2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Congress Center Hamburg
Sept 28 - Oct 2, 2015. Hamburg, Germany

Optimizing the Layout of Multiple Mobile Robots for Cooperative Sound Source Separation

Kouhei Sekiguchi, Yoshiaki Bando, Katsutoshi Itoyama, and Kazuyoshi Yoshii

***Abstract*— This paper presents a novel active audition method that enables multiple mobile robots to move to optimal positions for improving the performance of sound source separation. A main advantage of our distributed system is that each robot has its own microphone array and all mobile robots can collaborate on source separation by regarding a set of movable microphone arrays as a big reconfigurable array. To incrementally optimize the positions of the robots (the layout of the big microphone array) in an active-audition manner, it is necessary to predict the source separation performance from a possible layout of the next time step although true source signals are unknown. To solve this problem, our method simulates delay-and-sum beamforming from a possible layout for theoretically calculating the gain for each frequency component of a source signal in the corresponding separated signal. The robots are moved into a layout with the highest average gain over all sources and the whole frequency range. The experimental results showed that the harmonic mean of signal-to-distortion ratios (SDRs) was improved by 6.0 dB in simulations and by 5.7 dB in a real environment.**

I. Introduction

Simultaneous localization and mapping (SLAM) has actively been studied in recent years as one of the most fundamental techniques for mobile robots that work autonomously in an unknown environment [1]–[3]. In general it is difficult for a robot to directly estimate its own absolute position without using a GPS system. To make a near-field map and estimate its own relative position on the map, the robot needs to gather information of a surrounding area. A scanned area is gradually expanded by moving the robot. Several kinds of visual information obtained from cameras [1] and laser rangefinders [2], [3] have commonly been used for attaining accurate SLAM. Those sensors, however, can neither work effectively if obstacles exist in the field of vision nor detect notable sounds occurring around the robot.

Recently, audio-based SLAM has gained a lot of attention for making a map of sound objects [4]–[6]. If multiple sound sources exist in an environment, a robot needs to localize and separate mixture sounds recorded by its own microphones. Although this is attained by using a technique of microphone array processing [7]–[11], its performance is limited by the layout of sound sources, *e.g.*, if multiple sound sources exist in the same direction, the separation performance is degraded [12]. One solution to this problem is to move the robot to a better position. Although such *active audition* has been considered to be promising [13], it is often difficult to find a position that sparsifies the directions of all sound sources.

The authors are with the Graduate School of Informatics, Kyoto University, Sakyo, Kyoto, 606-8501, Japan. {sekiguch, yoshiaki, itoyama, yoshii}@kuis.kyoto-u.ac.jp

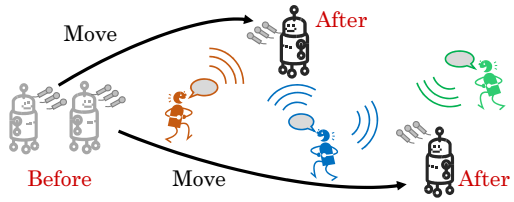


Fig. 1. Optimizing the layout of multiple mobile robots for cooperative sound source separation that regards a set of movable microphone arrays as a big reconfigurable microphone array.

A new approach to audio-based SLAM is to use *multiple* mobile robots, each of which has a microphone array, for cooperative localization and separation of *multiple* sound sources existing in a real environment. Such a distributed system has an advantage that each robot can independently work or all robots can cooperate by regarding a set of movable microphone arrays as a big reconfigurable array [14]. While each robot can estimate only the directions of sound sources, for example, the two-dimensional positions of those sources can be estimated from the localization results of multiple robots by using a triangulation method. Since a lot of effort has been devoted to sound source localization [15]–[17], in this paper we focus on cooperative sound source separation using multiple mobile robots.

One of the main problems of cooperative sound source separation is how to determine the optimal layout of multiple mobile robots that maximizes the performance of source separation. The actual performance of source separation cannot be calculated because true source signals are unknown. It is thus necessary to *predict* the source separation performance from a possible layout of the next time step before actually moving into the layout.

In this paper we propose a novel active audition method that moves multiple robots to optimal positions by simulating delay-and-sum beamforming from a possible layout under a condition that the positions of sound sources are already known (Fig. 1). If this type of beamforming is used for source separation, the gain, which is the expected ratio of a target sound source and the other sound sources in the corresponding separated signal, can be theoretically calculated from the positions of the robots and sound sources. The separation performance is then predicted from the gains by taking into account of the directions of microphone arrays and the distances between the sound sources and the robots. Since it is often difficult to find the best layout with the highest predictive performance via local search, we use a genetic algorithm that tends to avoid the local optimal solution.

*Demo page: http://winnie.kuis.kyoto-u.ac.jp/members/sekiguch/iros2015/

978-1-4799-9993-4/15/$31.00 ©2015 IEEE

II. Related Work

This section introduces several studies on active audition and sound source separation.

A. Active Audition

Active audition is a technique that aims to improve the performance of auditory scene analysis (analysis of surrounding sound objects) by making effective use of the movement of a robot equipped with microphones. Several studies have tried to accurately estimate the directions of sound sources by turning the head of a humanoid robot. Nakadai *et al.* [18], for example, developed a humanoid robot with two microphones that can track sound source directions by integrating audio, visual, and motor control. Berglund and Sitte [19] developed a robot with two microphones that learns how to orient toward a sound source via reinforcement learning. Kim *et al.* [15] proposed a method that can reduce the errors of sound source localization by taking into account the results of voice activity detection (VAD) and face tracking.

Active audition has often been used for a single moving robot that estimates the position of a sound source. Reid and Milios [16], for example, developed a robot that estimates the 3D position of a sound source by moving two microphones. Sasaki *et al.* [17] developed a mobile robot with a microphone array that estimates the positions of multiple sound sources. Since the robot can move around sound sources, the positions of those sources are obtained from source directions estimated from different observation positions in a way of triangulation. Yoshida and Nakadai [20] integrated the audio, visual, and active motion functions of a robot to estimate how the active motion affects the VAD.

If multiple robots are used, the positions of sound sources can be obtained quickly without moving robots. Martinson *et al.* [14] optimized a layout of multiple robots to improve the performance of sound source localization in a two-dimensional space. In this work, each robot was equipped with a single microphone and the optimal layout was determined such that each robot was distant from both the other robots and obstacles and close to sound sources.

B. Sound Source Separation

Delay-and-sum beamforming (DSBF) is a basic technique of microphone-array-based sound source separation [17], [21]. Sasaki *et al.* [21] attempted to optimize the layout of a 32-channel microphone array for improving the performance of DSBF. The optimal layout is determined such that it has high directivity to all directions.

Independent component analysis (ICA) is an another popular technique of source separation that aims to discover statistically independent source signals from given mixed signals. Although such time-domain ICA can separate convolutionally mixed signals, it requires high computational costs. Frequency-domain ICA, on the other hand, is more efficient by performing standard ICA at each frequency band. Since frequency-domain ICA has a problem about permutation of frequency bands, many studies have been conducted in an attempt to solve this problem [8]–[11], [22], [23].

III. Proposed Method

This section describes a proposed method that optimizes the layout of multiple mobile robots for cooperative sound source separation. Each robot is equipped with a standard microphone array. Multi-channel audio signals are recorded by regarding a set of the distributed microphone arrays of the robots as a big microphone array. The method uses delay-and-sum beamforming (DSBF) for extracting audio signals coming from a particular direction.

To optimize the layout of the robots, we need to design an objective function to be maximized with respect to a layout. In this study we can theoretically predict the performance of DSBF-based source separation in advance of actually moving the robots. More specifically, the ratio of a source signal in the corresponding separated signal (separation performance) is determined by specifying a *mixing process* that represents propagation of source signals to microphones and a *filtering process* that represents extraction of source signals from observed signals. Since it is often difficult to find the best layout with the highest performance via local search, we use a genetic algorithm that tends to avoid the local optima.

A. Problem Specification

Our goal is to find a layout of multiple robots that enables high-quality separation of all sound sources existing in a test environment. Let M be the total number of microphones over all robots (the number of channels of the big microphone array), N the number of sound sources, and R the number of robots. The optimization problem is defined as follows:

- **Input**: $\boldsymbol{x}(t) = [x_1(t), \cdots, x_M(t)]^T \in \mathbb{R}^M$
  M-channel audio signals recorded by using the M-channel big reconfigurable microphone array
- **Output**:
  (1) $\boldsymbol{y}(t) = [y_1(t), \cdots, y_N(t)]^T \in \mathbb{R}^N$
  N separated signals corresponding to sound sources
  (2) $\boldsymbol{A}^* = [\boldsymbol{a}_1^*, \cdots, \boldsymbol{a}_R^*] \in \mathbb{R}^{R\times 2}$
  The optimized positions of multiple mobile robots
- **Assumptions**:
  All microphones are synchronized and the correct positions of sound sources $B = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_N] \in \mathbb{R}^{N\times 2}$ are already estimated by using a triangulation method [17].

B. Mixing Process

We explain how observed signals $\boldsymbol{x}(t)$ are associated with source signals $\boldsymbol{s}(t) = [s_1(t), \cdots, s_N(t)]$, where $s_n(t)$ is the signal of the n-th sound source. Suppose that neither noise nor reverberation exists and that sound propagation can be represented as a linear time-invariant system as follows:

$$\boldsymbol{x}(\omega) = \boldsymbol{H}(\omega)\boldsymbol{s}(\omega), \tag{1}$$

where $\boldsymbol{x}(\omega) = [X_1(\omega), \cdots, X_M(\omega)]^T \in \mathbb{C}^M$ is the spatial spectrum of the observed signals at frequency ω, $\boldsymbol{s}(\omega) = [S_1(\omega), \cdots, S_N(\omega)]^T \in \mathbb{C}^N$ is that of the source signals at frequency ω, and $\boldsymbol{H}(\omega) \in \mathbb{C}^{M\times N}$ is a mixing matrix. $X_m(\omega)$ is the Fourier transform of the observed signal $x_m(t)$

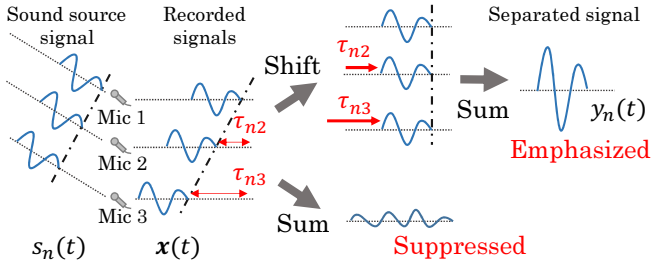


Fig. 2. Overview of delay-and-sum beamforming (DSBF)

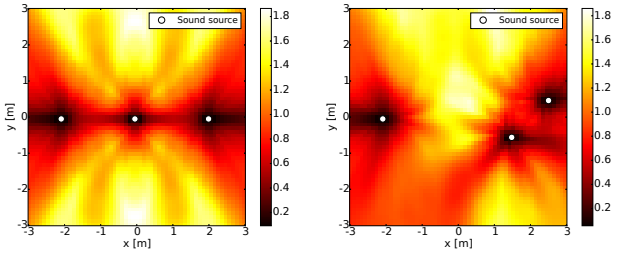


Fig. 3. The examples of the objective function at each position. Circles indicate the sound source positions. The position with the high function value is good.

and $S_n(\omega)$ is that of the source signal $s_n(t)$. The relationship between $X_m(\omega)$ and $S_n(\omega)$ is given by

$$X_m(\omega) = \sum_{n=1}^{N} \frac{1}{d_{nm}} S_n(\omega) e^{-j\omega\tau_{nm}}, \tag{2}$$

where d_{nm} is the distance between the n-th sound source and the m-th microphone and τ_{nm} is the delay time of the m-th observed signal $x_m(t)$ from the n-th source signal $s_n(t)$, *i.e.*, $x_m(t) = s_n(t - \tau_{nm})$. $1/d_{nm}$ indicates the amplitude decay (the amplitude of a propagated signal is inversely proportional to the distance). Note that τ_{nm} can be calculated in advance according to the positional relationship between the robot and the source as $\tau_{nm} = d_{nm}/c$ (c is the speed of sound). Comparing Eq. (1) with Eq. (2), we get

$$h_{mn}(\omega) = \frac{1}{d_{nm}} e^{-j\omega\tau_{nm}}. \tag{3}$$

C. Filtering Process

We explain how separated signals $\boldsymbol{y}(t)$ are obtained from observed signals $\boldsymbol{x}(t)$. As in the mixing process, we assume that $\boldsymbol{y}(t)$ can be represented as a linear system as follows:

$$\boldsymbol{y}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega), \tag{4}$$

where $\boldsymbol{y}(\omega) = [Y_1(\omega), \cdots, Y_N(\omega)]^T \in \mathbb{C}^N$ is the spatial spectrum of the separated signals at frequency ω and $\boldsymbol{W}(\omega) \in \mathbb{C}^{N\times M}$ is a filtering matrix. Here Eqs. (1) and (4) indicate that if $\boldsymbol{W}(\omega) = \boldsymbol{H}(\omega)^{-1}$, the separated signals are equal to the true source signals, *i.e.*, $\boldsymbol{y}(\omega) = \boldsymbol{H}(\omega)^{-1}\boldsymbol{x}(\omega) = \boldsymbol{H}(\omega)^{-1}\boldsymbol{H}(\omega)\boldsymbol{s}(\omega) = \boldsymbol{s}(\omega)$.

We use a standard source separation method called delay-and-sum beamforming (DSBF) for estimating the filtering matrix $\boldsymbol{H}(\omega)$. It focuses on the time differences of arrivals (TDOAs) of a source signal at the microphones (Fig. 2). To obtain the separated signal $y_n(t)$ corresponding to the n-th source, each observed signal $x_m(t)$ is time-shifted by the corresponding TDOA τ_{nm} and then all the shifted signals are summed up. The shifting operation aligns the phases of the target source signal and cancels out the phases of other sounds. This emphasizes only the target source signal and suppresses other sounds. The equivalent frequency-domain representation of DSBF is given by

$$Y_n(\omega) = \sum_{m} \frac{1}{d_{nm}} X_m(\omega) e^{j\omega\tau_{nm}}, \tag{5}$$

where $1/d_{nm}$ is a weighting coefficient. We put more emphasis on the observed signal recorded by a microphone that is closer to the target sound source.

In order to take advantage of using multiple mobile robots, a set of the distributed microphone arrays is regarded as one big microphone array. This means that all the observed signals recorded by the robots are used for cooperative sound source separation. Comparing Eq. (4) with Eq. (5), we get

$$w_{nm}(\omega) = \frac{1}{d_{nm}} e^{j\omega\tau_{nm}}. \tag{6}$$

D. Objective Function

We define the objective function that should be maximized for layout optimization as the harmonic mean of the gains obtained by DSBF. Let $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_R]$ be a set of the positions of R robots. The objective function $f(\boldsymbol{A})$ w.r.t. $\boldsymbol{A}$ is defined as follows:

$$f(\boldsymbol{A}) = \frac{N}{\sum_{n=1}^{N} \frac{1}{g_n(\boldsymbol{A})}}, \tag{7}$$

where $g_n(\boldsymbol{A})$ is the gain of the n-th sound source signal in the n-th separated signal. A reason why the harmonic mean is used instead of the standard average is that we aim to find the layout that enables high-quality source separation such that the gains are balanced over all sound sources. If one of the source signals is poorly estimated, the value of the objective function is significantly decreased.

Using Eqs. (1) and (4), the relationship between the separated signals $\boldsymbol{y}(t)$ and the source signals $\boldsymbol{s}(t)$ is represented in the frequency domain as follows:

$$\boldsymbol{y}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{s}(\omega), \tag{8}$$

where $\boldsymbol{A}(\omega) \in \mathbb{C}^{N\times N}$ is a gain matrix obtained by $\boldsymbol{A}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{H}(\omega)$. If $\boldsymbol{A}(\omega) = \boldsymbol{I}$ is achieved, the separated signals are equal to the true source signals (perfect separation). In reality, $\boldsymbol{A}(\omega)$ has off-diagonal elements that represent the crosstalks between the source signals. Therefore, the gain of the n-th source signal at frequency ω is given by

$$g_n(\boldsymbol{A}, \omega) = \frac{a_{nn}(\omega)}{\sum_{n\neq k} a_{nk}(\omega)}, \tag{9}$$

where $a_{nn}(\omega)$ and $a_{nk}(\omega)$ represent the weight of the n-th source signal and that of the k-th source signal in the n-th separated signal, respectively. In this paper, we take the

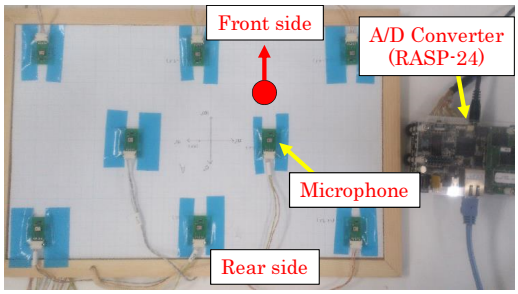


Fig. 4. The layout of an 8-channel microphone array on each mobile robot.

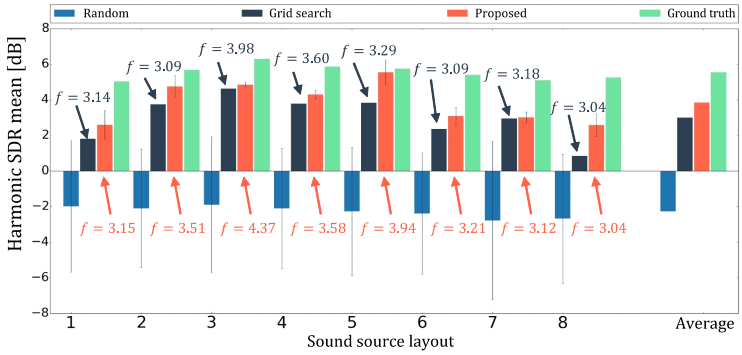


Fig. 5. The harmonic mean of SDRs for each layout of sound sources in a simulated room. f means the objective function value.

average of the gains over all frequency bands and define $g_n(\boldsymbol{A})$ as follows:

$$g_n(\boldsymbol{A}) = \frac{\sum_\omega a_{nn}(\omega)}{\sum_{n \neq k} \sum_\omega a_{nk}(\omega)}, \tag{10}$$

When DSBF is used for sound source separation, $a_{nk}(\omega)$ is obtained by using Eqs. (3) and (6) as follows:

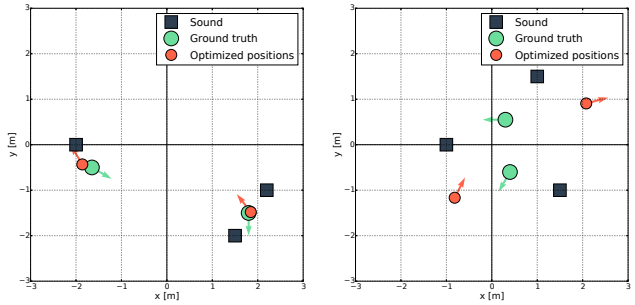
$$a_{nk}(\omega) = \left| \sum_{m=1}^{M} \frac{1}{d_{nm} d_{km}} \exp(j\omega(\tau_{nm} - \tau_{km})) \right|. \tag{11}$$

The frequency bins from 1 [Hz] to 8000 [Hz] (L bins) are taken into account as the range of ω. Figure 3 shows the values of the objective function in a 6 [m] square room when a single robot with an 8-channel microphone array is used for separating three sound sources. The function takes small values in several cases. If multiple sound sources are in the same direction (the robot and the sound sources get into line), the TDOAs at the microphones are close to each other and non-target sound sources are scarcely suppressed. If the robot is excessively close to sound sources, the separation performance of the other sources is critically degraded although the close sources can be accurately separated. The objective function thus takes a small value because it is defined as the harmonic mean of the gains over all sources.

E. Layout Optimization

We use a genetic algorithm (GA) for optimizing the layout of multiple mobile robots. This is because if a grid search algorithm is used, the computational cost exponentially increases as the number of robots increases. In the context of GA, candidate layouts are often called *creatures*. There are two types of creation of next-generation creatures: small modification of the current generation with a high probability (*crossover*) and drastic change from the current generation with a low probability (*mutation*). After creating a certain number of creatures, the objective function is calculated for each creature and creatures are selected with a probability based on the function values. This process is repeated until a certain termination condition is met.

In this paper, crossover is achieved by randomly moving robots to nearby positions and turning the robots to random directions. Mutation is achieved by randomly choosing the positions and directions of robots from a test environment. The objective function is defined as the harmonic mean of the gains obtained by simulating delay-and-sum beamforming from a possible layout. The creatures of a new generation are selected according to elitist selection or roulette-wheel selection. In elitist selection, creatures with larger function values are selected from the top of the ranking. In roulette-wheel selection, the creatures are selected with probabilities proportional to the values of the objective function, and hence the creatures with lower function values are selected with low probabilities. If a fixed number of generations is reached, a creature with the highest function value is selected as the optimal creature (optimal layout of the robots).



(a) Sound source layout 3 (b) Sound source layout 8

Fig. 6. Example of robot layout optimization. Red and green circles indicate the positions obtained by the proposed method and the ground truth positions, respectively.

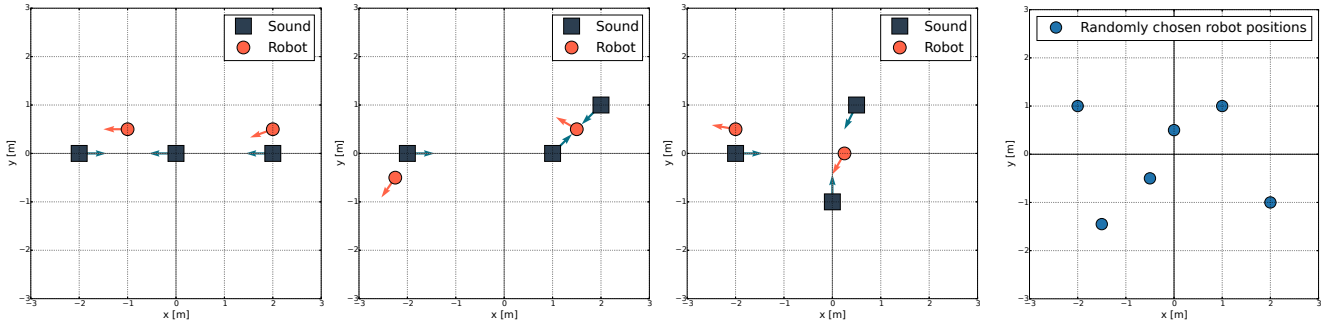
IV. Experimental Evaluation

This section reports experiments conducted to evaluate the improvement of source separation performance in simulated and real rooms. If the number of robots is more than that of sound sources, the optimal layout is trivial, *i.e.*, the best performance can be achieved by moving a robot to each sound source. We therefore assume that the number of robots is less than that of sound sources.

A. Experiment in a Simulated Room

We conducted an experiment in a simulated room to evaluate the effectiveness of the proposed method.

1) Experimental Conditions: We supposed that there were three sound sources and two robots, each of which had an 8-channel microphone array (Fig. 4), in a room of 6 m^2 ($M =$



(a) Sound source layout 1 (b) Sound source layout 2 (c) Sound source layout 3

Fig. 7. The results of robot layout optimization for three source layouts. Circles and squares indicate the positions of robots and sound sources, respectively. Arrows indicate the directions of robots and sound sources.

Fig. 8. Randomly chosen candidate positions of the robots.

16, $N = 3$, and $R = 2$). Eight layouts of sound sources were tested. In the layouts 1 and 2, the three sound sources got into line. In the layouts 3, 4, and 5, two sound sources were close to each other and the remainder was distant. In the layouts 6, 7, and 8, the three sound sources were distant from each other. Source signals were randomly selected from JNAS phonetically-balanced Japanese utterances [24]. The observed signal of each microphone was synthesized by convoluting an geometrically-calculated impulse response.

We compared the proposed method that uses the GA for layout optimization with a random method that randomly chooses the layout of the robots and a grid search method that finds the layout that maximizes the objective function by using a grid search algorithm, and calculated a ground truth layout that maximizes the harmonic mean of SDRs by using a grid search algorithm under a condition that sound source signals are known. The configuration of the GA was that the number of creatures of each generation was 1000 and the GA stopped when the 200th generation was reached. In the grid search algorithm, the interval of grid points was 0.2 [m] and the direction of each robot was selected from $0°, 45°, 90°$, and $135°$, because the layout of an 8-ch microphone array we used was symmetric.

The separation performance was measured with the harmonic mean of the signal-to-distortion ratios (SDRs) for the separated signals corresponding to the three sound sources. The SDR is the ratio of a target signal to the other sounds in a separated signal. A higher SDR means better separation performance [25], [26]. Since the proposed and random methods involve randomness, we ran 30 trials and calculated the average of the harmonic mean of SDRs.

2) Experimental Results: Figure 5 shows the experimental results indicating the harmonic mean of SDRs for the eight layouts of sound sources and f means the objective function value. In all cases, the SDRs obtained by the proposed method were superior to those obtained by the random method by 6.0 dB on average. Comparing the proposed method with the grid search method, in all cases, the objective function values obtained by the GA were almost the same or larger than those obtained by the grid search, though the computational cost of the GA (1000 creatures $\times$ 200 generations $=$ 0.2 million layouts) is smaller than that of the grid search algorithm ($(31$ grid points $\times$ 31 $\times$ 4 directions$)^2 \approx 14$ million layouts). This was because the grid search algorithm calculates the objective function at only grid points and cannot calculate all the possible layouts and directions of robots. This indicates that GA is suited for the layout optimization.

Figure 6 shows an example of the results of the robots layout optimization. As shown in Fig. 6(a) (the case of the sound source layout 3), the robot layout obtained by the proposed method was close to the ground truth layout and the source separation performance at the layout obtained by the proposed method was as high as that at the ground truth layout. Fig. 6(b) (the case of the sound source layout 8), on the other hand, shows that the layout obtained by the proposed method was significantly different from the ground truth one, and the harmonic mean of SDRs of the proposed method was lower than that of the ground truth by 2.7 [dB]. In addition, comparing the proposed method with the grid search method, though the objective function values were almost the same, the harmonic mean of SDRs of the proposed method was higher than that of the grid search method. A main reason was that although the amplitudes of the source signals varied according to frequency bands, the proposed method put the same emphasis on all frequency bands. To improve the source separation performance, we plan to consider the frequency characteristics of sound sources.

B. Experiment in a Real Room

We conducted an experiment using real recordings to evaluate the effectiveness of the proposed method.

1) Experimental Conditions: Three sound sources and two robots, each of which had an 8-channel microphone array (Fig. 4), were put in a wide room with a reverberation time (RT_{60}) of 800 ms ($M = 16$, $N = 3$, and $R = 2$). The source signals used in this experiment were the same as those used in the simulated experiment (Section IV-A). Three layouts of sound sources were tested (Fig. 7). In this experiment, the directions of speakers were set as shown in Fig. 7 because speakers actually have directivity. To adjust the height of each microphone array to the sound sources, the microphone array was attached to a pole (Fig. 9). An impulse response was actually measured for each microphone. The

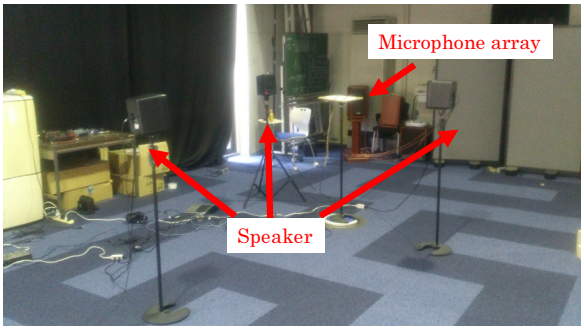


Fig. 9. Measuring real impulse responses at each position in an environmental room.

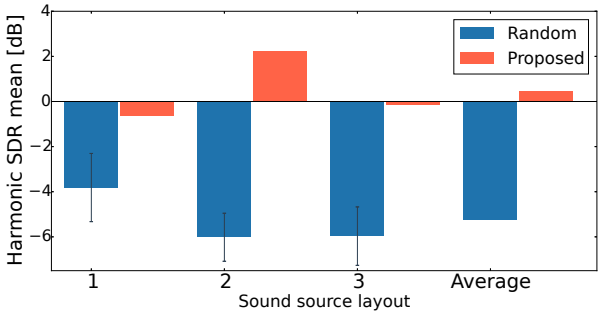


Fig. 10. The harmonic mean of SDRs for each layout of sound sources in an experimental room.

observed signals were synthesized by convoluting the real impulse responses of the corresponding positions with the source signals. Note that those synthesized signals can be considered to be quite similar to real recordings. The microphone array equipped on the robot was synchronized by using a multichannel A/D converter (RASP-24 manufactured by Systems In Frontier Corp) with a sampling rate of 16 kHz and a quantization of 16 bits (Fig. 4).

We compared the proposed method with a random method that randomly chooses two positions from six candidate positions. These candidates in the room were chosen randomly, as shown in Fig. 8. The source separation performance was evaluated as in the simulated experiment (Section IV-A).

2) Experimental Results: Figure 10 shows the experimental results obtained by the random method and those obtained by the proposed method. In all sound source layouts, the proposed method achieved better SDRs by 5.7 dB on average. The proposed method scored particularly well in the sound source layout 2 with an improvement of 8.2 dB. This is because, taking advantage of using two robots, the robot on the right mainly recorded the two right-side sound sources and the robot on the left mainly recorded the left-side sound source. Therefore, the separation performance of all sound sources was significantly improved.

In comparison with the experiment in the simulated room, the harmonic means of SDRs were significantly degraded in both methods. There are two main reasons. Since delay-and-sum beamforming is formulated in an ideal condition, the separation performance was strongly influenced by reverberation and by errors related to the positions of the sound sources and the robots. In addition, the directivity of a sound source matters. In the experiment in the simulated room, we assumed that sound sources had no directivity. In fact, however, real sound sources did have directivity, and the time differences of arrivals differed according to the directions of sound sources.

A main limitation of the proposed method is that an objective function cannot be defined appropriately if sophisticated methods such as ICA, IVA, and GHDSS [7]–[9] are used for source separation. These methods aim to estimate a separation matrix $\boldsymbol{W}(\omega)$ such that it is as close as possible to the inverse of a mixing matrix $\boldsymbol{H}(\omega)$ (*i.e.*, $\boldsymbol{W}(\omega)\boldsymbol{H}(\omega) \approx \boldsymbol{I}$), and theoretically the gain then becomes infinite in any layout. Nonetheless, the optimal layout obtained by the proposed method based on the DSBF-based objective function could be expected to improve the performance of sophisticated separation methods. As the error of sound source localization has an influence on the source separation performance, the objective function could be improved by focusing on the ambiguity of sound source localization.

In order to solve the directivity problem, robots should move to the front side of a sound source. This is because the TDOA is different from the expected TDOAs at the sides and rear of a sound source due to diffraction and reverberation. A promising solution here would be to estimate the directions of sound sources by audio-visual integration and to use an objective function that takes directivity into account.

V. Conclusion

This paper presented an active-audition method that optimizes the layout of multiple mobile robots for cooperative sound source separation in an environment with multiple sound sources. To take advantage of using multiple mobile robots, when robots separate the recorded signals, multiple microphones are regarded as one big microphone array, and each recorded signal is given a weight inversely proportional to the distance between a sound source and the microphone. The optimal layout is determined by theoretically predicting the performance of source separation (gain) based on delay-and-sum beamforming from a possible layout. We conducted two experiments to evaluate the performance of the proposed method in simulations and in a real environment. Compared with a random method that randomly chooses the positions of robots, the source separation performance was improved by 6.0 dB on average in simulations and by 5.7 dB on average in a real environment.

We plan to make the proposed method applicable to more sophisticated source separation methods, since in the current method, in theory the gain of each source signal becomes infinite for any robot layout if those separation methods are used. To solve this problem, the objective function needs to consider the error of sound source localization. We also plan to estimate sound source positions, robot positions, and optimal positions simultaneously via reinforcement learning in order to remove the assumption that robots and sound source positions are given.

ACKNOWLEDGMENT

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

REFERENCES

[1] K. Wang *et al.*, "The SLAM algorithm of mobile robot with omnidirectional vision based on EKF," in *ICIA*, 2012, pp. 13–18.

[2] H. Kretzschmar *et al.*, "Efficient information-theoretic graph pruning for graph-based SLAM with laser range finders," in *IROS*, 2011, pp. 865–871.

[3] T. Nguyen *et al.*, "Interactive syntactic modeling with a single-point laser range finder and camera," in *ISMAR*, 2013, pp. 107–116.

[4] H. Miura *et al.*, "SLAM-based online calibration of asynchronous microphone array for robot audition," in *IROS*, 2011, pp. 524–529.

[5] H. Sun *et al.*, "Microphone array based auditory localization for resque robot," in *CCDC*, 2011, pp. 606–609.

[6] J. Hu *et al.*, "Simultaneous localization of a mobile robot and multiple sound sources using a microphone array," *J. Advanced Robotics*, vol. 25, no. 1-2, pp. 135–152, 2011.

[7] S. Makino *et al.*, "Blind source separation of convolutive mixtures of speech in frequency domain," *IEEE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. 88, pp. 1640–1655, 2005.

[8] 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.

[9] I. Lee *et al.*, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *J. Signal Processing*, vol. 87, no. 8, pp. 1859–1871, 2007.

[10] J. Hao *et al.*, "Independent vector analysis for source separation using a mixture of gaussians prior," *J. Neural Computatation*, vol. 22, no. 6, pp. 1646–1673, 2010.

[11] H. Saruwatari *et al.*, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 666–678, 2006.

[12] K. Nakadai *et al.*, "Real-time sound source localization and separation for robot audition," in *ICSLP*, 2002, pp. 193–196.

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

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

[15] H. Kim *et al.*, "Human-robot interaction in real environment by audio-visual integration," *J. Control, Automation and Systems*, vol. 5, no. 1, pp. 61–69, 2007.

[16] G. L. Reid and E. Milios, "Active stereo sound localization," *J. Acoustical Society of America*, vol. 113, no. 1, pp. 185–193, 2003.

[17] Y. Sasaki *et al.*, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *IROS*, 2006, pp. 380–385.

[18] K. Nakadai *et al.*, "Active audition for humanoid," in *AAAI*, 2000, pp. 832–839.

[19] E. Berglund and J. Sitte, "Sound source localisation through active audition," in *IROS*, 2005, pp. 509–514.

[20] T. Yoshida and K. Nakadai, "Active audio-visual integration for voice activity detection based on a causal Bayesian network," in *Humanoids*, 2012, pp. 370–375.

[21] Y. Sasaki *et al.*, "32-channel omni-directional microphone array design and implementation," *J. Robotics and Mechatronics*, vol. 23, no. 3, pp. 378–385, 2011.

[22] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[23] N. Mitianoudis and M. Davies, *Independent Component Analysis and Blind Signal Separation*. Springer, 2004, vol. 3195, ch. Permutation Alignment for Frequency Domain ICA Using Subspace Beamforming Methods, pp. 669–676.

[24] Y. Sagisaka and N. Uratani, "ATR spoken language database," *J. The Acoustic Society of Japan*, vol. 48, no. 12, pp. 878–882, 1992.

[25] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[26] C. Raffel *et al.*, "mir_eval: A transparent implementation of common MIR metrics," in *ISMIR*, 2014, pp. 367–372.