Paper:

Layout Optimization of Cooperative Distributed Microphone Arrays Based on Estimation of Source Separation Performance

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

Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606-8501, Japan
E-mail: {sekiguch, yoshiaki, itoyama, yoshii}@kuis.kyoto-u.ac.jp

[Received July 20, 2016; accepted November 4, 2016]

The active audition method presented here improves source separation performance by moving multiple mobile robots to optimal positions. One advantage of using multiple mobile robots that each has a microphone array is that each robot can work independently or as part of a big reconfigurable array. To determine optimal layout of the robots, we must be able to predict source separation performance from source position information because actual source signals are unknown and actual separation performance cannot be calculated. Our method thus simulates delay-and-sum beamforming from a possible layout to calculate gain theoretically, i.e., the expected ratio of a target sound source to other sound sources in the corresponding separated signal. Robots are moved into the layout with the highest average gain over target sources. Experimental results showed that our method improved the harmonic mean of signal-to-distortion ratios (SDRs) by 5.5 dB in simulation and by 3.5 dB in a real environment.

Keywords: cooperative source separation, multiple mobile robots

1. Introduction

Computational auditory scene analysis based on information obtained using mobile robots equipped with microphones has been studied actively in recent years [1]. Robots recognize what an individual says in a noisy environment such as a crowded event site by microphone array processing such as sound source localization or separation. Since a robot equipped with a microphone array can estimate the direction of sound sources, two-dimensional (2D) positions of sound sources can be estimated at one time by using multiple robots and triangulation [2, 3]. Moreover, multiple robots can conduct cooperative source separation by regarding multiple robots as a single microphone array [4].

Source separation performance is related to the positional relationship between robots and sound sources [5]. For example, the performance is degraded when there are multiple sound sources in the same direction. This problem can be solved by moving the robots to better positions. If a robot cannot perform speech recognition correctly due to the bad positional relationship between the robot and sound sources or many noise sources, the robot can achieve this by giving some instructions to other robots near itself to move to better positions and cooperating with them. A major problem the robots face is determining the optimal robot layout to maximize separation performance. Actual source separation performance cannot be calculated because the true source signals are unknown. It is thus necessary to predict the source separation performance before actually moving the robots. Another problem in cooperative source separation is that the synchronization between each robot and the positions of each sound source and robot are necessary. Although these are difficult to estimate without a special device such as a GPS, several studies have explored the feasibility of estimating them simultaneously without a special device by using a SLAM framework [6, 7].

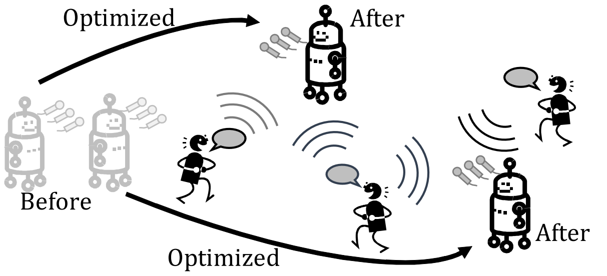


Fig. 1. Optimizing the layout of multiple mobile robots for cooperative source separation that regards multiple mobile robots equipped with microphone arrays as one big microphone array.

We propose active audition that optimizes the positions of multiple robots by simulating delay-and-sum beamforming (DSBF) from a possible layout under the condition that positions of sound sources are already known (**Fig. 1**). When DSBF is used to separate sources, gain, which is the expected ratio of a target sound source and the other sound sources in the corresponding separated signal, is calculated from the positions of sound sources and robots. Although in the conventional method [4] robots are moved to the positions that are optimal for separating all sound sources, in many cases, robots need to focus on particular sound sources. The proposed method enables robots to focus on only the target sound sources, which are selected by a user or robots, by using a cost function calculated from the gains of only the target sources. Since it is often difficult to find the layout with the highest gain via local search, we use a genetic algorithm (GA) to avoid getting stuck at a local optimum. In an experimental evaluation of the source separation performance, we tested not only the DSBF but also geometrically constrained independent component analysis (GICA) and geometrically constrained high-order decorrelation-based source separation (GHDSS) [8].

2. Related Work

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

2.1. Active Audition

Active audition is a technique that aims to improve the performance of auditory scene analysis (analysis of surrounding sound objects) by effectively using the movement of a robot equipped with microphones. Several studies have tried estimating the directions of sound sources by turning the head of a humanoid robot. Nakadai et al. [9], for example, developed a humanoid robot that has two microphones and tracks sound source directions by integrating audio, visual, and motor control. Berglund and Sitte [10] developed a robot with two microphones that learns how to orient itself toward a sound source via reinforcement learning. Kim et al. [11] proposed reducing errors in sound source localization by accounting for the results of voice activity detection (VAD) and face tracking.

Active audition has often been used with a single moving robot to estimate the position of a sound source. Reid and Milios [12], for example, developed a robot that estimates the 3D position of a sound source by moving two microphones. Sasaki et al. [13] developed a mobile robot that has a microphone array and estimates the positions of multiple sound sources. Since the robot can move around sound sources, the positions of these sources are determined from source direction estimated from different observation positions in a way of triangulation. Yoshida and Nakadai [14] integrated the audio, visual, and active motion functions of a robot to estimate how active motion affects VAD.

If multiple robots are used, the positions of sound sources can be obtained quickly without moving any robots. Martinson et al. [2] optimized multiple robot layout to improve the performance of sound source localization in 2D space. Individual robots were equipped with a single microphone and the optimal layout was determined so that robots were distant from both other robots and obstacles and close to sound sources.

2.2. Sound Source Separation

DSBF is basic microphone-array-based source separation [13, 15]. Sasaki et al. [15] attempted to optimize a 32-channel microphone array layout to improve the source separation performance of DSBF. The optimal layout was determined to have high directivity to all directions.

Independent component analysis (ICA), another widely used source separation technique, discovers statistically independent source signals from given mixed signals [16, 17]. Time-domain ICA separates convolutionally mixed signals, but its computational costs are high. This is lowered by using graphical processing units (GPUs) to calculate in parallel [18]. Frequency-domain ICA, however, conducts standard ICA at individual frequency bands and is more efficient. Since frequency-domain ICA has problems of frequency-band permutation and scaling, many studies have explored ways to solve these problems [19–21].

3. Proposed Method

The method we propose here optimizes multiple mobile robot layout for cooperative source separation. Individual robots have standard microphone arrays, and multi-channel audio signals are recorded by regarding a set of the robots' distributed microphone arrays as a big microphone array. This extracts audio signals from a particular direction by using geometric-constrained independent component analysis (GICA) and geometric-constrained high-order decorrelation-based source separation (GHDSS). Since the source separation performance of these methods is high and their computational costs low, these methods are suitable for robot audition requiring real-time processing.

To optimize the robot layout, we designed an objective function to be maximized with respect to a layout. We then predicted source separation performance theoretically by simulating DSBF without actually moving any robots. Specifically, the ratio of a source signal in the corresponding separated signal (separation performance) is determined by specifying a *mixing process* that represents the propagation of source signals to microphones and a *filtering process* that represents the extraction of source signals from observed signals. We use DSBF instead of GICA or GHDSS for two reasons. One is that predicting the source separation performance is difficult when GICA or GHDSS is used, and the other is that separation performances of these methods are correlated with that of DSBF. Since it is often difficult to find the layout with the highest gain via local search, we use a genetic algorithm that tends to avoid local optima.

3.1. Problem Specification

Our goal is to find a multiple robot layout enabling the high-quality separation of all target sound sources in a test environment. Let M be the total number of microphones on robots (the number of channels of the big microphone array), N the number of sound sources, N' the number of target sound sources, and R the number of robots. We assume that sound sources and robots are on a 2D plane. The optimization problem is defined as follows:

- **Input**: $\boldsymbol{x}(t) = [x_1(t), \ldots, x_M(t)]^T \in \mathbb{R}^M$
  M-channel audio signals recorded by using an M-channel big reconfigurable microphone array, where t is the index of the sample.
- **Output**:
  (1) $\boldsymbol{y}(t) = [y_1(t), \ldots, y_{N'}(t)]^T \in \mathbb{R}^{N'}$
  N' separated signals corresponding to sound sources.

  (2) $\boldsymbol{A}^* = [\boldsymbol{a}_1^*, \ldots, \boldsymbol{a}_R^*] \in \mathbb{R}^{R \times 3}$
  Optimized 2D positions and directions of multiple mobile robots.
- **Assumptions**:
  All microphones are synchronized and the correct positions of sound sources $B = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N] \in \mathbb{R}^{N \times 2}$ are already estimated by using triangulation [13].

3.2. Mixing Process

We explain the relationship between observed signals $\boldsymbol{x}(t)$ and source signals $\boldsymbol{s}(t) = [s_1(t), \ldots, s_N(t)]$, where $s_n(t)$ is the signal of the n-th sound source. Assume that neither reverberation nor noise exists and that sound propagation is represented as the following linear time-invariant system:

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

where $\boldsymbol{x}(\omega) = [X_1(\omega), \ldots, X_M(\omega)]^T \in \mathbb{C}^M$ is the spatial spectrum of observed signals at frequency ω, $\boldsymbol{s}(\omega) = [S_1(\omega), \ldots, S_N(\omega)]^T \in \mathbb{C}^N$ is that of 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 observed signal $x_m(t)$ and $S_n(\omega)$ is that of source signal $s_n(t)$. The relationship between $X_m(\omega)$ and $S_n(\omega)$ is given by

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

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 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 propagation distance). Note that τ_{nm} is calculated in advance based on the positional relationship between the sound source and robot as $\tau_{nm} = d_{nm}/c$, where c is the speed of sound. Comparing Eqs. (1) and (2), we get

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

3.3. Filtering Process

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

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

where $\boldsymbol{W}(\omega) \in \mathbb{C}^{N \times M}$ is a separation matrix, and $\boldsymbol{y}(\omega) = [Y_1(\omega), \ldots, Y_N(\omega)]^T \in \mathbb{C}^N$ is the spatial spectrum of separated signals at frequency ω. Eqs. (1) and (4) indicate that if $\boldsymbol{W}(\omega) = \boldsymbol{H}(\omega)^{-1}$, separated signals are equal to true source signals, i.e., $\boldsymbol{y}(\omega) = \boldsymbol{H}(\omega)^{-1}\boldsymbol{x}(\omega) = \boldsymbol{H}(\omega)^{-1}\boldsymbol{H}(\omega)\boldsymbol{s}(\omega) = \boldsymbol{s}(\omega)$. Null beamforming is a source separation method that uses the inverse of the mixing matrix as the separation matrix. Since the actual mixing matrix is not available in practice, the mixing matrix is prepared in advance, for example, by recording a time-stretched pulse (TSP) signal in an anechoic chamber. The performance of null beamforming, however, largely deteriorates when the actual and prepared mixing matrices differ due to reverberation or source localization error, etc., and is hardly ever used.

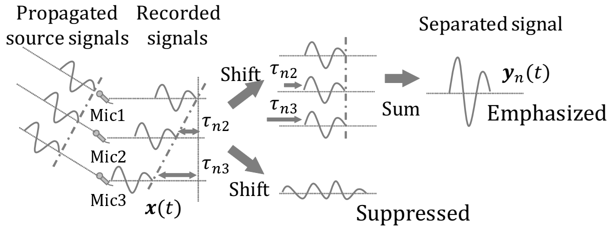


Fig. 2. Overview of delay-and-sum beamforming (DSBF).

DSBF is a standard source separation method that uses only time differences of arrivals (TDOAs) of a source signal at microphones (**Fig. 2**). Separated signal $y_n(t)$ corresponding to the n-th source is obtained by time-shifting each observed signal $x_m(t)$ by the corresponding TDOA τ_{nm} and then summing up all shifted signals. The shifting operation aligns phases of target source signals included in recorded signals and cancels out phases of other sound sources. This emphasizes the target sound source and suppresses other sounds. The equivalent frequency-domain representation of DSBF is given by

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

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

The source separation performance of DSBF is degraded when actual TDOAs differ from expected TDOAs due to reverberation, diffraction, or estimation error of the sound source direction. To solve this problem, adaptive beamforming methods have been developed, e.g., GICA and GHDSS. These methods use both TDOAs and the properties of source signals.

GICA, based on frequency-domain ICA, estimates the separation matrix online so that the independence of separated source signals becomes high. Permutation and scaling problems are solved by using geometrical restrictions.

The following two cost functions are used to estimate the separation matrix $\boldsymbol{W}$

$$\boldsymbol{J}_{\mathrm{ICA}}(\boldsymbol{W}) = \int p(\boldsymbol{y}) \log \frac{p(\boldsymbol{y})}{\prod_k p(y_k)} dy \quad . \quad . \quad . \quad . \quad . \quad (6)$$

$$\boldsymbol{J}_{\mathrm{GC}}(\boldsymbol{W}) = \|\boldsymbol{WH} - \boldsymbol{I}\|^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

where $p(\boldsymbol{y}) = p(y_1, \ldots, y_N)$ is the joint distribution of all separated signals. $\boldsymbol{J}_{\mathrm{ICA}}(\boldsymbol{W})$, the KL-divergence between $p(\boldsymbol{y})$ and $p(y_1, \ldots, y_N)$, becomes small when $p(\boldsymbol{y})$ is close to $\prod_k p(y_k)$. It follows that $\boldsymbol{J}_{\mathrm{ICA}}(\boldsymbol{W})$ is a measure of the independence of separated signals. $\boldsymbol{J}_{\mathrm{GC}}(\boldsymbol{W})$, the geometric restriction, becomes small when $\boldsymbol{W}$ is close to the inverse of mixing matrix $\boldsymbol{H}^{-1}$. In practice, $\boldsymbol{H}$ is unknown, however, so it is calculated by using the impulse response recorded beforehand or simulated from the positions of microphones. We want to conduct real-time source separation, so $\boldsymbol{W}$ is updated sequentially using the update equation given by

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \alpha \boldsymbol{J}'_{\mathrm{ICA}} - \beta \boldsymbol{J}'_{\mathrm{GC}}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

where α and β are step-size parameters, and $\boldsymbol{J}'_{\mathrm{ICA}} = \nabla_{\boldsymbol{W}*} \boldsymbol{J}_{\mathrm{ICA}}$ and $\boldsymbol{J}'_{\mathrm{GC}} = \nabla_{\boldsymbol{W}*} \boldsymbol{J}_{\mathrm{GC}}$.

GHDSS is similar to GICA but differs in that it uses a high-order correlation instead of an independency as the cost function. Its cost function is defined as follows:

$$\boldsymbol{J}_{HDSS}(\boldsymbol{W}) = \|E[\boldsymbol{E}_\phi]\|^2, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

$$\boldsymbol{E}_\phi = \phi(\boldsymbol{y})\boldsymbol{y}^H - \mathrm{diag}[\phi(\boldsymbol{y})\boldsymbol{y}^H], \quad . \quad . \quad . \quad (10)$$

$$\phi(y_i) = \tanh(\eta |y_i|) e^{j\theta(y_i)}, \quad . \quad . \quad . \quad . \quad . \quad (11)$$

where $E[]$ denotes the expectation operator, and $\boldsymbol{E}_\phi$ is a high-order correlation matrix, and η is a scaling parameter.

The set of distributed microphone arrays is regarded as a single microphone array, meaning that all observed signals recorded by robots are used for cooperative source separation. Comparing Eqs. (4) and (5), we get

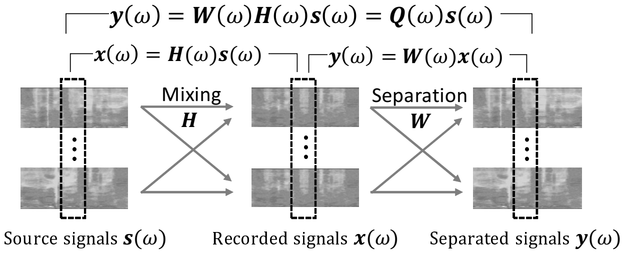
$$w_{nm}(\omega) = \frac{1}{d_{nm}} e^{j\omega\tau_{nm}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

3.4. Objective Function

We define an objective function to be maximized for layout optimization as the harmonic mean of gains of target sound sources obtained by DSBF. Let $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_R] \in \mathbb{R}^{R \times 3}$ be a set of positions and directions of R robots. Objective function $f(\boldsymbol{A})$ is defined as follows:

$$f(\boldsymbol{A}) = \frac{N'}{\displaystyle\sum_{n \in D}^{N} \frac{1}{g_n(\boldsymbol{A})}}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (13)$$

where D is a set of target sound sources and $g_n(\boldsymbol{A})$ is the gain of the n-th sound source signal. Summation of gains w.r.t. target sound sources means that the source separation performance of non-target sound sources is not considered. One reason that the harmonic mean is used instead of the standard average is that we want to find a layout that enables high-quality source separation so that gains are balanced over all target sound sources: if one of the gains of target source signals is low, the objective function value is decreased significantly.



$\boldsymbol{y}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{H}(\omega)\boldsymbol{s}(\omega) = \boldsymbol{Q}(\omega)\boldsymbol{s}(\omega)$
$\boldsymbol{x}(\omega) = \boldsymbol{H}(\omega)\boldsymbol{s}(\omega)$
$\boldsymbol{y}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega)$
Mixing
$\boldsymbol{H}$
Separation
$\boldsymbol{W}$
Source signals $\boldsymbol{s}(\omega)$
Recorded signals $\boldsymbol{x}(\omega)$
Separated signals $\boldsymbol{y}(\omega)$

Fig. 3. Relationship between source and separated signals.

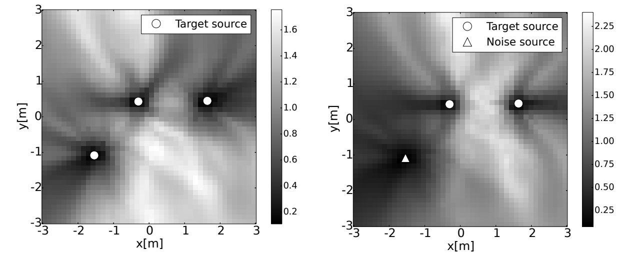


Fig. 4. Examples of the objective function at each position in a room. Circles and triangles indicate positions of target sound and noise sound. Positions with high function values are good ones.

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

$$\boldsymbol{y}(\omega) = \boldsymbol{Q}(\omega)\boldsymbol{s}(\omega), \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

where $\boldsymbol{Q}(\omega) \in \mathbb{C}^{N \times N}$ is a gain matrix obtained by $\boldsymbol{Q}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{H}(\omega)$ (as shown in **Fig. 3**). If $\boldsymbol{Q}(\omega) = \boldsymbol{I}$ is achieved, separated signals are equal to the true source signals. In practice, $\boldsymbol{Q}(\omega)$ has off-diagonal elements that represent crosstalk between source signals. The gain of the n-th source signal at frequency ω is therefore as follows:

$$g_n(\boldsymbol{A}, \omega) = \frac{q_{nn}(\omega)}{\displaystyle\sum_{k \neq n} q_{nk}(\omega)}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (15)$$

where $q_{ij}(\omega)$ is an element of the i-th row and j-th column of $\boldsymbol{Q}(\omega)$ and represent the weight of the j-th source signal in the i-th separated signal. We average gains over all frequency bands and define $g_n(\boldsymbol{A})$ as follows:

$$g_n(\boldsymbol{A}) = \frac{\sum_\omega q_{nn}(\omega)}{\displaystyle\sum_{k \neq n} \sum_{\omega} q_{nk}(\omega)}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

When DSBF is used for source separation, $q_{nk}(\omega)$ is obtained by using Eqs. (3) and (12) as follows:

$$q_{nk}(\omega) = \left| \sum_{m=1}^{M} \frac{1}{d_{nm} d_{km}} \exp(j\omega(\tau_{nm} - \tau_{km})) \right|. \quad (17)$$

Frequency bins from 1 Hz to 8000 Hz (L bins) are taken into account as the range of ω. **Fig. 4** shows values of the objective function in a room 6 m square when a single robot with an 8-channel microphone array is used to separate sound sources. In the left figure, all sound sources are target, and in the right figure, two sound sources on the right are target. The function takes small values in some cases. If multiple sound sources are in the same direction, meaning that the robot and sound sources are in a line, TDOAs at microphones are close to each other and noise sound sources are poorly suppressed. If the robot is too close to some of the sources, such sources may be separated accurately but the separation performance of other sources is degraded significantly. The objective function then takes a small value because it is defined as the harmonic mean of gains over all target sources.

When adaptive beamforming such as GICA or GHDSS is used for source separation, however, it is difficult to calculate gain, because these methods estimate separation matrix $\boldsymbol{W}(\omega)$ that is as close as possible to the inverse of mixing matrix $\boldsymbol{H}(\omega)$ (i.e., $\boldsymbol{W}(\omega)\boldsymbol{H}(\omega) \sim \boldsymbol{I}$), and gain theoretically becomes infinite in any layout. The source separation performance of GICA or GHDSS and that of DSBF have a correlation, however, and the optimal layout calculated by using the gains of DSBF is also good for source separation using GICA or GHDSS, as shown in the experiments described in Section 4.2 and 4.3.

3.5. Layout Optimization

We use GA to optimize the multiple robot layout. This is because using a grid search algorithm would increase computational cost exponentially as the number of robots increases, and if hill-climbing or gradient descent is used, the result is sometimes a local optimum (Section 4.1.2). The GA can avoid falling into a local optimum at lower computational cost than the grid search algorithm. In the GA context, candidate layouts are often called *creatures*. There are two types of creation of next-generation creatures: small modifications of the current generation with a high probability (*crossover*) and drastic changes 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 at a probability based on function values. This process is repeated until a termination condition is met.

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

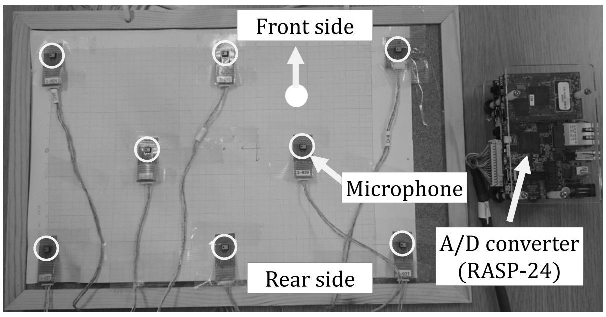


Fig. 5. 8-channel microphone array layout for each mobile robot.

4. Experimental Evaluation

We conducted experiments to evaluate improvements in source separation performance in simulated and real rooms. In a simulated room, we conducted two kinds of experiments. In one experiment, all sound sources were regarded as target sources, and in the other, three of six sources were regarded as target sources.

4.1. Optimization Method Evaluation

We conducted an experiment to compare the three optimization methods hill climbing, gradient descent, and GA.

4.1.1. Experimental Conditions

In experiments, we assumed that in a room 6 m square, there were four sound sources and two robots, each of which had an 8-channel microphone array ($M = 16$, $N = 4$, and $R = 2$). All sound sources were target sources. We tested six patterns of sound source layout. Sound source layouts were determined randomly so that the distance between each sound source pair exceeded 1 m.

We compared GA with hill-climbing and gradient descent. In the GA configuration, the number of creatures of each generation was 900 and the GA stopped when the 30th generation was reached. In the hill-climbing method, updating was as follows: (1) move all robots from current positions toward a certain direction out of the eight directions, including up and down, left and right, and slanted directions, and select the direction of each robot from 0°, 45°, 90°, and 135°, because the layout of the 8-ch microphone array we used was symmetrical (**Fig. 5**), (2) calculate the gain w.r.t. each robot layout ($(8*4)^R$ patterns) and select the optimal robot layout, and (3) repeat steps (1) and (2) until gain converges. In the gradient descent method, we updated each parameter (the x-coordinate, the y-coordinate, and the angle of each robot) one by one by using numerical differentiation.

Optimization was implemented using Python and a desktop computer with an Intel Core i7-4790 CPU (4 cores, 3.4 GHz) and 8 GB of memory. Performance measures were the processing time and average gain over 20 trials.

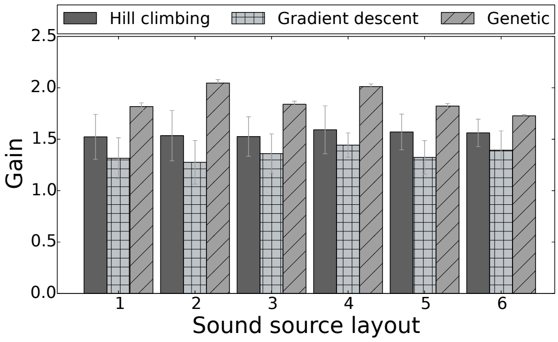


Fig. 6. Average gains for sound source layout.

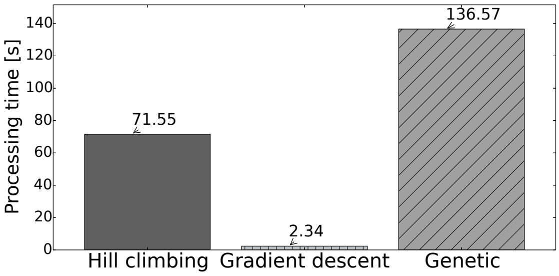


Fig. 7. Average processing time for each method.

4.1.2. Experimental Results

Figure 6 shows average gains for each sound source layout. In all cases, GA gains were higher than those of the other two optimizations, and standard deviations of GA gains were small. Gains of other methods were lower than those of GA, and standard deviations of those gains were large. This means that hill-climbing and gradient descent often fall into local optima.

Figure 7 shows processing time for individual methods. Gradient descent was 58 times faster than the GA, ending before the first GA update ended, though its performance was not good. When we must determine optimal robot positions as fast as possible, we could use gradient descent. Although hill-climbing was only about twice as fast as the GA, GA performance when hill-climbing ended was higher than that of hill-climbing, so using the GA with a smaller number of iterations is more effective than using hill-climbing. In later experiments, we used the GA as optimization.

In addition to the GA, an evolution strategy (ES) is often used to solve nonlinear optimization problems. In the ES, robot positions are updated by adding Gaussian noise $N(0, \sigma^2)$. The covariance parameter σ^2 changes based on how often the position updates succeeds. Although sampling such as Markov chain Monte Carlo (MCMC) can be used [22] to solve a nonlinear optimization problem effectively, the optimization problem must be formulated as a probabilistic problem. In the future we plan to compare these methods to the GA.

4.2. Experiment in a Simulated Room

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

4.2.1. Experimental Conditions

We arranged two types of experimental setup:

1. **Type 1:** Four sound sources and two robots were in a room 6 m square. All sources were target sources ($M = 16$, $N = 4$, and $R = 2$).

2. **Type 2:** Six sound sources and three robots were in a room 6 m square. Three of the sources were target sources, and the other three were noise sources ($M = 24$, $N = 6$, $N' = 3$, and $R = 3$).

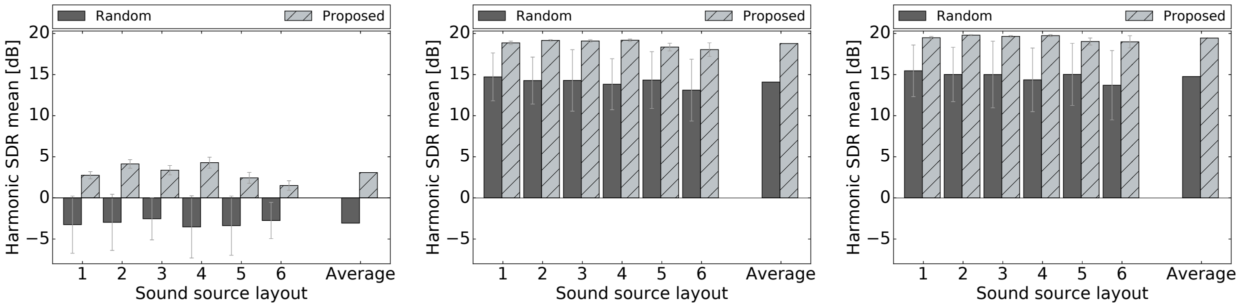
Each robot had an 8-channel microphone array (**Fig. 5**). We tested six patterns of sound source layout for each experimental setup. Type 1 sound source layout were the same as those of the experiment in Section 4.1. Type 2 sound source layout were determined randomly so that the distance between each pair of sound sources exceeded 1 m. Source signals were selected randomly from JNAS phonetically-balanced Japanese utterances [23]. The observed signal of each microphone was synthesized by convoluting geometrically calculated impulse responses that were calculated by $\frac{1}{d_{nm}} e^{-j\omega\tau_{nm}}$, where d_{nm} was the distance between the sound source n and microphone m.

Under type 1 experimental conditions, we compared the proposed method that uses the GA for layout optimization with a method that chooses the layout of robots randomly. Under type 2 experimental conditions, we compared the proposed method with the random method and the conventional method that calculates the objective function by regarding all sound sources including noise sources as target sources. In the GA configuration, each generation had 900 creatures and the GA stopped when the 30th generation was reached. With all methods, we used DSBF, GICA, or GHDSS for source separation. α and β in Eq. (8) were both set to 0.5, and η in Eq. (11) was set to 1. The initial value of separation matrix $\boldsymbol{W}$ in GICA and GHDSS was the inverse of the mixing matrix $\boldsymbol{H}^{-1}$, which was calculated from geometrically calculated impulse responses. The window size of the short-time Fourier transform was 512 samples.

Separation performance was measured with the harmonic mean of signal-to-distortion ratios (SDRs) for 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, and a higher SDR means better separation performance [24, 25]. The SDR is calculated as follows:

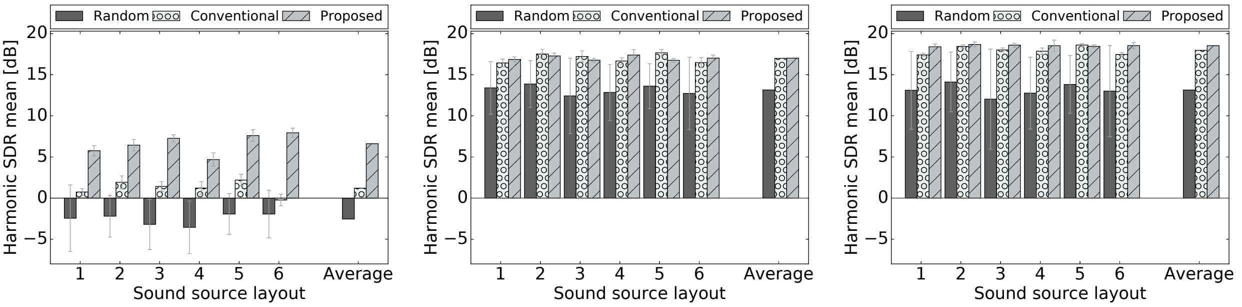
$$\mathrm{SDR} = 10 \log_{10} \frac{\|s_{\mathrm{target}}\|^2}{\|e_{\mathrm{interf}} + e_{\mathrm{noise}} + e_{\mathrm{artif}}\|^2}, \quad . \; . \; (18)$$

where s_{target} is a version of true source signal modified by an allowed distortion, and e_{interf}, e_{noise}, and e_{artif} are interferences, noise and artifact error terms. Since proposed and random methods involve randomness, we ran 20 trials and calculated the average of the harmonic mean of the corresponding SDRs.



(a) DSBF (b) GICA (c) GHDSS

Fig. 8. Harmonic mean of SDRs for each source separation method in the type 1 experimental setup in a simulated room.



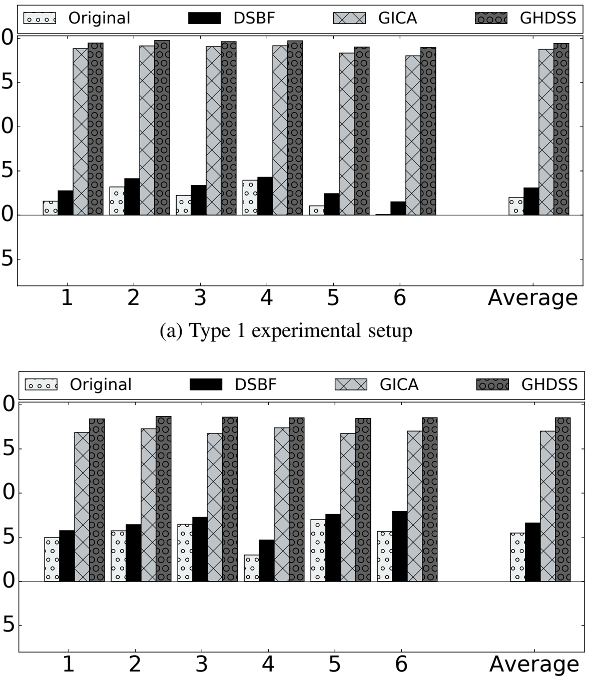
(a) DSBF (b) GICA (c) GHDSS

Fig. 9. Harmonic mean of SDRs for each source separation method in the type 2 experimental setup in a simulated room.

4.2.2. Experimental Results

Figure 8 shows experimental results obtained using the type 1 experimental setup. It shows the harmonic mean of SDRs for the six layouts of sound sources. In all cases, SDRs obtained by the proposed method were better than those obtained by the random method. The average SDR improvement of DSBF was 6.1 dB, that of GICA was 5.4 dB, and that of GHDSS was 3.0 dB.

Figure 9 shows experimental results obtained using the type 2 setup. As with the type 1 setup, the SDRs obtained by the proposed method were in all cases better than those obtained by the random method. In both setups, SDRs of GICA and GHDSS were significantly higher than those of DSBF because the correct mixing matrix was used to calculate J'_{GC} in GICA and GHDSS. The average SDR improvement of DSBF was 9.2 dB, that of GICA 2.2 dB, and that of GHDSS 3.9 dB. That of DSBF was higher than that obtained in the type 1 setup. This means that the degree of freedom of the robot layouts increased as the number of robots increased, and layout optimization had a strong impact on the source separation performance. Comparing the proposed method with the conventional method with regard to DSBF, the average SDR improvement was 5.4 dB.

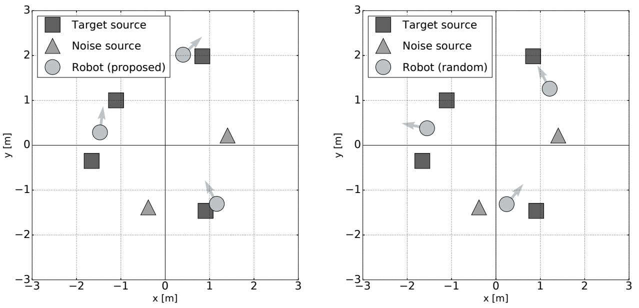


(a) Type 1 experimental setup

(b) Type 2 experimental setup

Fig. 10. Harmonic mean of SDRs of the original observed signal and that of separated signals obtained by each source separation method at optimal robot positions.

Figures 10(a) and **(b)** show the harmonic mean of SDRs of original observed signals (baseline) and that of separated signals at optimal robot layouts. The SDRs of observed signals were calculated by evaluating all of the observed signals and selecting the best SDR for each sound source. Comparing the observed signals with the separated signals of DSBF, the differences were small, because when a robot was very close to a single sound source, the SDR of the observed signal recorded by the robot with respect to the sound source was very high. When DSBF was used, the SDR of the sound source



(a) The robot positions calculated by using the proposed method

(b) The robot positions calculated by using the conventional method

Fig. 11. Example of robot layout optimization for sound source layout 2 (Type 2). Circles at left indicate positions obtained by the proposed method and circles at right indicate positions obtained by the random method.

would be worse than that of the observed signal because the recorded signals of the other robots contain the signals of noise sources and deteriorate separation performance. When there was a sound source around which no robot exists, however, the SDRs of all observed signals with respect to the sound source would be much lower than the separated signals obtained by using DSBF.

Figure 11(a) shows an example of the robot layout calculated using the proposed method in the type 2 experimental setup. The harmonic means of SDRs of DSBF, GICA, and GHDSS were 7.3 dB, 16.3 dB, and 18.3 dB, respectively. **Fig. 11(b)** shows an example of the robot layout calculated using the conventional method in the type 2 setup. The harmonic means of SDRs of DSBF, GICA, and GHDSS were 2.1 dB, 17.6 dB, 18.5 dB, respectively. Comparing these two patterns, the source separation performance of DSBF was high when robots were close to target sound sources and far away from the noise sound. The source separation performance of GICA and GHDSS, in contrast, were high when the robots could listen to all sound sources, including noise sounds. This is because adaptive beamforming methods can cancel noise by directing a null beam to them. Thus, to predict the source separation performance of GICA or GHDSS by using gain calculated by the DSBF may sometimes be difficult, even though gain correlates with the source separation performance. We thus must develop a prediction method that considers adaptive beamforming properties.

4.3. Experiment in a Real Room

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

4.3.1. Experimental Conditions

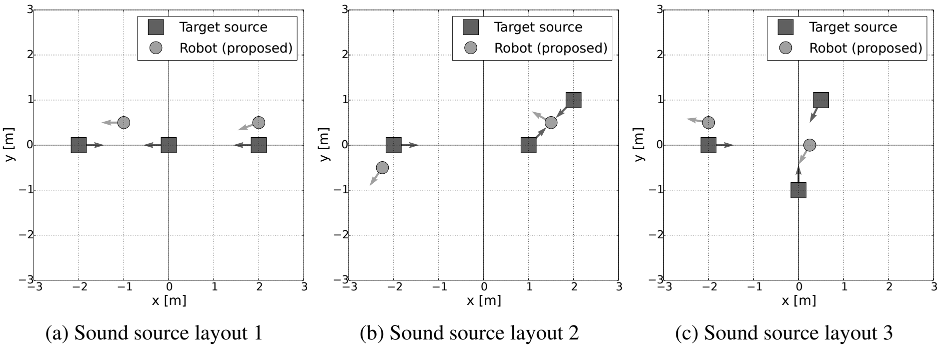
Three sound sources and two robots, each of which had an 8-channel microphone array (**Fig. 5**), were put in a wide room with a reverberation time (RT_{60}) of 800 ms ($M = 16, N = 3$, and $R = 2$). We defined all sound sources as target sound sources ($N' = 3$). Source signals were the same as those used in the simulated experiment (Section 4.2). Three layouts of sound sources were tested (**Fig. 12**). Speakers were directed as shown in **Fig. 12** because speakers actually have directivity. To adjust the height of each microphone array to sound sources, the microphone array was attached to a pole (**Fig. 13**). The microphone array on each robot was synchronized by using a multichannel A/D converter (RASP-24 manufactured by Systems In Frontier Corp) with a quantization of 16 bits and a sampling rate of 16 kHz (**Fig. 5**).

We compared the proposed method with one that randomly chooses two positions from six candidate robot positions. These candidates were chosen randomly, as shown in **Fig. 14**. An actual impulse response was recorded for each microphone at the candidate positions and at the positions calculated by using the proposed method, and observed signals were synthesized by convoluting the recorded impulse responses of the corresponding positions with the source signals. Note that these synthesized signals are considered to be quite similar to real recordings. We used DSBF, GICA, GHDSS, and null beamforming as source separation and evaluated the source separation performance as in the simulated experiment (Section 4.2). The parameters of the separation methods were the same as those used in the simulated experiment.

4.3.2. Experimental Results

Figure 15 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. The average SDR improvement of DSBF was 4.6 dB, that of GICA 6.3 dB, that of GHDSS 5.4 dB, and that of null beamforming 5.9 dB. The proposed method scored particularly well in sound source layout 2. This is because, taking advantage of using two robots, the robot on the right mainly recorded the right-side sound sources and the robot on the left mainly recorded the left-side sound source. Therefore, the separation performance for all the sources was significantly improved. **Fig. 16** shows the harmonic mean of SDRs of separated signals and that of original recorded signals (baseline) in optimal robot layouts. Since GICA and GHDSS are adaptive beamforming methods, their performances were better than DSBF and null beamforming. The separation performance of null beamforming was relatively high because in this experiment, we used correct sound source positions. Since null beamforming has sharp directivity, separation performance would deteriorate significantly if localization results had errors.

In this experiment, the harmonic means of SDRs obtained using any of the separation methods were worse for two reasons than those obtained when the methods were used in the experiment in the simulated room. The first reason is that in the simulation experiment, the impulse response used for source separation was the same as that used for synthesizing observed signals. In this experiment, however, for source separation, we used the impulse



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

Fig. 12. Results of robot layout optimization for three source layouts. Circles and squares indicate the positions of robots and target sound sources, respectively. Arrows indicate the directions of robots and sound sources.

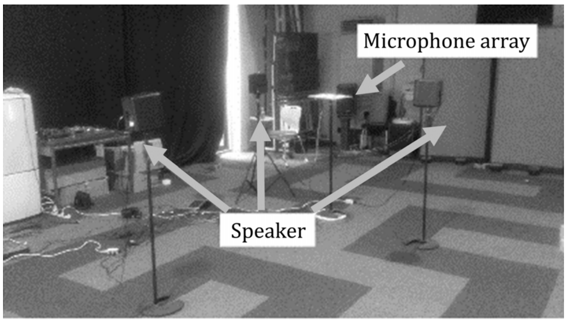


Fig. 13. Measuring real impulse responses in an environmental room.

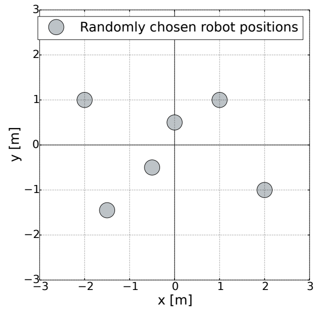
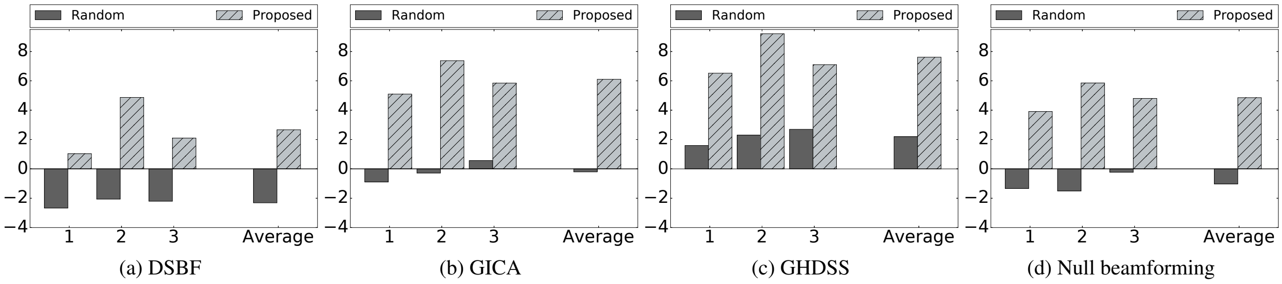


Fig. 14. Randomly chosen candidate positions of the robots.



(a) DSBF (b) GICA (c) GHDSS (d) Null beamforming

Fig. 15. Harmonic mean of SDRs in three sound source layouts for each source separation method in a real environment.

responses calculated geometrically from microphone array positions. The second reason is the directivity of a sound source. In the experiment in the simulated room, we assumed that sound sources had no directivity. Real sound sources, however, have directivity, and time differences of arrivals differed based on the directions of sound sources.

This directivity problem could be solved by having robots move to the front side of a sound source. This is because reverberation and diffraction cause the TDOAs at the sides and rear of a sound source to differ from the expected TDOAs. Another promising solution would be to estimate the directions of sound sources by audio-visual integration and to use an objective function that takes directivity into account.

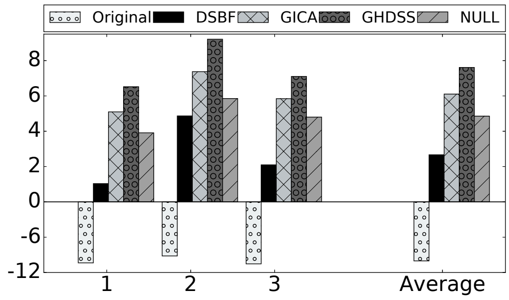


Fig. 16. Harmonic mean of SDRs of the original observed signal and that of separated signals obtained by each source separation method in optimal robot positions.

5. Conclusion

This paper presented an active-audition method that optimizes the layout of multiple mobile robots for separating the target sound sources highly accurately. To take advantage of using multiple mobile robots when separating recorded signals, we regarded them as one big microphone array. The optimal layout is determined by theoretically predicting the source separation performance (gain) based on DSBF from a possible layout. We conducted three experiments: (1) an evaluation of the layout optimization method (hill-climbing, gradient descent, or genetic algorithm), (2) an evaluation of the source separation performance in simulations, and (3) an evaluation of the source separation performance in a real environment. In experiment (1), we confirmed the effectiveness of GA. In experiments (2) and (3), we compared the method we presented with one that chooses the positions of robots randomly and found that it improved the average source separation performance by 5.7 dB in simulation and by 5.6 dB in a real environment.

The proposed method estimates the optimal layout of multiple robots even if each robot has a different microphone array. If each robot has a different number of microphones, for example, the position of a robot having fewer microphones has a small impact on the objective function. If there are obstacles between microphones, it is possible to calculate the objective function by modifying the amplitudes and phases of impulse responses recorded at an interval of 5° based on the positional relationship between sound sources and robots.

We are now planning to develop a prediction method that considers the properties of adaptive beamforming. To deal with moving sound sources whose positions are not given in advance, we plan to combine a method of simultaneous localization and mapping (SLAM) with dynamic motion planning for multiple robots. Specifically, a partially observable Markov decision process (POMDP) would be useful for dynamically updating the robot paths in real time so that the performance of source separation and localization is maximized.

Acknowledgements

This study was partially supported by KAKENHI 24220006 and 15K12063 and JST ImPACT Tough Robotics Challenge (TRC).

References:

[1] H. G. Okuno, K. Nakadai, and H.-D. Kim, "Robot Audition: Missing Feature Theory Approach and Active Audition," Robotics Research, Vol.70, pp. 227-244, Springer, 2011.

[2] E. Martinson, T. Apker, and M. Bugajska, "Optimizing a Reconfigurable Robotic Microphone Array," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 125-130, 2011.

[3] T. Nakashima, K. Komatani, and S. Sato, "Natural Interaction with Robots, Knowbots and Smartphones," chapter Integration of Multiple Sound Source Localization Results for Speaker Identification in Multiparty Dialogue System, pp. 153-165, Springer, 2014.

[4] K. Sekiguchi, Y. Bando, K. Itoyama, and K. Yoshii, "Optimizing the Layout of Multiple Mobile Robots for Cooperative Sound Source Separation," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 5548-5554, 2015.

[5] K. Nakadai, H. G. Okuno, and H. Kitano, "Real-Time Sound Source Localization and Separation for Robot Audition," IEEE Int. Conf. on Spoken Language Processing, pp. 193-196, 2002.

[6] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "SLAM-based Online Calibration of Asynchronous Microphone Array for Robot Audition," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 524-529, 2011.

[7] D. Su, T. Vidal-Calleja, and J. V. Miro, "Simultaneous Asynchronous Microphone Array Calibration and Sound Source Localisation," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 5561-5567, 2015.

[8] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Blind Source Separation With Parameter-Free Adaptive Step-Size Method for Robot Audition," IEEE Trans. on Audio, Speech and Language Processing, Vol.18, No.6, pp. 1476-1485, 2010.

[9] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active Audition for Humanoid," American Association for Artificial Intelligence, pp. 832-839, 2000.

[10] E. Berglund and J. Sitte, "Sound Source Localisation Through Active Audition," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 509-514, 2005.

[11] H.-D. Kim, J.-S. Choi, and M. Kim, "Human-Robot Interaction in Real Environment by Audio-Visual Integration," J. of Control, Automation and Systems, Vol.5, No.1, pp. 61-69, 2007.

[12] G. L. Reid and E. Milios, "Active Stereo Sound Localization," J. of Acoustical Society of America, Vol.113, No.1, pp. 185-193, 2003.

[13] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Multiple Sound Source Mapping for a Mobile Robot by Self-motion Triangulation," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 380-385, 2006.

[14] T. Yoshida and K. Nakadai, "Active Audio-Visual Integration for Voice Activity Detection based on a Causal Bayesian Network," IEEE RAS Int. Conf. on Humanoid Robots, pp. 370-375, 2012.

[15] Y. Sasaki, T. Fujihara, S. Kagami, H. Mizoguchi, and K. Oro, "32-Channel Omni-Directional Microphone Array Design and Implementation," J. of Robotics and Mechatronics, Vol.23, No.3, pp. 378-385, 2011.

[16] N. Mitianoudis and M. Davies, "Independent Component Analysis and Blind Signal Separation," Lecture Notes in Computer Science, Vol.3195, chapter Permutation Alignment for Frequency Domain ICA Using Subspace Beamforming Methods, pp. 669-676, Springer, 2004.

[17] L. Parra and C. Spence, "Convolutive Blind Separation of Nonstationary Sources," IEEE Trans. on Speech and Audio Processing, Vol.8, No.3, pp. 320-327, 2000.

[18] R. Mazur and A. Mertins, "A CUDA Implementation of Independent Component Analysis in the Time-frequency Domain," European Signal Processing Conf., 2011.

[19] J. Hao, I. Lee, T.-W. Lee, and T. J. Sejnowski, "Independent Vector Analysis for Source Separation Using a Mixture of Gaussians Prior," J. of Neural Computatation, Vol.22, No.6, ppp. 1646-1673, 2010.

[20] I. Lee, T. Kim, and T.-W. Lee, "Fast Fixed-point Independent Vector Analysis Algorithms for Convolutive Blind Source Separation," J. of Signal Processing, Vol.87, No.8, pp. 1859-1871, 2007.

[21] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind Source Separation Based on a Fast-Convergence Algorithm Combining ICA and Beamforming," IEEE Trans. on Audio, Speech and Language Processing, Vol.14, No.2, pp. 666-678, 2006.

[22] J. Mockus, "On Bayesian Methods for Seeking the Extremum," Proceedings of the IFIP Technical Conf., pp. 400-404, Springer-Verlag, 1974.

[23] Y. Sagisaka and N. Uratani, "ATR Spoken Language Database," J. of the Acoustic Society of Japan, Vol.48, No.12, pp. 878-882, 1992.

[24] C. Raffel, B. McFee, E. J. humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir eval: A Transparent Implementation of Common MIR Metrics," The Int. Society of Music Information Retrieval, pp. 367-372, 2014.

[25] E. Vincent, R. Gribonval, and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," IEEE Trans. on Audio, Speech and Language Processing, Vol.14, No.4, pp. 1462-1469, 2006.

Name:
Kouhei Sekiguchi

Affiliation:
Speech and Audio Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Address:
Room 417, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Brief Biographical History:

2015- Graduate School of Informatics, Kyoto University

Main Works:

- "Optimizing the Layout of Multiple Mobile Robots for Cooperative Sound Source Separation," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp. 5548-5554, 2015.
- "Online Simultaneous Localization and Mapping of Multiple Sound Sources and Asynchronous Microphone Arrays," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp. 1973-1979, 2016.

Membership in Academic Societies:

- Information Processing Society of Japan (IPSJ)

Name:
Yoshiaki Bando

Affiliation:
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
JSPS Research Fellow DC1

Address:
Room 417, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Brief Biographical History:

2014 Received M.Inf. degree from Graduate School of Informatics, Kyoto University

2015- Ph.D. Candidate, Graduate School of Informatics, Kyoto University, JSPS Reseach Fellow DC1

Main Works:

- "Posture estimation of hose-shaped robot by using active microphone array," Advanced Robotics, Vol.29, No.1, pp. 35-49, 2015 (Advanced Robotics Best Paper Award).
- "Variational Bayesian Multi-channel Robust NMF for Human-voice Enhancement with a Deformable and Partially-occluded Microphone Array," European Signal Processing Conf. (EUSIPCO), pp. 1018-1022, 2016.
- "Microphone-accelerometer based 3D posture estimation for a hose-shaped rescue robot," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp. 5580-5586, 2015.

Membership in Academic Societies:

- The Institute of Electrical and Electronic Engineers (IEEE) Robot Automation Society (RAS)
- The Robotics Society of Japan (RSJ)
- Information Processing Society of Japan (IPSJ)

Name:
Katsutoshi Itoyama

Affiliation:
Assistant Professor, Speech and Audio Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Address:
Room 417, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Brief Biographical History:

2011- Received Ph.D. degree from Graduate School of Informatics, Kyoto University

2011- Assistant Professor, Graduate School of Informatics, Kyoto University

Main Works:

- "Query-by-Example Music Information Retrieval by Score-Informed Source Separation and Remixing Technologies," EURASIP J. on Advances in Signal Processing, Vol.2010, No.1 pp. 1-14, January 17, 2011.

Membership in Academic Societies:

- The Institute of Electrical and Electronics Engineers (IEEE)
- The Acoustical Society of Japan (ASJ)
- Information Processing Society of Japan (IPSJ)

Name:
Kazuyoshi Yoshii

Affiliation:
Senior Lecturer, Speech and Audio Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Address:
Room 412, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Brief Biographical History:

2008- Received Ph.D. degree from Graduate School of Informatics, Kyoto University

2008- Research Scientist, Information Technology Research Institute (ITRI), National Institute of Advanced Industrial Science and Technology (AIST)

2013- Senior Researcher, AIST

2014- Senior Lecturer, Graduate School of Informatics, Kyoto University

Main Works:

- "A Nonparametric Bayesian Multipitch Analyzer Based on Infinite Latent Harmonic Allocation," IEEE Trans. on Audio, Speech, and Language Processing, Vol.20, No.3, pp. 717-730, 2012.

Membership in Academic Societies:

- The Institute of Electrical and Electronic Engineers (IEEE)
- Information Processing Society of Japan (IPSJ)
- The Institute of Electronics, Information, and Communication Engineers (IEICE)