DIRECTION-OF-ARRIVAL ESTIMATION UNDER NOISY CONDITION USING FOUR-LINE OMNI-DIRECTIONAL MICROPHONES MOUNTED ON A ROBOT HEAD

Tetsuji Ogawa[1], *Kosuke Hosoya*[2], *Kenzo Akagiri*[2], *and Tetsunori Kobayashi*[2]

[1]Waseda Institute for Advanced Study. [2]Dept. of Computer Science, Waseda University.
3-4-1 Okubo, Shinjuku-ku, 169-8555, Tokyo, JAPAN

ABSTRACT

We propose a new direction-of-arrival (DOA) estimation method suitable for autonomous mobile robots. Autonomous mobile robots have to meet physical constraints of signal processing devices, such as a space-saving microphone arrangement and few computational resources. In addition, DOA estimation of the robots needs to be robust to noise around the robots. In order to cope with the physical constraints, we used four-line omni-directional micro mechanical systems (MEMS) microphones. DOA estimation was conducted using statistical pattern recognition in which normalized spectral amplitudes, which were free from sound sources, were used as DOA features. In the proposed method, strict head related transfer function estimation, which is not practically feasible, is not needed. In addition, unlike many conventional methods, phase information is not explicitly used because the phase information is unreliable in the situation that we deal with, i.e., situations in which the microphone spacings are small, or strong reflections and diffractions occur around the microphones. The feature vectors we used can cope with these problems. Effectiveness of the proposed method was experimentally investigated in recognition of 19 DOAs in the presence of diffuse noise: the proposed method achieved a DOA correct of approximately 99% at a SNR of 0 dB.

1. INTRODUCTION

We attempt to achieve high-performance direction-of-arrival (DOA) estimation, which is a basis of robot audition, using the compact and light-weighted devices, which can be mounted on autonomous mobile robots.

The problems addressed in the present study are common in the research field of microphone array signal processing. Multiple signal classification (MUSIC) is frequently applied to sound source localization and DOA estimation[1]. Although this method works effectively in noisy environments, steering vectors from a sound source to microphones are required. If the microphones are placed on free-fields, we can easily compute these steering vectors with the characteristics of the delays between the designated source position and the microphones. However, when the microphones are placed on the robot head (or body), the effects of the reflections and diffractions induced by the robot cannot be ignored. In order to cope with these effects, precise head related transfer functions (HRTFs) of the robot were measured in all possible areas around the robot[2]. However, the measurement of such data is not practically feasible. Nakadai et al. approximated the shape of the robot head by a simple sphere for computing the HRTFs geometrically[3]. However, in most cases, robot heads are far from spherical.

Methods using time delays or phase differences between microphones (e.g., crosspower-spectrum phase (CSP)) were also frequently applied to sound source localization and DOA estimation[4, 5, 6]. However, phase differences cannot be precisely estimated when the microphone spacings are small as in the case of the present study. In addition, when the microphones are mounted on the robot, precise phase difference estimation is difficult because of the reflections and diffractions induced by the robot. In the case of near-field, where sound sources and microphones do not exist on the same plane, Sato et al. improved the performance of a sound source localization system by using both the distance between the sound source and the microphones and the heights of the sound source and the microphones[7]. However, it is difficult to estimate the source height precisely while the robot is moving.

In a previous paper, we proposed a DOA estimation method that was free from strict HRTF measurements, by using four-line directional microphones mounted on a robot head[8]. This method required the use of directional microphones. However, it is difficult to develop directional microphones using micro electro mechanical systems (MEMS) technologies, which are necessary for the miniaturization and weight-saving of microphone systems of autonomous mobile robots. Therefore, the microphones and signal processing devices could not be miniaturized easily.

In the present paper, we propose a new DOA estimation method using omni-directional microphones, which are suitable for MEMS technologies. In the present study, four-line analog MEMS omni-directional microphones were placed on the top of the robot head. In the proposed method, DOAs are estimated by pattern recognition: feature extraction, which consists of multiple beamforming, spectral amplitude normalization, temporal averaging, and filter-bank analysis, is performed, and static pattern recognition with Gaussian mixture models (GMM) is then carried out. In this case, the normalized spectral amplitudes of the outputs of multiple beamformers, which correspond to directional microphone observations, are discriminative for each DOA, irrespective of sound source spectra. It should be noted that the DOA features we used can cope with the influences of the reflections and diffractions naturally, while the conventional methods, using unreliable phase difference estimation, degrade the performance of DOA estimation. In addition, the proposed DOA estimation method is consistent with the noise reduction method we proposed in [9], which uses the same beamformer outputs as used in the present study.

The rest of the present paper is organized as follows: The microphone system used is described in Section 2. In Section 3, we describe a feature extraction method in detail. In Section 4, we describe a method for estimating the DOAs using pattern recognition. In Section 5, we present conditions and results of DOA estimation experiments. Finally, in Section 6, we present our concluding remarks.

2. MICROPHONE SYSTEMS

We used the compact and light-weighted microphones and signal processing devices, which were suitable to be mounted on autonomous mobile robots.

2.1 MEMS microphone

We used four-line analog MEMS microphones manufactured by a semiconductor integrated technology; they were significantly compact and light-weighted. We used SPM0208HD5 as the microphone. The width, depth, and height of this microphone is 4.72 mm, 3.76 mm, and 1.25 mm, respectively. We made 1.5-cm-square substrates. Each of these substrates comprises a MEMS microphone and peripheral circuits with a pre-amplifier. These substrates were

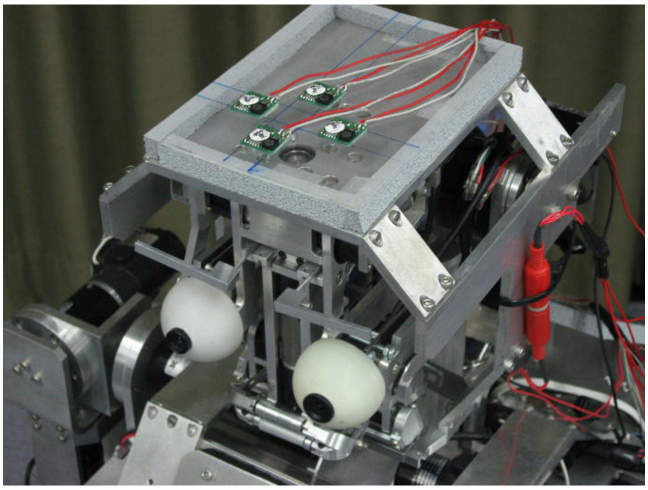

Figure 1: Robot and microphone substrates.

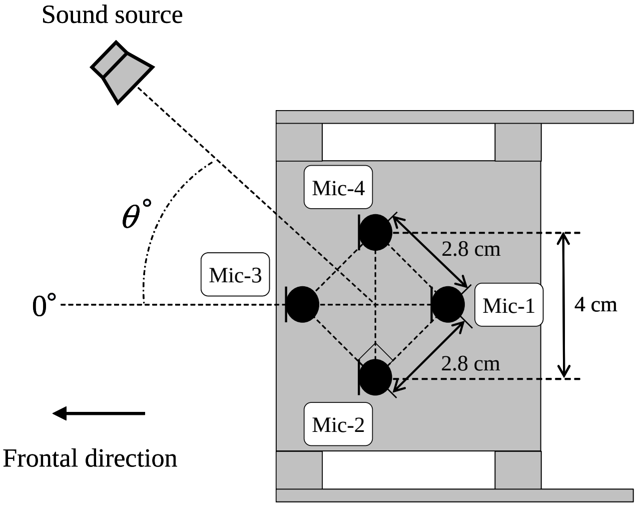

Figure 2: Microphone arrangement. This figure shows the top view of the robot.

placed on the top of the robot head, as shown in Figure 1.

2.2 Microphone arrangement

Figure 2 shows a microphone arrangement. The microphones were placed in a squared form, where the spacing between adjacent microphones was 2.8 cm and that between diagonally opposite microphones was 4 cm. The microphone channels are labelled as shown in Figure 2. In the present study, the front, right, and left direction of the robot are defined as zero, positive, and negative degrees, respectively.

2.3 A/D conversion system

Four-channel analog signals received by the microphones were converted into digital signals using a compact embedded device. The device consists of SUZAKU-V.SZ310 and SID00-U00. SUZAKU-V.SZ310 is an universal embedded device platform, which is based on the combination of FPGA and Linux (with a PowerPC405 CPU core) with a 10 BASE-T/100 BASE-TX Ethernet connector. SID00-U00 is an eight-channel A/D conversion system, which is used as an extension of the SUZAKU board. A resolution of the SID00-U00 was refined from its original 12 bits to 16 bits. The digital signals were transfered via Ethernet to a laptop PC mounted on the robot. DOA estimation was then carried out on the laptop PC.

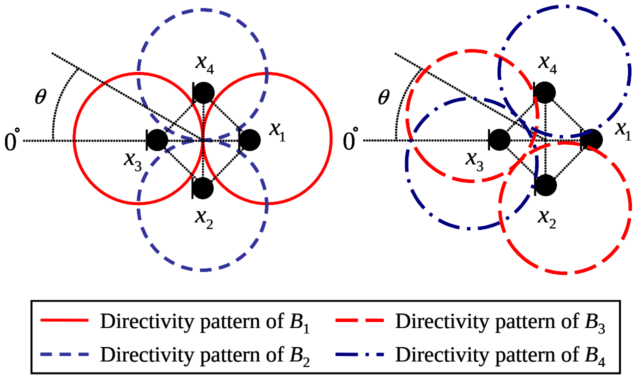
θ
0°
x_4
x_1
x_3
x_2
θ
0°
x_4
x_1
x_3
x_2
Directivity pattern of B_1
Directivity pattern of B_3
Directivity pattern of B_2
Directivity pattern of B_4

Figure 4: Directivity patterns of subtractive beamformers.

3. FEATURE EXTRACTION IN DOA ESTIMATION

DOA estimation was carried out by pattern recognition. We attempted to extract DOA features that were free from sound source spectra (i.e., speech utterances). Figure 3 shows a schematic diagram of the feature extraction method. This method consists of four stages of signal processing as follows: 1) multiple beamforming for developing directivities, 2) spectral amplitude normalization for eliminating the influence of the sound source spectra on the DOA features, 3) temporal averaging for improving reliability of DOA estimation, and 4) filter-bank analysis for reducing dimensionality of feature vectors.

3.1 Multiple beamforming

In the present study, we developed four subtractive beamformers using the microphone observations as follows:

$$B_1(\omega,k) = X_1(\omega,k) - X_3(\omega,k) \tag{1}$$

$$B_2(\omega,k) = X_4(\omega,k) - X_2(\omega,k) \tag{2}$$

$$B_3(\omega,k) = X_3(\omega,k) - X_2(\omega,k) \tag{3}$$

$$B_4(\omega,k) = X_3(\omega,k) - X_4(\omega,k) \tag{4}$$

where ω denotes the discrete frequency; k denotes the discrete frame; $X_i(\omega,k)$ denotes the spectral component of the observation received by Mic-i; and $B_j(\omega,k)$ denotes the spectral component of the j-th beamformer output. The directivity patterns of these beamformers are shown in Figure 4.

3.2 Spectral amplitude normalization

We assume the situation where a sound source exists and a sound field is observed by the microphones. In this case, microphone observations are described as follows:

$$|X_1(\omega,k)| = |G_1(\omega,\theta)| \cdot |S(\omega,k)| \tag{5}$$

$$|X_2(\omega,k)| = |G_2(\omega,\theta)| \cdot |S(\omega,k)| \tag{6}$$

$$|X_3(\omega,k)| = |G_3(\omega,\theta)| \cdot |S(\omega,k)| \tag{7}$$

$$|X_4(\omega,k)| = |G_4(\omega,\theta)| \cdot |S(\omega,k)| \tag{8}$$

where $S(\omega,k)$ denotes the spectral component of a sound source; $G_i(\omega,\theta)$ denotes the HRTF from the sound source to Mic-i; and θ denotes the DOA.

We define the observation at Mic-3 as a reference signal. The spectral amplitudes of the beamformer outputs, $|B_j(\omega,k)|$, were normalized by the spectral amplitude of the reference signal,

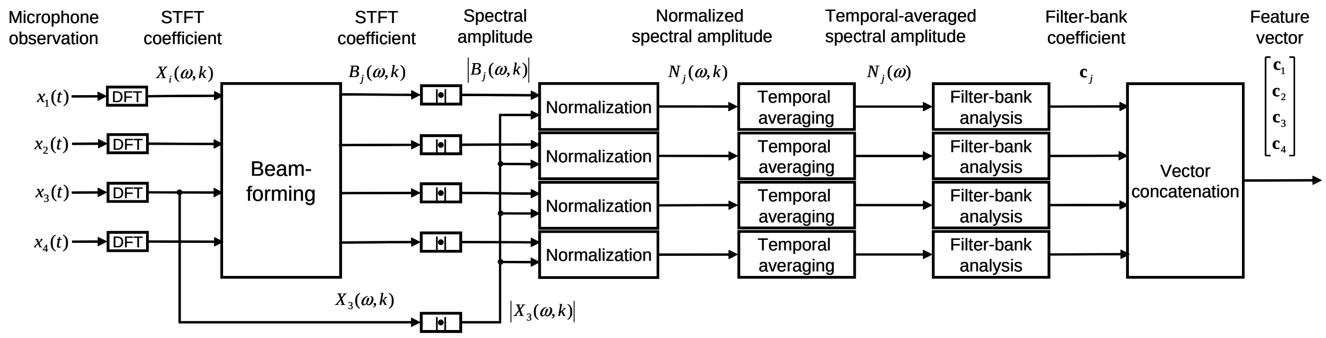

Figure 3: Schematic diagram of feature extraction method.

$|X_3(\omega,k)|$, as follows:

$$N_1(\omega,k) = \frac{|B_1(\omega,k)|}{|X_3(\omega,k)|} = \frac{|G_1(\omega,\theta)-G_3(\omega,\theta)|}{|G_3(\omega,\theta)|} \quad (9)$$

$$N_2(\omega,k) = \frac{|B_2(\omega,k)|}{|X_3(\omega,k)|} = \frac{|G_4(\omega,\theta)-G_2(\omega,\theta)|}{|G_3(\omega,\theta)|} \quad (10)$$

$$N_3(\omega,k) = \frac{|B_3(\omega,k)|}{|X_3(\omega,k)|} = \frac{|G_3(\omega,\theta)-G_2(\omega,\theta)|}{|G_3(\omega,\theta)|} \quad (11)$$

$$N_4(\omega,k) = \frac{|B_4(\omega,k)|}{|X_3(\omega,k)|} = \frac{|G_3(\omega,\theta)-G_4(\omega,\theta)|}{|G_3(\omega,\theta)|} \quad (12)$$

where $N_j(\omega,k)$ denotes the spectral component of the normalized spectral amplitude of the output of the j-th beamformer. In this case, features used in DOA estimation should have only the DOA information, and should be free from the sound source spectra. Spectral amplitude normalization described in Equations 9, 10, 11, and 12 aims at removing the influence of the sound source spectra on the DOA features. In this case, HRTFs are functions of DOAs. Ideally, the normalized spectral amplitudes are determined by only the DOAs, irrespective of the observed speech utterances, because they are expressed with only the HRTFs as described in Equations 9, 10, 11, and 12. Therefore, the normalized spectral amplitudes are suitable for the DOA features. It should be noted that these features can cope with the influence of the reflection and diffraction induced by the robot head or body.

The influence of the sound sources can be eliminated even when the omni-directional microphone observations are used instead of the beamformer outputs. However, the features extracted from the directional microphone observations (i.e., beamformer outputs) are expected to be more discriminative as compared to those extracted from the omni-directional microphone observations[8]. In addition, the performance of DOA estimation can be improved by using additional beamformers that are different in directivity from B_1, B_2, B_3, and B_4.

3.3 Temporal averaging

The normalized spectral amplitudes were extracted every frame. For the case in which these amplitudes are used as DOA features, DOA estimation can be more reliable by integrating these amplitudes over multiple frames. Therefore, the normalized spectral amplitudes were averaged over all frames covered by a speech utterance as follows:

$$N_j(\omega) = \frac{1}{K}\sum_{k=1}^{K} N_j(\omega,k) \quad (j=1,\cdots,4) \quad (13)$$

where K denotes the number of frames in a speech utterance. In this case, speech parts were detected with signal powers and zero-crossing rates.

3.4 Filter-bank analysis

In order to reduce dimensionality of feature vectors, filter-bank analysis was conducted. A filter bank that consists of L-channel triangular filters arranged on the frequency axis at regular intervals was applied to $N_1(\omega)$, $N_2(\omega)$, $N_3(\omega)$, and $N_4(\omega)$. Each spectral amplitude was multiplied by the corresponding filter gain, and the results were then summed in the filter. Therefore, each filter-bank coefficient $c_i(l)$ holds a weighted sum of the normalized spectral amplitudes in that filter-bank channel. $c_i(l)$ is described as follows:

$$c_j(l) = \sum_{\omega=\omega_{\mathrm{lo}}}^{\omega_{\mathrm{hi}}} W(\omega;l)\cdot \log N_j(\omega), \quad (l=1,\cdots,L) \quad (14)$$

$$W(\omega;l) = \begin{cases} \frac{\omega-\omega_{\mathrm{lo}}(l)}{\omega_{\mathrm{c}}(l)-\omega_{\mathrm{lo}}(l)}, & (\omega_{\mathrm{lo}}(l) \le \omega \le \omega_{\mathrm{c}}(l)) \\ \frac{\omega_{\mathrm{hi}}(l)-\omega}{\omega_{\mathrm{hi}}(l)-\omega_{\mathrm{c}}(l)}, & (\omega_{\mathrm{c}}(l) \le \omega \le \omega_{\mathrm{hi}}(l)) \end{cases} \quad (15)$$

where $\omega_{\mathrm{lo}}(l)$, $\omega_{\mathrm{c}}(l)$, and $\omega_{\mathrm{hi}}(l)$ are the low, center, and high frequency of the l-th filter-bank channel. In this case, $N_j(\omega)$ was compressed into a L-dimensional vector $\boldsymbol{c}_j$ using Equation 14 as follows:

$$\boldsymbol{c}_1 = (c_1(1),\cdots,c_1(L)) \quad (16)$$

$$\boldsymbol{c}_2 = (c_2(1),\cdots,c_2(L)) \quad (17)$$

$$\boldsymbol{c}_3 = (c_3(1),\cdots,c_3(L)) \quad (18)$$

$$\boldsymbol{c}_4 = (c_4(1),\cdots,c_4(L)) \quad (19)$$

Consequently, a $4\cdot L$-dimensional vector $(\boldsymbol{c}_1,\boldsymbol{c}_2,\boldsymbol{c}_3,\boldsymbol{c}_4)$ was extracted every utterance.

4. DOA ESTIMATION USING PATTERN RECOGNITION

Pattern recognition was carried out under the maximum likelihood (ML) criterion. In the present study, a DOA was regarded as a categorical class, and a statistical model was trained for each DOA.

In the training stage, parameters of a Gaussian mixture model (GMM) were estimated with speech utterances coming from the corresponding DOA. A likelihood of the DOA class C, given to a feature vector $\boldsymbol{x}_n$, which was extracted from the n-th speech utterance, was computed as follows:

$$P(\boldsymbol{x}_n|\mathscr{M}^C) = \sum_{m=1}^{M} w_m^C \cdot \mathscr{N}(\boldsymbol{x}_n;\boldsymbol{\mu}_m^C,\boldsymbol{\Sigma}_m^C) \quad (20)$$

$$\mathscr{N}(\boldsymbol{x}_n;\boldsymbol{\mu}_m^C,\boldsymbol{\Sigma}_m^C) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}_m^C|^{-\frac{1}{2}}\cdot \exp\left[(\boldsymbol{x}_n-\boldsymbol{\mu}_m^C)^{\mathrm{T}}(\boldsymbol{\Sigma}_m^C)^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_m^C)\right] \quad (21)$$

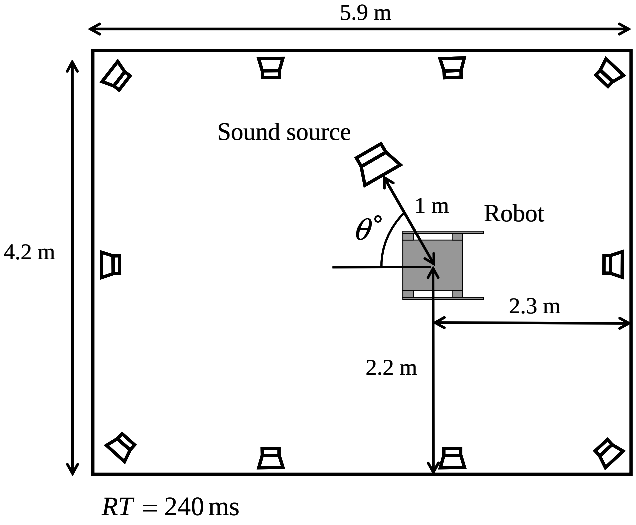

Figure 5: Recording environment.

where C denotes the DOA class; m denotes the mixture component; M denotes the number of mixtures in GMMs; w_m^C, $\mathscr{N}(\cdot)$, $\boldsymbol{\mu}_m^C$, and $\boldsymbol{\Sigma}_m^C$ denote the mixture weight, the probabilistic distribution function, the mean vector, and the covariance matrix in the m-th component of the GMM with a DOA class of C, respectively; $\mathscr{M}^C$ denotes the parameter assembly of the GMM, which includes $\boldsymbol{\mu}_m^C$, $\boldsymbol{\Sigma}_m^C$, and w_m^C; and D denotes the dimensionality of $\boldsymbol{x}_n$.

In the classification stage, likelihoods of all DOA GMMs to a speech utterance were computed using Equations 20 and 21. The best matching class $\hat{C}$, which gave the highest likelihood for all classes, was determined as the DOA of the speech utterance as follows:

$$\hat{C} = \arg\max_{C} \left[\log P(\boldsymbol{x}_n | \mathscr{M}^C) \right] \tag{22}$$

5. DOA ESTIMATION EXPERIMENT

In the present study, DOA estimation systems were evaluated in terms of their automatic DOA estimation performance, which is based on the DOA correct, in noisy environment. The DOA correct was calculated by using a well-known formula, as follows:

$$\text{DOA correct} = \frac{N - S}{N} \times 100\ (\%) \tag{23}$$

where N and S denote the number of utterances we used in evaluation and that of utterances misestimated in DOA, respectively.

5.1 Speech materials

Figure 5 shows the recording environment. Speech utterances were 100 phonetically-balanced isolated word sentences, which were spoken by 10 male speakers. Each speaker uttered 10 words. The speech data were played on the loudspeaker, and recorded by the microphones placed on the head of the robot involved in conversation, “ROBISUKE”[10]. In this recording, the distance between the sound source and the robot was 100 cm, and the height of the loudspeaker was 125 cm. In this experiment, 19 DOAs were radially placed every 10 degrees from -90 ° to 90 ° around the robot. The speech data of 100 words were recorded for each of 19 directions. Diffuse noise was simulated as follows: noise from a large air-conditioning machine was played on 10 loudspeakers placed around the room. The diffuse noise recorded by the microphones on the robot head was then superposed on the word utterances so that the SNR of the word utterance to the diffuse noise would be -5, 0, 5, 10, and 15 dB.

Table 1: Setup of feature extraction.

sampling frequency	16 kHz
frame length	128 ms
frame shift	32 ms
analysis window	Hamming window
number of filter-banks	24
analysis range of frequencies	1500–6000 Hz

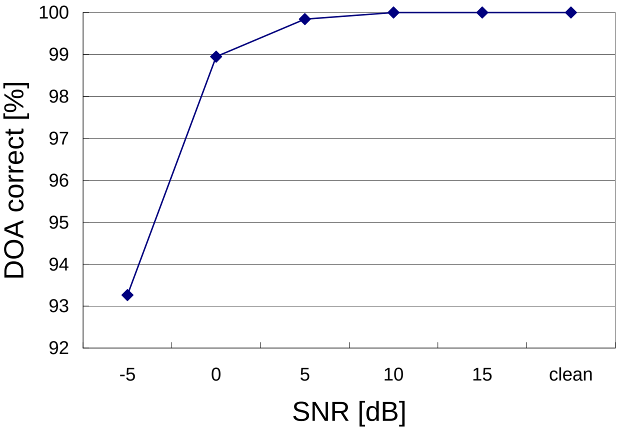

Figure 6: DOA corrects as a function of SNRs, averaged over all DOAs.

5.2 Experimental condition

Experimental conditions of feature extraction is shown in Table 1. In this case, 96-dimensional feature parameters were extracted because each of $N_1(\omega)$, $N_2(\omega)$, $N_3(\omega)$, and $N_4(\omega)$ was analyzed with 24-channel filter-banks.

A statistical model for each DOA was represented by a 2-mixture Gaussian distribution with diagonal covariances. Evaluation was carried out by 10-fold cross-validation tests: 90 words (spoken by nine male speakers) were used for training the DOA model, and the remaining 10 words (spoken by one male speaker) were used in evaluation, for each DOA. As a result, a total of 100 words were used in the evaluation. It should be noted that this experiment was conducted under the “open” conditions in terms of speakers and vocabularies, i.e., both the speakers and vocabularies used in the evaluation were different from those used for training the models in each fold.

5.3 Experimental result

Figure 6 shows DOA corrects of the proposed method, averaged over all DOAs, as a function of SNRs of speech utterances to diffuse noise. In this figure, “clean” denotes the quiet environment where only the speech utterances are observed. Figure 7 shows DOA corrects for each DOA in the cases of SNRs of -5, 0, and 5 dB. In this experiment, we arbitrarily determined the microphone spacings (i.e., 4 cm in a diagonal position), the number of filter-bank channels used in feature extraction (i.e., 24), the analysis range of frequencies (i.e., 1500–6000 Hz), and the number of mixtures in DOA models (i.e., 2) so that the best performance could be achieved.

Figure 6 shows that the proposed method can estimate DOAs precisely under the noisy condition: the proposed method achieved DOA corrects of 100% when diffuse noise was observed at higher SNRs than 10 dB, a DOA correct of approximately 99% at a SNR of 0 dB, and a DOA correct of approximately 93% even at a SNR

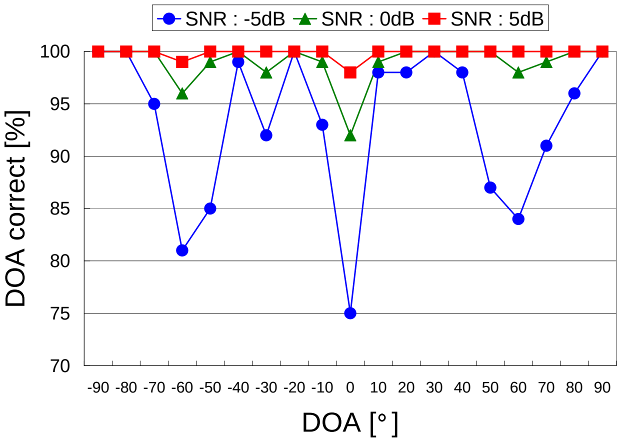

Figure 7: DOA corrects for each DOA.

of -5 dB.

As shown in Figure 7, the performances of DOA estimation were degraded in the DOAs of -60°, 0°, and 60°: the sound sources of the diffuse noise (i.e., loudspeakers) exist in around those DOAs. In order to improve the performances, we can carry out adaptation of DOA models to noise. The DOA features we used are suitable for such adaptation, which should be carried out with few speech utterances, because the proposed DOA features are free from sound source spectra.

In contrast, the performance of the CSP-based method was evaluated using the microphone observations of Mic-2 and Mic-4 under the assumption of the near-field[8]: this method gave a DOA correct of approximately 26% at most even in the quiet environment. Therefore, phase differences cannot be estimated precisely, when the microphone spacings are small and the reflections and diffractions induced by the robot occur around the microphones.

6. CONCLUSION

We proposed a new DOA estimation method that does not need estimation of strict HRTFs, using the compact and light-weighted MEMS microphones, which were suitable for autonomous mobile robots. In the proposed method, statistical pattern recognition was carried out using normalized spectral amplitudes as DOA features. These features were irrespective of sound sources, and could cope with the reflections and diffractions induced by the robot naturally. The experimental results in the presence of diffuse noise showed the effectiveness of the proposed method: it achieved the DOA corrects over 99% when the SNRs were higher than 0 dB, and a DOA correct of approximately 93% at a SNR of -5 dB.

REFERENCES

[1] R. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. Antennas Propagation, vol.AP-34, no.3, pp.276-280, March 1986.

[2] I. Hara et al., “Robust speech interface based on audio and video information fusion for humanoid HRP-2,” Proc. IROS2004, pp.2404-2410, Sept. 2004.

[3] K. Nakadai et al., “Applying scattering theory to robot audition system,” Proc. IROS, pp.1147-1152, Oct. 2003.

[4] C. H. Knapp et al., “The generalized correlation method for estimation of time delay,” IEEE Trans. Acoust. Speech and Signal Process., vol.ASSP-24, no.4, pp.320-327, 1976.

[5] M. Omologo et al., “Acoustic event localization using a crosspower-spectrum phase based technique,” Proc. ICASSP, vol.2, pp.273-276, April 1994.

[6] T. Nishiura et al., “Localization of multiple sound sources based on a CSP analysis with a microphone array,” Proc. ICASSP, vol.2, pp.1053-1056, June 2000.

[7] M. Sato et al., “Near-field sound-source localization based on a signed binary code,” IEICE Trans. Fundamentals, vol.E88-A, no.8, pp.2078-2086, 2005.

[8] N. Mochiki et al., “Ears of the robot: Direction of arrival estimation based on pattern recognition using robot-mounted microphones,” IEICE Trans. Inf.& Syst., vol.E91-D, no.5, pp.1522-1530, May 2007.

[9] T. Ogawa et al., “Ears of the robot: Noise reduction using four-line ultra-micro omni-directional microphones mounted on a robot head,” Proc. EUSIPCO, Aug. 2008.

[10] S. Fujie et al., “Multi-modal integration for personalized conversation: Towards a humanoid in daily life,” Proc. Humanoids, pp.617-622, Dec. 2008.