SUMMARY We propose a new type of direction-of-arrival estimation method for robot audition that is free from strict head related transfer function estimation. The proposed method is based on statistical pattern recognition that employs a ratio of power spectrum amplitudes occurring for a microphone pair as a feature vector. It does not require any phase information explicitly, which is frequently used in conventional techniques, because the phase information is unreliable for the case in which strong reflections and diffractions occur around the microphones. The feature vectors we adopted can treat these influences naturally. The effectiveness of the proposed method was shown from direction-of-arrival estimation tests for 19 kinds of directions: 92.4% of errors were reduced compared with the conventional phase-based method.

**key words:** robot audition, direction-of-arrival estimation, pattern recognition, multiple signal classification, crosspower-spectrum phase

1. Introduction

We propose a direction-of-arrival (DOA) estimation method that is suitable for an implementation of robot audition that is able to cope with changes of acoustic environments around the robot.

The problems we addressed here are somewhat common in the research field of microphone array signal processing, which has a long history. Multiple signal classification (MUSIC), which is based on a correlation matrix, is frequently applied to sound source localization and DOA estimation [1]. This method works effectively in the presence of multiple sound sources or under noisy conditions, but it requires steering vectors from a sound source to microphones. Under the assumption that the microphones are placed on an almost free field as applied to conventional microphone array techniques, we can easily calculate these steering vectors by using only the characteristics of delays between the microphones and the designated source positions because the effects of reflections and diffractions can be ignored. In the case of a robot, on the other hand, since the microphones are placed on the robot head (or body), the effects of the reflections and the diffractions that occur around the robot degrade performances of DOA estimation. In this case, it is not realistic to approximate the steering vectors under the assumption of only the direct waves. In order to solve this problem, precise head related transfer functions (HRTFs) of the robot were measured in all possible areas around the robot [2], [3]. However, the measurement of the HRTFs is indeed troublesome work. Nakada et al. tried to calculate the HRTFs geometrically by making an approximation on the shape of their robot head: they regard it as a simple sphere [4], [5]. However, in most cases, robot heads are far from spherical.

On the other hand, the method based on time delay estimation or phase difference estimation using cross-power-spectrum phase (CSP) are also frequently applied to sound source localization and DOA estimation [6]–[9]. However, for the case in which the microphones are mounted on the robot head, it is difficult to obtain a precise phase difference occurring between microphones because the reflections and the diffractions caused by the robot head cannot be avoided. Furthermore, the problem involves near-fields in which sound sources and microphones do not exist on a same plane. For example, a conversation at which our robot aims is the case. Sato et al., geometrically calculate both the distance between the sound source and the microphones and the difference of heights between the sound source and the microphones [10]. In this method, the source localization performance highly depends on the preciseness of source height estimation. However, it is difficult to estimate the source height precisely in the environment in which the robot actually operates. In order to remove echos and reverberations, the source localization method based on time difference estimation is integrated with echo avoidance [11]. This method requires a generalized pattern of impulse responses that have some features such as delay time of first echo and the time constant of echo decay. However, it is not realistic to apply such approximated impulse responses for the case in which the microphones are placed on the robot head with a complicated structure.

In the present paper, we propose a new type of DOA estimation method using directivity microphones mounted on the robot head that is able to cope with environmental changes without strict HRTFs. A ratio of power spectrum amplitudes occurring between any pair of two microphones gives a discriminative pattern for each DOA regardless of sound source spectra because of the influences of the reflections and the diffractions caused by the robot head. Thus, the DOA can be estimated using statistical pattern recognition in which feature parameters consist of these power spectrum amplitude ratios. In this method, the DOA is estimated by choosing the best matching DOA, which gives the highest likelihood for all DOAs, using statistical models that are preliminarily trained for each DOA. It should be noted...
that the influences of the reflections and the diffractions can be naturally treated in this method while they cause a degradation of the DOA estimation performance in conventional methods that are based on unreliable phase difference estimation.

The rest of the present paper is organized as follows. The microphone arrangement of the robot head is described in Sect. 2. In Sect. 3, the algorithm of the proposed DOA estimation method is described in detail. Section 4 gives conditions and results of DOA estimation experiments in a real environment. Finally in Sect. 5, we give conclusions.

2. Microphone Arrangement

As depicted in Fig. 1, two microphones are mounted on each side of the robot head. One is directed to the frontal side of the face and the other is directed perpendicularly sideward. We refer to the two frontal-directed microphones as right-mounted frontal-directed microphone (RF-Mic) and left-mounted frontal-directed microphone (LF-Mic), and the two side-directed microphones as right-mounted right-directed microphone (RR-Mic) and left-mounted left-directed microphone (LL-Mic). The directivity microphones we used are audio-technica ATM15a.

3. DOA Estimation

3.1 A Ratio of Power Spectrum Amplitudes

A ratio of power spectrum amplitudes is calculated between input signals received from different microphones. We assume an environment in which a sound source exists and a sound field is observed by the microphones. In the present paper, \( S(\omega) \) denotes the spectrum of the sound source, and \( X_{LL}(\omega), X_{LF}(\omega), X_{RF}(\omega) \) and \( X_{RR}(\omega) \) denote the spectra of the input signals received from the LL-Mic, the LF-Mic, the RF-Mic and the RR-Mic, respectively, where \( \omega \) denotes a discrete frequency. Using HRTFs, \( X_i(\omega) \) is described as follows.

\[
|X_{LL}(\omega)| = |G_{LL}(\omega, \theta)| \cdot |S(\omega)| \\
|X_{LF}(\omega)| = |G_{LF}(\omega, \theta)| \cdot |S(\omega)| \\
|X_{RF}(\omega)| = |G_{RF}(\omega, \theta)| \cdot |S(\omega)| \\
|X_{RR}(\omega)| = |G_{RR}(\omega, \theta)| \cdot |S(\omega)|
\]

where \( G_{LL}(\omega, \theta), G_{LF}(\omega, \theta), G_{RF}(\omega, \theta) \) and \( G_{RR}(\omega, \theta) \) denote HRTFs from the sound source to the LL-Mic, the LF-Mic, the RF-Mic and the RR-Mic, respectively, and \( \theta \) denotes a DOA. Figure 2 shows amplitude characteristics of the HRTFs that are observed from the four directivity microphones mounted on the robot head. The amplitude characteristics of the HRTFs have poles and zeros due to the influence of the reflection and the diffraction waves caused by the robot head. The frequencies and the amplitude levels at the poles and the zeros are different for each DOA. Thus, the HRTFs can be regarded as a function of the DOA.

Here, we define \( X_{RF}(\omega) \) as a spectrum of a reference signal. Spectra of the remaining microphone inputs are divided by the spectrum of this reference signal, and then the spectrum of the sound source \( S(\omega) \) can be canceled. As a result, the ratio of power spectrum amplitudes for each microphone pair can be expressed as the ratio of HRTF amplitudes.

![Fig. 1](image1.png)

Fig. 1 The robot head and the microphones mounted on the robot head.

![Fig. 2](image2.png)

Fig. 2 Amplitude characteristics of HRTFs. Each figure shows the amplitudes as a function of frequencies.
regardless of the sound source spectrum $S(\omega)$ as follows.

\[
\begin{align*}
|X_{LL}(\omega)| &= |G_{LL}(\omega, \theta)| \\
|X_{RF}(\omega)| &= |G_{RF}(\omega)| \\
|X_{LF}(\omega)| &= |G_{LF}(\omega)| \\
|X_{RR}(\omega)| &= |G_{RR}(\omega)| \\
|X_{RF}(\omega)| &= |G_{RF}(\omega)| \\
|X_{RF}(\omega)| &= |G_{RF}(\omega)|
\end{align*}
\]  

The ratios of the power spectrum amplitudes for each microphone pair that are described in Eqs. (5), (6), (7) can also be expressed as the functions of the DOA because the HRTF is expressed as the function of the DOA as previously mentioned. From the above viewpoints, we propose a DOA estimation method based on statistical pattern recognition in which the ratio of power spectrum amplitudes for each microphone pair, $\frac{X_i(\omega)}{|X_{RF}(\omega)|}$ where $i=LL, LF, and RR$, is used as a feature vector. This method can estimate the DOA regardless of the sound source spectrum, which represents a speech utterance, and does not require troublesome strict HRTF measurements. In addition, the influence of the reflections and the diffractions caused by the robot head are naturally treated in the proposed method while they degrade DOA estimation performance in the conventional method based on unreliable phase difference estimation.

3.2 Feature Vector Extraction

A feature vector used for pattern recognition is extracted every utterance. In the present paper, the utterance is assumed to consist of a word. Figure 3 shows the diagram of the feature extraction process. Feature extraction consists of four-step signal processing as follows.

(1) Short time Fourier transformation

Short time Fourier transformation (STFT) is performed for the speech signals of the word utterances. $X_i(\omega,t)$ denotes a STFT coefficient of the signal received from the microphone $i$-Mic at a discrete frequency $\omega$ and a frame $t$.

(2) Normalization of power spectrum amplitudes

Power spectrum amplitudes of the speech signals received from the LL-Mic, the LF-Mic, and the RR-Mic are normalized using the reference spectrum $X_{RF}(\omega,t)$ as follows.

\[
N_i(\omega,t) = \frac{|X_i(\omega,t)|}{|X_{RF}(\omega,t)|} \quad (i=LL, LF, RR)
\]  

Here, for the frequency $\omega$ and the frame $t$ at which $|X_{RF}(\omega,t)|$ is close to zero, $N_i(\omega,t)$ cannot be calculated and thus it does not contribute to the temporal averaging of the next step.

(3) Temporal averaging

The normalized power spectrum amplitude $N_i(\omega,t)$ is averaged for all frames covered by a word utterance.

\[
N_i(\omega) = \frac{1}{T} \sum_{t=1}^{T} N_i(\omega,t) \quad (i=LL, LF, RR)
\]  

where $T$ denotes the number of the frames of the word utterance. Here, signals are determined as speech parts for the case in which signal powers and zero-crossing rates are higher than the predefined thresholds.

(4) Filter-bank analysis

A filter bank that consists of $L$-channel triangular filters arranged on the frequency axis at regular intervals is applied to three normalized power spectrum amplitudes, $N_{LL}(\omega)$, $N_{LF}(\omega)$, and $N_{RR}(\omega)$. Here, since amplitude information is generally dominant in high frequency bands, and thus it is desirable to get high resolutions in high frequency bands, the filter banks are arranged not on the frequency axis at mel-scaled intervals but on the frequency axis at the regular intervals. The triangular filter-bank we used is depicted in Fig. 4 in which the number of filter-bank channels is $L$. Each amplitude is multiplied by the corresponding filter gain and
the results are then summed in the filter. Thus, each filter-bank coefficient \( c_i(l) \) holds a weighted sum that represents a normalized power spectrum amplitude in that filter-bank channel. \( c_i(l) \) is described as follows.

\[
c_i(l) = \sum_{k=k_{lo}}^{k_{hi}} W(k; l) \cdot \log N_i(k) \quad (l = 1, \ldots, L)
\]

where \( k_{lo}, k_c \), and \( k_{hi} \) are the low, the center and the high frequency of the \( l \)-th filter-bank channel. The time-averaged spectrum of each channel is compressed into a \( L \)-dimensional vector as follows.

\[
c_{LL} = (c_{LL}(1), \ldots, c_{LL}(L)) \quad (12)
\]

\[
c_{LF} = (c_{LF}(1), \ldots, c_{LF}(L)) \quad (13)
\]

\[
c_{RR} = (c_{RR}(1), \ldots, c_{RR}(L)) \quad (14)
\]

We conclusively use a \( 3 \times L \)-dimensional feature vector \( (c_{LL}, c_{LF}, c_{RR}) \), which is a concatenation of \( c_{LL}, c_{LF}, \) and \( c_{RR} \).

### 3.3 DOA Estimation Using Pattern Recognition

Using the feature parameters described in 3.2, pattern recognition is performed on the basis of the maximum likelihood (ML) criterion. In this framework, the DOA is regarded as a categorical class and a statistical model is trained for each DOA.

In the training stage, parameters of Gaussian mixture model (GMM) are estimated using word utterances coming from the corresponding DOA. A likelihood that the GMM of the DOA class \( c \) gives to a feature vector \( x_i \) extracted from the \( i \)-th word utterance is described as follows.

\[
P(x_i | M^c) = \sum_{k=1}^{K} w_k^c \cdot N(x_i; \mu_k^c, \Sigma_k^c)
\]

\[
N(x_i; \mu_k^c, \Sigma_k^c) = (2\pi)^{-D/2} |\Sigma_k^c|^{-1/2} \exp \left[ -(x_i - \mu_k^c)^T (\Sigma_k^c)^{-1} (x_i - \mu_k^c) \right]
\]

where \( k \) denotes an index of a mixture component, \( c \) denotes an index of a DOA class, \( K \) denotes the number of mixtures of GMMs, \( w_k^c \), \( N() \), \( \mu_k^c \), and \( \Sigma_k^c \) denote the mixture weight, the probabilistic distribution function, the mean vector and the covariance matrix of the \( k \)-th component in the GMM of the DOA class \( c \), respectively. \( M^c \) denotes the parameter assembly of GMM, which consists of \( \mu_k^c, \Sigma_k^c \), and \( w_k^c \), and \( D \) denotes the dimensionality of \( x_i \).

In the classification stage, likelihoods are calculated for all DOA GMMs using Eq. (15) and Eq. (16), and then the best matching class \( \hat{c} \), which gives the highest likelihood for all the classes, is determined as the DOA of an input speech signal \( x_i \) as follows.

\[
\hat{c} = \arg \max_c \left[ \log P(x_i | M^c) \right]
\]

### 4. DOA Estimation Experiment

Experimental comparisons were conducted under various acoustic conditions around the robot. The proposed method was compared with conventional DOA estimation methods such as the MUSIC method [2], [3], which required measurements of strict HRTFs, and the CSP-based method [7]–[9], which was based on phase difference estimation. Here, since we target robot audition for conversations with people, the sound field can be assumed to be the near-field.

#### 4.1 Speech Materials

Figure 5 shows the recording environment. Speech data were sampled at 32kHz and quantized into 16bit data. We recorded word utterances through the four-line directivity microphones mounted on the robot head under 27 kinds of conditions, which consist of three kinds of arrangements of the robot, three kinds of reverberation times, and three kinds of heights of a sound source. The speech data were played back through a loudspeaker that acts as a replacement of people (sound source).

Three kinds of robot arrangements, A-1, A-2, and A-3, are described as follows.

A-1 The robot is placed at P-1 facing the parallel direction to the X-axis.

A-2 The robot is placed at P-1 facing the parallel direction to the Y-axis.

A-3 The robot is placed at P-2 facing the parallel direction to the Y-axis.

The reverberation times of the room were 238ms, 318ms, and 395ms, which were controlled by drawing heavy curtains around the room. The heights of the sound source (loudspeaker) were 104cm, 125cm, and 150cm. We have 27 different recording conditions by making any combination of them.
In addition, 19 kinds of DOAs were radially placed every 10 degrees around the robot from -90 degrees to 90 degrees. In the present paper, the front, the right and the left direction of the robot were defined as zero, positive, and negative degrees, respectively.

For the recordings, 100 isolated spoken words were chosen from the ATR phonetically balanced word database [12]. We selected ten speakers and ten words per speaker. The speech data of the 100 words were recorded for each of 19 directions. 90 words, which were spoken by nine male speakers, were used for a training of the DOA models, and the remaining words (ten words spoken by one male speaker) were used for a test. The evaluation was conducted using 10-fold cross-validation tests. As the result, a total of 100 words were used for the evaluation. Here, speakers and vocabularies used for the test were different from those used for the training in each evaluation set.

4.2 Experimental Setup for the Proposed Method

4.2.1 Feature Extraction

The speech data were analyzed using a Hamming window with a frame length of 128 ms and a frame shift of 32 ms. We investigated the DOA estimation performances for the cases in which the filter-bank with 8, 16, and 24 channels was applied. Here, 24, 48, and 72-dimensional feature parameters were extracted for the case in which four microphones were used in the proposed method, because the feature vector has \(3 \cdot L\)-dimensionality, where the number of filter-bank channels is \(L\). In addition, the feature parameters can be extracted using only two microphones, RF-Mic and LF-Mic, which are directed to the frontal side of the face. This aims at comparing the proposed method with the conventional CSP-based method. In this case, using the filter-bank with 8, 16 and 24 channels, the dimensionalities of the feature vector were 8, 16 and 24, respectively. In this experiments, a frequency range was restricted from 0 Hz to 4000 Hz.

4.2.2 DOA Recognition

A DOA model consists of a Gaussian mixture distribution with diagonal covariance matrices. We investigated the DOA estimation performance for the case in which 1, 2, 4, 6, 8, 10, 12, 14 were applied to the number of mixtures in a GMM used for the DOA model.

The speech data used for the DOA model training were recorded under 12 kinds of conditions, which consist of two kinds of robot arrangements, two kinds of reverberation times, and three kinds of sound source heights. On the other hand, evaluations were conducted under the source height-closed condition, which does not include the same robot arrangements and reverberation times as used in the training but include the same heights of the sound source as used in the training. Figure 6 shows an example of a combination of the training and the test data. In this case, the training were performed using the speech data recorded under the conditions of the robot arrangements of A-1 or A-2, the reverberation times of 238 ms or 318 ms, and the sound source heights of 104 cm, 125 cm or 150 cm. On the other hand, the evaluation was performed using the speech data recorded under the condition of the robot arrangement of A-3, the reverberation time of 395 ms, and the same kinds of source heights as the training. Nine-fold cross-validation tests were conducted using different combinations of the training and the test data with respect to the conditions of the robot arrangements and the reverberation times. Thus, the evaluations were “open” concerning both the robot arrangements and the reverberation times of the room but “closed” concerning the sound source heights.

4.3 Experimental Setup for the Conventional Method

The performance of the proposed method under the sound source height-closed condition was compared with the performances of the conventional methods such as the MUSIC method using strict HRTFs that were measured for each combination of DOAs and sound source heights, and the
CSP-based method under the assumption of the near-field, which requires the sound source height.

### 4.3.1 Overview of MUSIC Method

An overview of the MUSIC method for DOA estimation is described as follows. In this method, a DOA estimate is obtained as the DOA that gives the highest peak of MUSIC spectrum amplitudes as a function of DOAs. At first, a spatial correlation matrix \( \mathbf{R} = \mathbf{E} \mathbf{x}(\omega, t) \mathbf{x}(\omega, t)^H \) is calculated. Here, \( \mathbf{x}(\omega, t) \) denotes an input vector, which consists of STFT coefficients of the input signal received from microphones. Then, \( \mathbf{R} \) is decomposed into the signal subspace and the noise subspace. The MUSIC spectrum is calculated in each frequency band using eigenvalue decomposition, \( \mathbf{R} = \mathbf{E} \lambda \mathbf{E}^{-1} \), where \( \mathbf{E} \) denotes the eigenvector matrix, the columns of which consist of the eigenvectors of \( \mathbf{R} \), and \( \lambda \) denotes the eigenvalue matrix, the diagonal elements of which consist of the eigenvalues of \( \mathbf{R} \). The standard MUSIC spectrum is described as follows.

\[
P_{\text{music}}(\omega, \theta) = \frac{a^H(\omega, \theta) a(\omega, \theta)}{a^H(\omega, \theta) \mathcal{E}_N \mathcal{E}_N^H a(\omega, \theta)}
\]  

(18)

where

\[
\mathcal{E}_N = [\mathbf{e}_{N_s+1}, \mathbf{e}_{N_s+2}, \ldots, \mathbf{e}_{N_m}]
\]

Here, \( N_s \) and \( N_m \) are termed the number of sound sources and microphones, respectively, and they have to be known in advance. \( \mathcal{E}_N \) denotes the eigenvectors corresponding to the smallest \( N_m-N_s \) eigenvalues \( \mathbf{e}_{N_s+1}, \mathbf{e}_{N_s+2}, \ldots, \mathbf{e}_{N_m} \), which become the basis of the noise subspace. \( a \) denotes a steering vector, which consists of the measured HRTFs of the direct paths from the sound source in direction \( \theta \) to the microphones. The MUSIC spectra are added for a certain frequency range. The DOA estimates \( \theta_i (i = 1, \ldots, N_s) \) are obtained by choosing the DOAs that give the highest peaks of the MUSIC spectrum after frequency-averaging.

### 4.3.2 Setup for MUSIC Method

Experimental setup of the MUSIC method is described as follows. Speech was analyzed using a Hamming window with a frame length of 96 ms and a frame shift of 32 ms. In the present paper, the number of microphones was four and the number of sound sources was one. MUSIC spectrum was computed by using the spatial correlation matrix averaged for three analysis blocks, and then the peak of the MUSIC spectrum was regarded as the DOA estimate for these analysis blocks. The DOAs can be determined every 96 ms. The DOA for a word utterance was conclusively determined as the most frequent DOA for all analysis blocks included in the utterance. The analysis block, in which the input energy was small, was eliminated from the calculation of a spatial correlation matrix. Here, frequency addition was performed for the range from \( \omega_1 \) to \( \omega_2 \) as follows.

\[
P_{\text{music}}(\theta) = \sum_{\omega=\omega_1}^{\omega_2} P_{\text{music}}(\omega, \theta)
\]

(20)

In this experiment, \( \omega_1 \) was 2000 Hz and \( \omega_2 \) was 4000 Hz. The steering vectors had to be obtained for 19 points, which have intervals of 10 degrees from -90 to 90 degrees. Thus, the HRTFs were strictly measured at these 19 points with the distance between the sound source and the robot of 100 cm for each of three kinds of sound source heights. The DOA estimation system choose the highest peak of the MUSIC spectrum, which are sufficiently larger than that of both adjacent sides, out of these 19 positions.

### 4.3.3 Overview of CSP-Based Method

An overview of a CSP-based DOA estimation method is described as follows. The aim of CSP-based DOA estimation is to search the delay \( \tau \) that gives the highest frame-averaged CSP coefficient for all delays \( \tau \) as follows [13].

\[
\hat{\tau} = \arg \max_{\tau} \left( \frac{1}{T} \sum_{t=1}^{T} C(t, \tau) \right)
\]

(21)

where \( T \) denotes the number of frames of a word utterance, and \( C(t, \tau) \) denotes the CSP coefficient of the \( t \)-th frame, which is estimated as follows.

\[
C(t, \tau) = \text{IFFT} \left[ \frac{X_1(t, \omega) \cdot X_2^*(t, \omega)}{|X_1(t, \omega)\cdot|X_2(t, \omega)|} \right]
\]

(22)

where \( X_1(t, \omega) \) and \( X_2(t, \omega) \) denote spectra of signals received from microphones, and \( |\cdot|^* \) represents the complex conjugate. Finally, using the corresponding delay \( \hat{\tau} \), the DOA can be obtained as follows.

\[
\theta = \sin^{-1} \left( \frac{c \cdot \hat{\tau}}{M \cdot F_s} \right)
\]

(23)

where \( c \) denotes the acoustic velocity, \( M \) denotes the microphone spacing, and \( F_s \) denotes the sampling frequency. Here, Eq. (23) is realized under the assumption that the microphones are located in the far-field.

### 4.3.4 Setup for CSP-Based Method

In this experiment, we used two frontal-directed microphones, RF-Mic and LF-Mic. In order to achieve angular resolutions of 10 degrees even in the side of the robot, speech data were up-sampled at 64 kHz. The speech data were analyzed using a Hamming window with a frame length of 64 ms and a frame shift of 16 ms. In the present paper, a tolerance interval in DOA estimation was five degrees. This means that the estimates within errors of five degrees are regarded as correct. For the case in which a DOA could not be calculated because a time delay became too much long, the DOA was determined to be 90 degrees or -90 degrees.

In the case of application to robot audition we targets, the distance between the microphones and the sound source
is not sufficiently large compared with the microphone spacing, and the assumption of the far-field is not realized. Thus, near-field DOA estimation is required. Figure 7 illustrates a near-field placement of the microphones and the sound source. Here, the near-field DOA \( \phi \) can be estimated using the following geometrical compensation [10].

\[
\phi = \sin^{-1}\left( \frac{d \cdot \sin \theta}{\sqrt{d^2 + h^2}} \right)
\]

where \( d \) denotes the distance from the microphones to the sound source, \( h \) denotes the relative height, which represents the difference between the height of the microphones and the height of the sound source, and \( \theta \) denotes the DOA for \( h = 0 \) estimated by Eq. (23). In this experiment, the relative height \( h \) was 17.4 cm, 38.4 cm, and 63.2 cm because the height of the microphones was 86.6 cm, the microphone spacing \( M \) was 17.9 cm, and the distance between the microphones and the loudspeaker \( d \) was 100 cm.

From the above discussion, the CSP-based method requires a precise sound source height for each utterance under the assumption of the near-field. On the other hand, it is required for the proposed method that the training data of the DOA models include the data played back from the source which is same in height as used in the evaluation, but does not require the precise source height for each utterance. Thus, the experimental condition of the CSP-based method has an advantage over that of the proposed method.

4.4 Experimental Result

Figure 8 describes the DOA corrects of the proposed GMM-based method and the conventional methods. Each bar represents the DOA correct averaged for all DOAs. Figure 9 describes the DOA corrects for each DOA. Here, in the proposed method, the number of filter-bank channels used in feature extraction and the number of mixtures in a DOA GMM were arbitrarily determined so as to give the best performance of DOA estimation. For the case in which four microphones were used, the optimal number of filter-bank channels and mixtures in a GMM were 24 and 6, respectively. For the case in which two microphones were used, they were 24 and 12, respectively.

Figure 8 shows that the CSP-based method gave a DOA correct of 59.6% averaged for all DOAs under the assumption of the near-field. As we can see in Fig. 9, the CSP-based method achieved precise DOA estimation with respect to the frontal-directions of the robot. On the other hand, with respect to the sideward-directions of the robot head, since signals with multiple paths based on the reflections and the diffractions were derived by the robot head, this method had a tendency that phases were delayed compared with the case that only the direct wave was arrived. This caused a degradation of the DOA recognition performance. Thus, it is difficult to estimate the DOA using the phase information.

On the other hand, the MUSIC method and the proposed GMM-based method achieved precise DOA estimation for all DOAs. The proposed method gave the DOA corrects of 99.5% and 96.9% averaged for all DOAs using four microphones and two microphones, respectively.
for the case in which two microphones were used and the closed condition was applied to the sound source heights, the experimental condition of the proposed method could be regarded to be almost equivalent to the condition of the CSP-based method under the assumption of the near-field. The proposed method could achieve significant improvement in the DOA estimation performance compared with the conventional CSP-based method. Here, the proposed method reduced 92.4% of errors of the CSP-based method. In addition, the proposed method gave almost the same performance compared to the MUSIC method (97.9%), which required the measurements of the strict HRTFs for each combination of DOAs and sound source heights. Thus, the proposed method achieved precise DOA estimation by only applying not the strict HRTFs, which were required for the MUSIC method, but the approximated HRTFs, which were estimated from arbitrary word utterances coming from each DOA.

These results show that the proposed method can deal with the influence of the reflections and the diffractions naturally and can cope with changes of the reverberation times and the positions of the robot.

In addition, preliminary experiments show that the proposed method can cope with changes of the distance between the microphones and the sound source. The proposed method gave good performances (DOA corrects of approximately 100%) not only under the distance-closed conditions but also under the distance-open conditions in which the training data included both the information of smaller distances and the information of larger distances than the evaluation data.

5. Conclusion

We proposed the robust DOA estimation method that does not require an estimate of a strict HRTF. The proposed method is based on statistical pattern recognition using a ratio of power spectrum amplitudes for each microphone pair as a feature vector. The feature parameters used for the proposed method can treat reflections and diffractions derived from the robot head naturally, and the parameters can be independent of the sound source spectra. The DOA estimation experiments under the various acoustic conditions showed that the effectiveness of the proposed method: it reduced 92% and 99% of errors of the conventional CSP-based method using two microphones and four microphones, respectively, and it achieved almost the same performance compared to MUSIC method, which required measurements of the strict HRTFs. Thus, the proposed method achieved high-performance DOA estimation by leveraging the influences of the reflections and the diffractions derived from the robot head while the conventional CSP-based method was adversely affected by those influences.

References


280, March 1986.


Tetsuji Ogawa received the B.S., M.S. and Ph.D. degrees in Department of Electric, Electronics and Computer engineering, Waseda University, Tokyo, Japan, in 2000, 2002 and 2005, respectively. He was a Research Associate (2004–2007) and a Visiting Lecturer (2007) at Waseda University. He has been an Assistant Professor in Waseda Institute for Advanced Study. His research interest includes stochastic modeling for pattern recognition, speech enhancement and speech recognition. He is a member of Acoustic Society in Japan.

Tetsunori Kobayashi received the B.S., M.S., and Ph.D. degrees in electrical engineering from Waseda University, Tokyo, Japan, in 1980, 1982 and 1985, respectively. He was a Lecturer (1985–1987) and an Associate Professor (1987–1991) at Hosei University, Tokyo, Japan. In 1991, he joined Waseda University as an Associate Professor. Since 1997, he has been a Professor. He was a visiting scientist at Spoken Language System Group, Laboratory for Computer Science, Massachusetts Institute of Technology from 1994 through 1995. His research interests include perceptual computing and intelligent robotics. He is a member of Information Processing Society of Japan, and Institute of Electrical and Electronics Engineers.