}_{k,r_2}\tan\psi^{[m]}_{k,r_1}}{\tan\psi^{[m]}_{k,r_2} - \tan\psi^{[m]}_{k,r_1}} \end{bmatrix}, \tag{12}$$

where $\psi^{[m]}_{k,r_i}$ and $\alpha^{[m]}_{k,r_i}$ is defined as follows:

$$\psi^{[m]}_{k,r_i} = r^{[m]\theta}_{k-1,r_i} + \phi_{k,r_i,\mathrm{new}} \tag{13}$$

$$\alpha^{[m]}_{k,r_i} = r^{[m]y}_{k-1,r_i} - \tan(\psi^{[m]}_{k,r_i}) r^{[m]x}_{k-1,r_i} \tag{14}$$

If the maximum likelihood is larger than a threshold, the measurement is considered to be generated from the known sound source, and the robot positions and the time offsets are updated by the EKF. Sound sources whose states are not updated for a prescribed period of time are deleted. This process is conducted in order to deal with the pseudo sound sources generated by an improper measurement.

The final estimation results for the robot positions and the time offsets are obtained by calculating the weighted average of each particle. Since the estimated number of sound sources is different for each particle, we cannot calculate the weighted average of the sound source positions, and the estimation results for the sound source positions are

calculated as follows. First, the position of each sound source are classified by using a K-means algorithm based on the direction of the sound source from a centroid of the robot positions. The parameter K is calculated by rounding out the weighted average of the number of sound sources of each particle. Second, with regard to each class, the estimation result is calculated as the weighted mean of each sound source classified into the class.

IV. Experimental Evaluation

This section reports experimental results of the proposed method by using three robots and two sound sources.

A. Experimental Conditions

This experiment was conducted in an anechoic chamber in which there were two sound sources and three robots (Fig. 4). Each of the robots had an eight-channel microphone array whose layout is shown in Fig. 6. The following three patterns of the movements of sound sources were tested (Fig. 5).

1) **Pattern 1:** One sound source was stationary and the other moved along a circular route. The recording time was 40 seconds.
2) **Pattern 2:** Same as the Pattern 1 except that the route of the moving source was changed to a square and the stationary sound source was put 0.5 m away from the position in Pattern 1. The recording time was 45 seconds.
3) **Pattern 3:** Both sound sources moved along the same circular route with different start points. The recording time was 55 seconds.

At the points indicated by the square marks, sources emitted sounds almost simultaneously. To get the correct time offsets we conducted synchronous recording by using a multi-channel A/D converter (RASP-24 manufactured by Systems In Frontier Corp) with a sampling rate of 16 kHz and a quantization of 16 bits. We then intentionally shifted the signals recorded by robots 2 and 3 by 10 ms and -5 ms, respectively.

The configuration of the FastSLAM was that the number of particles was 50000, the initial states of each particle was generated randomly, the standard deviation of DOA and TDOA measurements were 5° and 0.1 ms respectively, and other parameters were determined experimentally. The sound source separation was conducted online by the geometric high-order dicorrelation-based source separation (GHDSS) method [16]. The open-source robot-audition software *HARK* [17] was used for conducting the MUSIC and GHDSS. The states were updated only when we got DOAs different from the DOAs observed within the previous two seconds.

We evaluated the estimation error of the robot positions, robot angles, time offsets, and sound positions. Since we didn't know the correspondence relations between the estimated and actual sound sources, the estimation error of a sound source was defined as the distance between the actual sound position and the estimated sound position closest to the actual one. The estimated number of the sound sources was not always the same as the actual number, and if it was smaller than the actual number, we didn't calculate the estimation error of sound sources that had no corresponding estimated sound source.

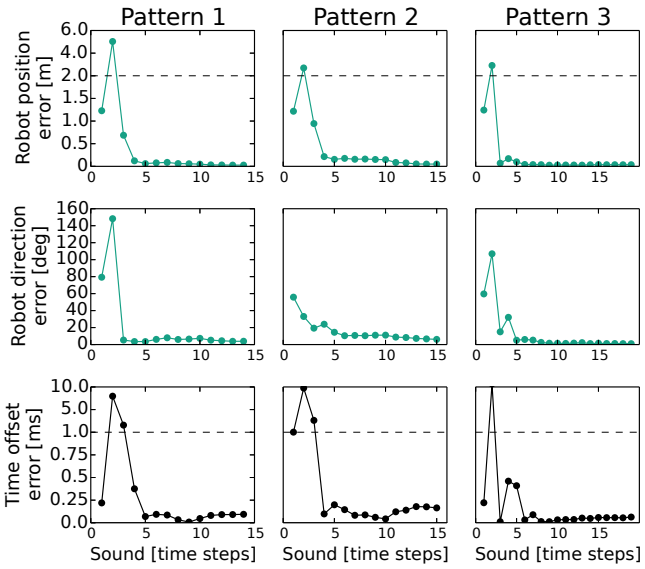


Fig. 7. Estimation errors of the robot positions, the robot angles, and the time offsets.

B. Experimental Results

Fig. 7 shows the estimation error of the robot positions, the robot directions, and the time offsets. In all the patterns they were estimated with high accuracy, and, after the last measurement, the mean errors of the robot positions, the robot directions, and the time offsets were less than 0.05 m, 10 degree, and 0.2 ms, respectively. Since the sampling rate of the recording was 16 kHz, 0.2 ms of the time offset estimation error was equivalent to 3.2 samples. This value is so small that it does not matter in the adaptive sound source separation methods [17].

Fig. 8 shows the estimation error of the sound source positions and the estimated number of sound sources. The estimated number of sound sources is the weighted mean of each particle. In the first pattern the final estimation errors of the both sound sources were less than 15 cm and the estimated number of the sound sources was almost always 2 except at the second measurement. In the second pattern, the estimated sound source directions were almost correct, although the accuracy of the estimated sound source positions was low and the estimated number of sound sources were often more than 2.

The reason why the estimated number of sources often became more than 2 is due to the estimation error of the correspondence relations. When the performance of source separation is low, the cross correlation between the source signals from the different sources also becomes high, and the correspondence relations would be mistakenly decided. Then, at the decision step of the data association in FastSLAM 2.0 algorithm, the likelihood $p(\boldsymbol{\phi}_k, \boldsymbol{\xi}_k | \hat{\boldsymbol{s}}_k^{[m]}, \boldsymbol{r}_{k-1}^{[m]}, \boldsymbol{\tau}_{k-1}^{[m]}, c_k^{[m]})$ becomes small, and a pseudo sound source is created.

One reason why the estimation of sound source positions

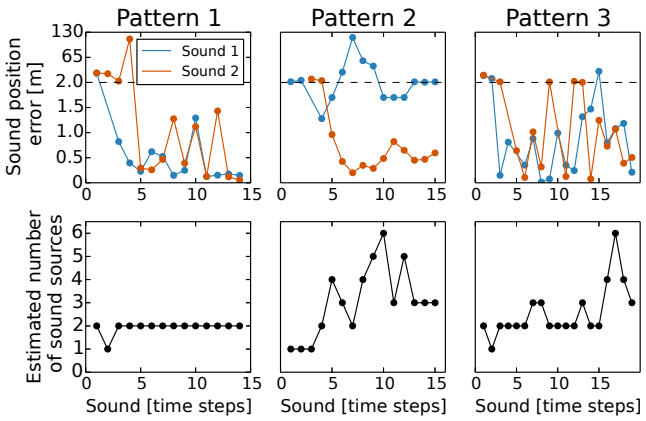


Fig. 8. Estimation errors of the sound source positions and the estimated number of sound sources.

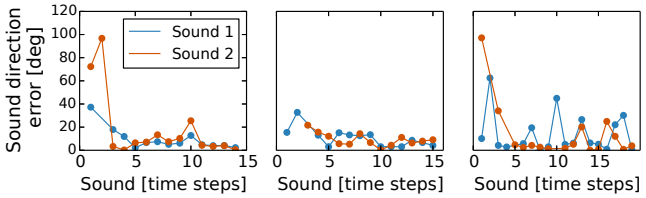


Fig. 9. Estimation errors of the sound source directions viewed from the centroid of the robot positions.

failed in some cases is that the distance between a robot and a sound source is relatively long compared to the distances between the robots. Although the direction of the estimated sound source is almost correct, a slight error of the DOA estimation results in a large estimation error. Fig. 9 shows the estimation errors of the sound source directions. These directions mean those of the sources measured on the centroid of the robot positions. These results show that the estimation errors of the source directions were almost less than 20 deg. Fig. 10 shows the sound source positions of each particle and the estimation results after the 14th measurement. We also see that the particles were distributed on the correct sound source directions although the estimated position was not correct. In this case, even if we increase the number of particles, the estimation error would not become small.

One way to improve the proposed method is to make the robots move around the sound sources. By using the robot movements, we can correct the sound source positions with the different relative directions of the sound sources. This approach has been studied in the context of active audition [18]–[20]. These studies will be effective for our extension.

V. Conclusion

This paper presented a method that in an environment with multiple sound sources conducts audio-based SLAM and synchronizes multiple microphone arrays simultaneously by using multiple robots that each have a microphone array.

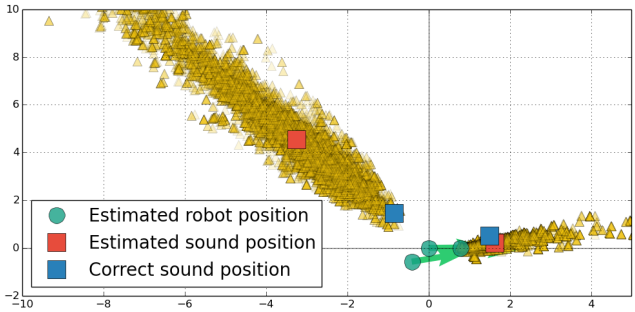


Fig. 10. The experimental result in the Pattern 2 after the 14-th measurement. Yellow triangles, red squares, and blue squares indicate the sound source positions of each particle, the weighted mean of the yellow triangles, and the correct sound source positions, respectively.

Conventional methods using asynchronous microphones assume that only one sound source is active at each time. In our method, taking advantage of using microphone arrays, we conduct sound source separation to estimate TDOAs from observed mixture signals and estimate DOAs by using a microphone array processing technique. We integrate estimated TDOAs and DOAs by using a state-space model, and we estimate the positions of sound sources and robots, the robot directions, and the time offsets between the microphone arrays by using a FastSLAM algorithm. We conducted an experiment to evaluate the estimation accuracy of the proposed method in anechoic chamber. In all three patterns, the estimation error of the robot positions, the robot directions, and the time offsets after the last measurement were less than 5 cm, 10 degree, and 0.2 ms, respectively. Although the estimation of the sound source position was difficult in some cases, the estimation error of the sound source positions after the last measurement was less than 20 cm in one pattern.

We plan to extend our method so that the robot can move. In the current method, there is a problem that the estimation of the sound sources is likely to fail in some cases. Although the uncertainty of the robots becomes larger when robots are moving, we can reduce the uncertainty of the sound sources by moving robots to optimal positions. Moreover, when the uncertainty of the sound sources is reduced, the uncertainty of the robots is also expected to be reduced.

Acknowledgment

This study was partially supported by JSPS KAKENHI Grant Number 24220006 and the Tough Robotics Challenge, ImPACT, Cabinet Office, Japan.

References

[1] H. G. Okuno *et al.*, "Robot audition: Missing feature theory approach and active audition," in *Robotics Research*. Springer, 2011, vol. 70, pp. 227–244.

[2] T. Nakashima *et al.*, *Natural Interaction with Robots, Knowbots and Smartphones*. Springer, 2014, ch. Integration of Multiple Sound Source Localization Results for Speaker Identification in Multiparty Dialogue System, pp. 153–165.

[3] E. Martinson *et al.*, "Optimizing a reconfigurable robotic microphone array," in *IEEE/RSJ IROS*, 2011, pp. 125–130.

[4] K. Sekiguchi *et al.*, "Optimizing the layout of multiple mobile robots for cooperative sound source separation," in *IEEE/RSJ IROS*, 2015, pp. 5548–5554.

[5] K. Hasegawa *et al.*, *Latent Variable Analysis and Signal Separation*. Springer, 2010, ch. Blind Estimation of Locations and Time Offsets for Distributed Recording Devices, pp. 57–64.
[6] H. Miura *et al.*, "SLAM-based online calibration of asynchronous microphone array for robot audition," in *IEEE/RSJ IROS*, 2011, pp. 524–529.
[7] D. Su *et al.*, "Simultaneous asynchronous microphone array calibration and sound source localisation," in *IEEE/RSJ IROS*, 2015, pp. 5561–5567.
[8] S. Thrun *et al.*, "FASTSLAM: An efficient solution to the simultaneous localization and mapping problem with unknown data association," *J. Machine Learning Research*, 2004.
[9] C. Peng *et al.*, "Beepbeep: A high accuracy acoustic ranging system using COTS mobile devices," in *Sensys*, 2007, pp. 1–14.
[10] P. Pertila *et al.*, "Closed-form self-localization of asynchronous microphone arrays," in *HSCMA*, 2011, pp. 139–144.
[11] M. H. Hennecke *et al.*, "Towards acoustic self-localization of ad hoc smartphone arrays," in *HSCMA*, 2011, pp. 127–132.
[12] H. H. Fan and C. Yan, "Asynchronous differential tdoa for sensor self-localization," in *IEEE ICASSP*, 2007, pp. 1109–1112.
[13] R. Schmidt *et al.*, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.
[14] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.
[15] N. Murata *et al.*, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.
[16] H. Nakajima *et al.*, "Blind source separation with parameter-free adaptive step-size method for robot audition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1476–1485, 2010.
[17] K. Nakadai *et al.*, "Design and implementation of robot audition system HARK – open source software for listening to three simultaneous speakers," *J. Advanced Robotics*, vol. 24, no. 5-6, pp. 739–761, 2010.
[18] ——, "Active audition for humanoid," in *IEEE AAAI*, 2000, pp. 832–839.
[19] G. L. Reid and E. Milios, "Active stereo sound localization," *J. Acoustical Society of America*, vol. 113, no. 1, pp. 185–193, 2003.
[20] E. Berglund and J. Sitte, "Sound source localisation through active audition," in *IEEE/RSJ IROS*, 2005, pp. 509–514.