The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Robot auditory system using head-mounted square microphone array

Kosuke Hosoya, Tetsuji Ogawa, and Tetsunori Kobayashi

Abstract— **A new noise reduction method suitable for autonomous mobile robots was proposed and applied to preprocessing of a hands-free spoken dialogue system. When a robot talks with a conversational partner in real environments, not only speech utterances by the partner but also various types of noise, such as directional noise, diffuse noise, and noise from the robot, are observed at microphones. We attempted to remove these types of noise simultaneously with small and light-weighted devices and low-computational-cost algorithms. We assumed that the conversational partner of the robot was in front of the robot. In this case, the aim of the proposed method is extracting speech signals coming from the frontal direction of the robot. The proposed noise reduction system was evaluated in the presence of various types of noise: the number of word errors was reduced by 69 % as compared to the conventional methods. The proposed robot auditory system can also cope with the case in which a conversational partner (i.e., a sound source) moves from the front of the robot: the sound source was localized by face detection and tracking using facial images obtained from a camera mounted on an eye of the robot. As a result, various types of noise could be reduced in real time, irrespective of the sound source positions, by combining speech information with image information.**

I. INTRODUCTION

We attempt to achieve high-performance robot auditory system that reduces various types of noise, such as directional noise, diffuse noise, and noise from the robot, simultaneously, irrespective of source positions, with the compact and light-weighted devices and the low-computational-cost algorithms.

When a robot talks with a conversational partner in real environments, only the speech utterances by the partner should be extracted from noisy speech utterances that include various types of noise, and then be precisely recognized. In order to achieve such functions for autonomous mobile robots, microphones and signal processing devices have to meet certain constraints in their sizes and weights so as to mount them on the robots. Miniaturization of these devices needs low-cost computations. It should be noted that the partner (i.e., sound source) can move in a conversation.

Noise reduction methods for robot auditory systems have been researched in recent years. In order to cope with the reflections and diffractions induced by the robot head or body, precise head related transfer functions (HRTFs) were measured in all possible areas around the robot[1]. However, the measurement of such data for each robot and each microphone arrangement is not practically feasible. In order to solve this problem, the shape of the robot head was approximated by a simple sphere for computing the HRTFs geometrically[2]. In most cases, however, the robot heads are far from spherical. An open-source software for robot audition, "HARK," was developed[3]. This software can cope with sound source localization, sound source separation, and speech recognition. Auditory systems for various robots can be developed by integrating multiple customizable modules provided by this software, irrespective of the microphone arrangements. However, geometric source separation used in this software needs the locations of the microphones and the sound sources. In addition, this software did not cope with the voice of the robot. In a child care robot "PaPeRo," direction-of-arrival (DOA) was estimated, and noise was reduced by adaptive noise canceller with five-channel microphones mounted on the robot head[4]. This robot could detect double talk, and eliminate the robot voice using a sound played on a loudspeaker. However, the performance of this method was not evaluated in the situation where the directional noise and the diffuse noise were observed simultaneously. In a humanoid robot "SIG2," spectra of a target speaker in front of the robot were estimated by using complex spectrum circle centroid (CSCC)[5]. The purpose of estimation of the target source spectra was voice activity detection (VAD)[6]. In addition, this method did not cope with diffuse noise.

In the present paper, we propose a new noise reduction method using a square microphone array with four-line omnidirectional MEMS microphones. The proposed method can reduce the directional noise, the diffuse noise[7][8], the robot internal noise (i.e., the noise induced by the movements of the robot and the noise from the motors mounted on the robot), and the robot voices (i.e., synthesized speech utterances played on a loudspeaker mounted on the robot) simultaneously under the constraints of the compact microphone arrangement and low-computational-cost algorithms: simple beamforming was carried out for developing multiple directivity patterns; time-frequency masking was carried out with the outputs of the beamformers for reducing the directional noise, i.e., extracting a sound coming from a specific area; multichannel Wiener filtering was carried out with the coherences between the beamformer outputs for reducing the diffuse noise; and the synthesized speech utterances by the robot were eliminated by time-frequency masking. Since the square microphone array was placed on the top of the robot head, a null-directivity was given to the whole robot during reduction of the directional noise. Therefore, the robot internal noise can be eliminated in this stage.

In the proposed method, we assume that a conversational partner of the robot is in front of the robot. In order to cope with the case in which the partner moves from the front of the robot, the sound sources are localized by face detection and tracking using facial images obtained from a camera mounted on an eye of the robot. Although adaptive filtering (e.g., independent component analysis; ICA) can perform sound source localization and separation, this method has unavoidable delays for making the filter converge, after the sound source positions are determined. Moreover, the sound sources

This work was partly supported by New Energy and Industrial Technology Development Organization (NEDO).

K. Hosoya and T. Kobayashi are with Dept. of Computer Science, Waseda University, 3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, JAPAN.

T. Ogawa is with Waseda Institute for Advanced Study, 1–6–1 Nishi-waseda, Shinjuku-ku, Tokyo 169–8050, JAPAN.

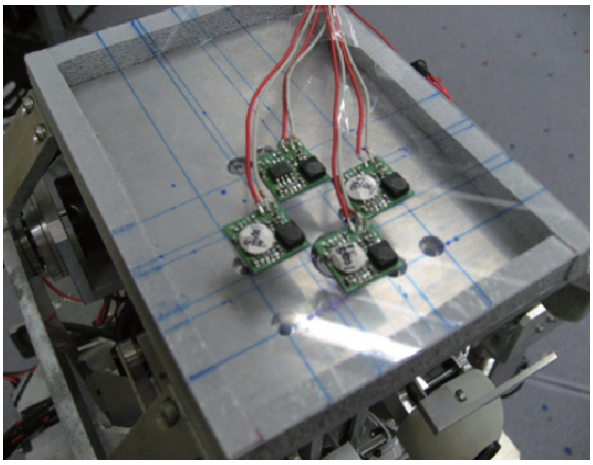

Fig. 1. Robot and microphone systems.

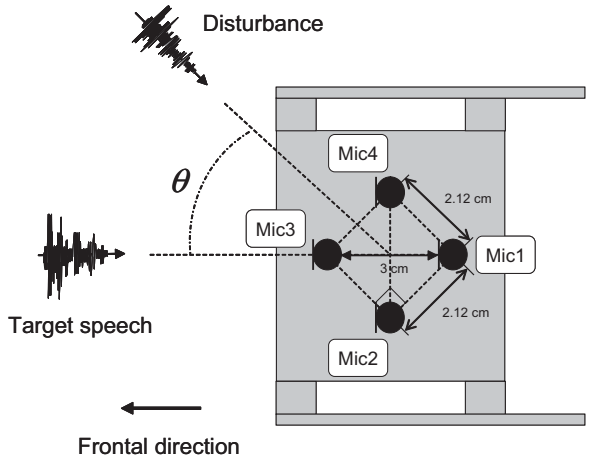

Fig. 2. Microphone arrangement. This figure shows the top view of the robot.

are not always at the same positions. Therefore, this method may not suitable for robot auditory systems. In contrast, the proposed system does not induce such delays by integrating simple sound source separation using speech information and simple sound source localization using image information.

The rest of the present paper is organized as follows: The robot auditory system used is described in Section II. In Section III, conditions and results of noise reduction experiments in real environments are given. In Section IV, sound source localization using face detection and tracking is described. Finally, in Section V, the concluding remarks are presented.

II. ROBOT AUDITORY SYSTEM

A. Microphone system

1) MEMS microphone: We used four-line analog MEMS microphones, which were constructed using a semiconductor integrated technology and were significantly compact and light-weighted. We used SPM0208HD5 made by Knowles Co., Ltd. The width, depth, and height of the 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 mounted on the robot head as shown in Fig. 1.

2) Microphone arrangement: Effects of the reflections and diffractions induced by the robot head or body can be eliminated by placing the microphones on the top of the robot head. The microphones were placed in a squared form. We call this arrangement of microphones "square microphone array." In the present study, each spacing of neighboring microphones was 2.12 cm, and that of microphones in a diagonal position was 3.0 cm. Notations of channels of the microphones are defined as shown in Fig. 2. In the present study, it is assumed that the target speech utterances come from the front of the robot. In this case, the front, right, and left direction of the robot are defined as zero, positive, and negative degrees, respectively.

B. Noise reduction system

Figure 3 shows a diagram of the proposed noise reduction method. The proposed method consists of four-stage signal processing: 1) time-frequency masking for directional noise reduction, 2) multichannel Wiener filtering for diffuse noise reduction, 3) single-channel Wiener filtering for residual noise reduction, and 4) time-frequency masking for robot voice reduction. In the present paper, $x_i(t)$ denotes a signal received by the Mic-i at a discrete time of t, and $X_i(\omega, k)$ denotes a STFT coefficient of x_i, where k and ω denote a discrete frame and a discrete frequency, respectively. $X_{\mathrm{r}}(\omega, k)$ denotes a spectral component of x_{r}, which is a voice of the robot.

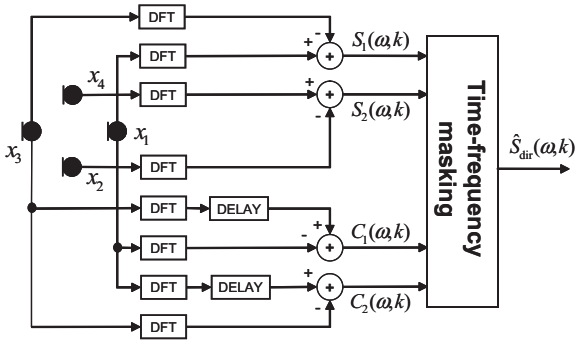

Fig. 4. Directional noise reduction system.

1) Directional noise reduction: Figure 4 shows a diagram of a directional noise reduction system. In this system, null beamformers and subtractive beamformers were developed, and then time-frequency masking was carried out using the outputs of these beamformers. $C_1(\omega, k)$ and $C_2(\omega, k)$ denote spectral components of the outputs of the null beamformers that were developed by delay addition followed by subtraction with $X_1(\omega, k)$ and $X_3(\omega, k)$. $C_1(\omega, k)$ and $C_2(\omega, k)$ were computed as follows:

$$C_1(\omega, k) = X_3(\omega, k) \cdot \exp(-j\omega\tau_{\mathrm{d}}) - X_1(\omega, k) \quad (1)$$

$$C_2(\omega, k) = X_1(\omega, k) \cdot \exp(-j\omega\tau_{\mathrm{d}}) - X_3(\omega, k) \quad (2)$$

where τ_{d} denotes a delay corresponding to the spacing of the microphones placed in a diagonal position. The directivity patterns of C_1 and C_2 are illustrated in Fig. 5. In this case, C_1 and C_2 form the directivity with a null in a direction of 0° and that with a null in a direction of 180°, respectively.

$S_1(\omega, k)$ and $S_2(\omega, k)$ denote a spectral component of the output of the subtractive beamformer developed with

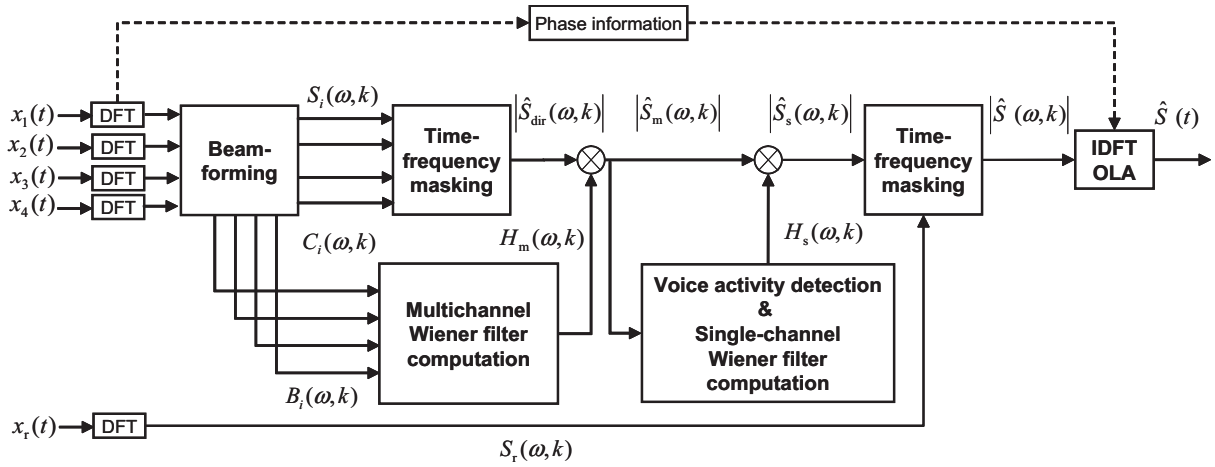

Fig. 3. Diagram of the proposed noise reduction system.

$X_1(\omega,k)$ and $X_3(\omega,k)$, and a spectral component of the output of the subtractive beamformer developed with $X_2(\omega,k)$ and $X_4(\omega,k)$, respectively. $S_1(\omega,k)$ and $S_2(\omega,k)$ were computed as follows:

$$S_1(\omega,k) = X_1(\omega,k) - X_3(\omega,k) \quad (3)$$
$$S_2(\omega,k) = X_4(\omega,k) - X_2(\omega,k) \quad (4)$$

The directivity patterns of S_1 and S_2 are illustrated in Fig. 6. S_1 forms the directivity that has maximum gains in directions of 0° and 180°, and nulls in directions of 90° and -90°. S_2 forms the directivity that has maximum gains in directions of 90° and -90°, and nulls in directions of 0° and 180°.

In this system, only the signals that come from the frontal direction of the robot, $\hat{S}_{\rm dir}$, were extracted by the following time-frequency masking:

$$\hat{S}_{\rm dir}(\omega,k) = \begin{cases} S_1(\omega,k), \\ \quad \text{if } |S_1(\omega,k)| > |S_2(\omega,k)| \\ \quad \text{and } |C_1(\omega,k)| < |C_2(\omega,k)| \\ \beta, \quad \text{otherwise} \end{cases} \quad (5)$$

where β denotes a flooring constant. In this time-frequency masking, the directional noise coming from the sides of the robot was suppressed by selecting the time-frequency components in which $S_1(\omega,k)$ was larger than $S_2(\omega,k)$ (as illustrated in Fig. 6), and then the noise from the backward of the robot was suppressed by selecting the components in which $C_2(\omega,k)$ was larger than $C_1(\omega,k)$ (as illustrated in Fig. 5). Consequently, the speech signals coming from the direction of the shaded area in Fig. 7 were extracted.

This directional noise reduction method separates sound sources both on the plane containing the square microphone array and in the frontal direction of the robot. In this method, a null-directivity was given to the whole robot. Therefore, this method can suppress robot internal noise, such as the noise induced by the movement of the robot and the noise from the motors of the robot.

2) Diffuse noise reduction: Diffuse noise included in $\hat{S}_{\rm dir}$ was reduced by multichannel Wiener filtering: spectral components with low interchannel correlations were regarded to be diffuse noise, and suppressed. First, four null beamformer outputs were computed as follows:

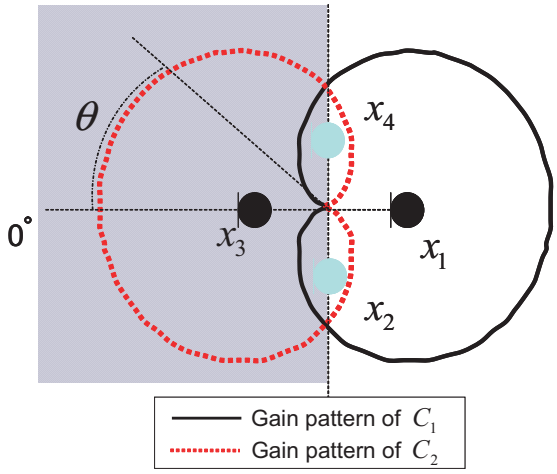

Fig. 5. Directivity patterns of null beamformer. Time-frequency masking removes spectral components of the signals coming from the backward of the robot (i.e., unshaded area).

$$B_1(\omega,k) = X_2(\omega,k)\cdot\exp(-j\omega\tau_{\rm n}) - X_1(\omega,k) \quad (6)$$
$$B_2(\omega,k) = X_3(\omega,k)\cdot\exp(-j\omega\tau_{\rm n}) - X_2(\omega,k) \quad (7)$$
$$B_3(\omega,k) = X_3(\omega,k) - X_4(\omega,k)\cdot\exp(-j\omega\tau_{\rm n}) \quad (8)$$
$$B_4(\omega,k) = X_4(\omega,k) - X_1(\omega,k)\cdot\exp(-j\omega\tau_{\rm n}) \quad (9)$$

where $\tau_{\rm n}$ denotes a delay corresponding to the spacing of neighboring microphones. Conventional methods compute the multichannel Wiener filter $H_{\rm m}(\omega,k)$ with the omni-directional microphone observations[9]. In contrast, the proposed method computes this filter with the null beamformer outputs instead of the microphone observations as follows:

$$H_{\rm m}(\omega,k) = \frac{\frac{1}{2}\sum[\text{abs}\{B_p(\omega,k)\cdot B_q^*(\omega,k)\}]}{\frac{1}{4}\sum_{i=1}^{4}[B_i(\omega,k)\cdot B_i^*(\omega,k)]} \quad (10)$$

where p and q were selected as $\{(p,q)\} = \{(1,2),(3,4)\}$ so that B_p and B_q could form line-symmetric directivity to the axis containing both the target source and the center of the microphone array: B_p and B_q have a difference of 90° in directivity. Figure 8 shows theoretical magnitude-squared coherences (MSCs) computed using the observations of the omni-directional microphones and those computed

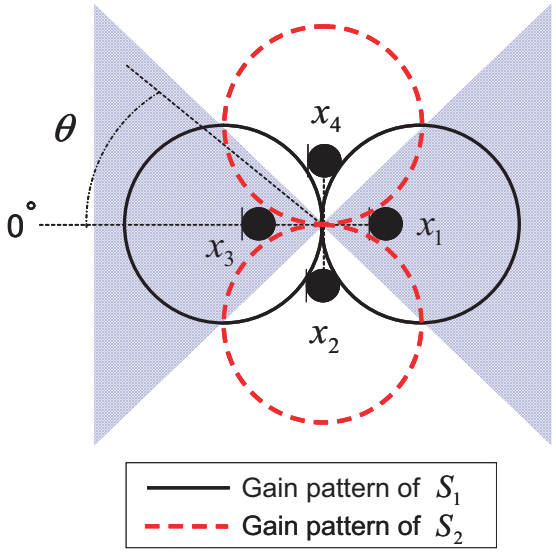

Fig. 6. Directivity patterns of subtractive beamformers. Time-frequency masking removes spectral components of the signals coming from the sides of the robot (i.e., unshaded area).

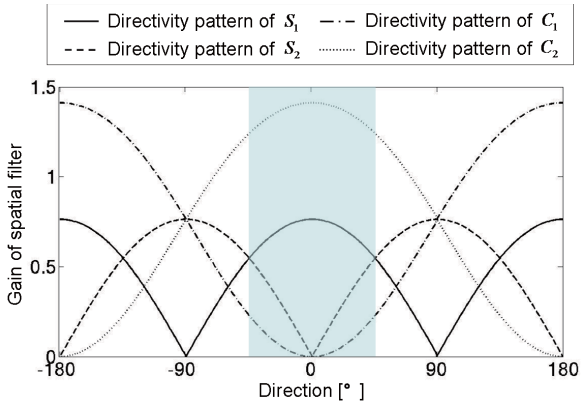

Fig. 7. Estimates of the directional noise reduction system. Spectral components of the signal coming from directions of the shaded area are extracted.

using the observations of the directional microphones with a interchannel difference of 90° in directivity, when the microphone spacing is 2.12 cm. This figure shows that the MSCs computed using the omni-directional microphones are not low but approximately one, especially in low frequencies, when the microphone spacings are small. In contrast, the MSCs can be eliminated by using the directional microphones. Therefore, the use of the directional microphones can be effective to diffuse noise reduction as compared to the use of the omni-directional microphones. In the proposed method, we used the outputs of the null beamformers that have a difference of 90° in directivity. Therefore, this method can eliminate the MSCs in diffuse noise fields as compared to the conventional multichannel Wiener filtering, which used the omni-directional microphone observations, and thus improve the performance of diffuse noise reduction[7]. The spectral amplitude of the result of the multichannel Wiener filtering was computed as follows:

$$|\hat{S}_{\mathrm{m}}(\omega,k)| = H_{\mathrm{m}}(\omega,k)\cdot|\hat{S}_{\mathrm{dir}}(\omega,k)| \tag{11}$$

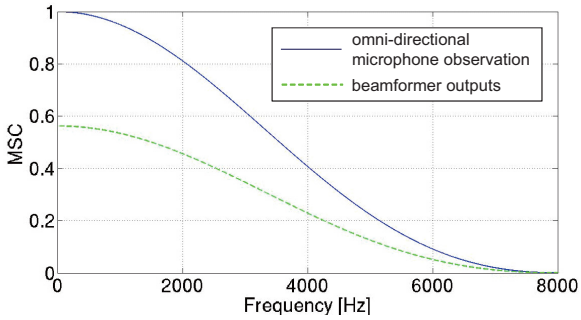

Fig. 8. Theoretical magnitude-squared coherences as a function of frequencies for the case in which microphones spacing is 2.12 cm.

3) Residual noise reduction: We attempted to suppress residual stationary noise remaining in $\hat{S}_{\mathrm{m}}$, in which the directional noise and the diffuse noise were approximately removed, by general single-channel Wiener filtering.

The residual noise was estimated as the signals in the non-speech parts that were detected with both the coherences between the outputs of the null beamformers, computed in the diffuse noise reduction stage, and the signal powers. The single-channel Wiener filter $H_{\mathrm{s}}(\omega,k)$ was computed with such residual noise. A target source spectrum $\hat{S}_{\mathrm{s}}(\omega,k)$ was estimated as follows:

$$|\hat{S}_{\mathrm{s}}(\omega,k)| = H_{\mathrm{s}}(\omega,k)\cdot|\hat{S}_{\mathrm{m}}(\omega,k)| \tag{12}$$

4) Robot voice reduction: Synthesized speech utterances, which represented the voices of the robot, were played on the loudspeaker mounted on the robot body. The powers of these utterances observed at the microphones were not low because the loudspeaker was close to the microphones. In this case, these utterances were residual even after abovementioned three-stage noise reduction was carried out. These residual signals significantly degrade the performances of VAD used in single-channel Wiener filtering and speech recognition. We attempted to reduce these robot voices by the following time-frequency masking:

$$|\hat{S}(\omega,k)| = \begin{cases} |\hat{S}_{\mathrm{s}}(\omega,k)|, & \text{if } |S_{\mathrm{r}}(\omega,k)| < \eta \\ \gamma, & \text{otherwise} \end{cases} \tag{13}$$

where $S_{\mathrm{r}}(\omega,k)$ denotes a spectral component of the robot voice s_{r}; η denotes a threshold for time-frequency masking; and γ denotes a flooring constant. In this experiment, η and γ were arbitrarily determined so as to achieve the best performance.

A phase of the observed signal was given to the spectral amplitude $|\hat{S}(\omega,k)|$ in order to recover the time-domain signal.

III. EXPERIMENT

We conducted the following two experiments: the first point to be investigated was the performances of the directional, diffuse, and residual noise reduction under the condition that both the directional noise and the diffuse noise were observed simultaneously; and the second argument dealt with the performance of reducing the noise from the robot in the situation where the robot uttered while moving, such as the robot internal noise (i.e., the noise induced by the movement of the robot and the noise from the

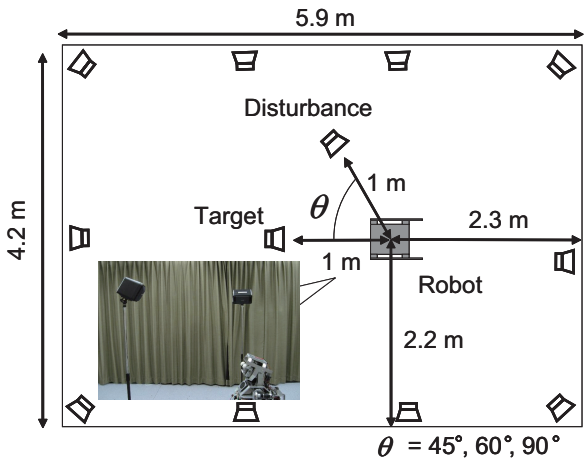

Fig. 9. Recording environment.

motors mounted on the robot) and the robot voices (i.e., the synthesized speech utterances played on the loudspeaker mounted on the robot).

A. Evaluation items and criteria

In the first experiment, the performances were investigated for the case without any processing (no-process), and the cases using four noise reduction methods, such as delay and sum (DS) method followed by Zelinski's post filtering, which was conventional multichannel Wiener filtering[9] (DS+MWF), generalized sidelobe canceller[10] (GSC), time-frequency masking using interchannel phase differences[11] (TFM), and the proposed method. In this case, DS+MWF, GSC, and TFM used the signals received by the Mic-2 and Mic-4. In this experiment, the performances of noise reduction were evaluated with the word accuracy, which is frequently used in assessments of automatic speech recognition systems, for separated speech utterances. The word accuracy was calculated in a common manner as follows:

$$\mathrm{WA} = \frac{N - D - S - I}{N} \times 100 \quad (\%) \tag{14}$$

where N, D, S, and I denote the number of words included in correct word sequences, deletion errors, substitution errors, and insertion errors, respectively.

In the second experiment, the performances of noise reduction were investigated for the case without any processing (no-process), the case using the directional, diffuse, and residual noise reduction in the proposed method (proposed-DDR), and the case using the robot voice reduction in addition to the proposed-DDR (proposed-DDRR). In this experiment, the performances were evaluated with noise attenuation (NA), which was calculated as follows:

$$\mathrm{NA} = 20 \log_{10} \frac{P_{\mathrm{proc}}}{P_{\mathrm{obs}}} \quad (dB) \tag{15}$$

where P_{proc} and P_{obs} are a temporal-averaged amplitude of the separated signal and that of the omni-directional microphone observation, respectively.

B. Speech materials

Figure 9 shows the recording environment. The microphones were placed on the head of the conversation robot "ROBISUKE"[12]. Both the distance between the target source and the robot and the distance between the disturbance and the robot were 100 cm. The target source was placed in a direction of $0°$ and the disturbance was placed in directions of $45°$, $60°$, and $90°$. The height of the target source and that of disturbance source were 153 cm. The robot head and the target loudspeaker were directed to each other as shown in Fig. 9.

The target speech utterances consisted of 100 sentences, which were spoken by 23 male speakers, taken from the Japanese newspaper article continuous speech database. As for the directional noise (i.e., disturbance speech utterances), 100 sentences were selected from the same database but different from the target speech utterances. In this case, a disturbance utterance was selected so as to be approximately same in duration and energy as the target utterance recorded simultaneously. Therefore, a SNR of the target speech utterance to the directional noise was approximately 0 dB.

The diffuse noise was simulated by playing the noise from a large air-conditioning machine on ten loudspeakers placed in a square round the room. The diffuse noise was recorded by the microphones on the robot head, and superposed on the target speech utterances with the directional noise so that a SNR of a target speech utterance to the diffuse noise would be 10 dB.

Five types of robot movement noise including the motor noise and 20 kinds of synthesized speech utterances by the robot were recorded separately. We had a total of 100 types of noise from the robot by superposing each type of robot movement noise on each robot speech utterance. These movements and speech utterances were used in the situations where the robot greet people.

TABLE I
EXPERIMENTAL CONDITION FOR SPEECH ANALYSIS.

sampling frequency	16 kHz
frame length	32 ms
frame shift	8 ms
analysis window	Hamming window
analysis range of frequencies	300–5500Hz

TABLE II
EXPERIMENTAL CONDITION FOR ACOUSTIC FEATURE EXTRACTION.

sampling frequency	16 kHz
frame length	25 ms
frame shift	10 ms
analysis window	Hamming window
pre-emphasis	$1–0.97z^{-1}$
feature parameters	12 MFCCs, 12 ΔMFCCs, and a Δlog energy

C. Experimental condition

Experimental conditions for noise reduction and acoustic feature extraction are shown in Tables I and II, respectively. Acoustic models were trained with 20414 sentences spoken by 133 male speakers, taken from the ASJ database, which consisted of Japanese newspaper article sentences (ASJ-JNAS) and phonetically-balanced sentences (ASJ-PB) recorded with close-talking microphones. We adopted tied-state triphones with 2000 states. The distribution function in each state of the models was represented by a 16-mixture Gaussian distribution with diagonal covariances. We used word trigram language models that were constructed using a lexicon with a vocabulary size of 20K.

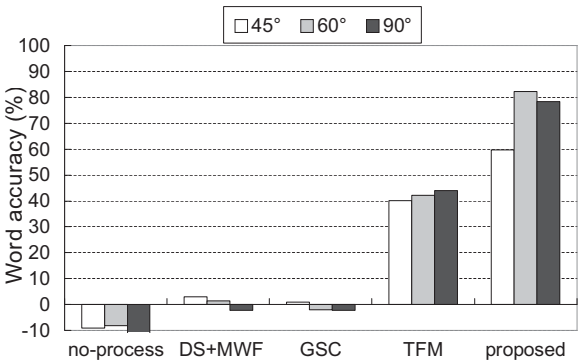

Fig. 10. Word accuracies for various noise reduction methods. A input sound consists of a target speech utterance, directional noise coming from the direction of 45°, 60°, or 90°, and diffuse noise.

TABLE III

NOISE ATTENUATIONS OF THE PROPOSED METHOD WITH AND WITHOUT ROBOT VOICE REDUCTION. A INPUT SOUND CONSISTS OF ROBOT INTERNAL NOISE AND A ROBOT SPEECH UTTERANCE.

Notations	Method	NA (dB)
proposed-DDR	w/o robot voice reduction	-8.7
proposed-DDRR	w/ robot voice reduction	-64.4

D. Experimental result

The results of the first investigation are shown in Fig. 10. This figure shows the word accuracies for various noise reduction methods. The proposed method achieved a word accuracy of better than 75% when the DOAs of the disturbance were 60° and 90°. In contrast, the word accuracy became below 0% without any noise reduction because insertion errors were significantly increased. In this case, speech recognition systems did not work. The word accuracies were not improved even when DS method followed by conventional multichannel Wiener filtering and GSC were applied. Since DS method, multichannel Wiener filtering, and GSC are difficult to achieve high performance under the condition that microphone spacings are small, the effectiveness of those methods was not observed in the case of the microphone arrangement we used. Time-frequency masking using interchannel phase differences gave good performances as compared to the DS method and the GSC. The performance of this method, however, was significantly degraded as compared to the proposed method.

The results of the second investigation are shown in Table III. This table shows noise attenuation (NA) of the proposed method with and without robot voice reduction over the case without any noise reduction. The proposed method with robot voice reduction eliminated the noise from the robot by approximately -64.4dB as compared to the case without any noise reduction. In contrast, the robot voices could not be reduced without robot voice reduction. In this case, the noise from the robot could be reduced by mere -8.7dB as compared to the case without any noise reduction.

IV. FACE DETECTION AND TRACKING USING IMAGE INFORMATION

In the present noise reduction system, it is assumed that the conversational partner of the robot is in front of the robot. When the target sound source moves from the frontal direction of the robot, it is localized by face detection and tracking with the facial images obtained from a camera mounted on an eye of the robot.

First, the robot seeks a conversational partner as a moving object. Such an object was detected by calculating optical flows over the captured image[13]. The robot then directs its eyes to the direction of that object. In this case, we assume that the detected object is the body or the waving hand of the partner. After that, the robot attempt to detect and track the face of the partner by reference to the position of the body and the waving hand. This face detection and tracking were carried out with the Haar-like features[14].

In the present study, a middle-ware for robot internal communication, "MONEA," was applied to the integration of multiple modules, such as an image processing module, a speech processing module, and a robot control module, and the timing control of these modules[15].

V. CONCLUSION

We proposed a new noise reduction method suitable for autonomous mobile robots, using the compact and light-weighted MEMS microphones and the low-computational-cost algorithm. The proposed method can cope with various types of noise, such as the directional noise, the diffuse noise, and the noise from the robot: the number of word errors was reduced by 69% as compared to the conventional method, and noise attenuation of -64.4dB in the noise from the robot was achieved. Moreover, we could developed a hands-free spoken dialogue system by integrating the proposed noise reduction system and the sound source localization system using face detection and tracking.

REFERENCES

[1] F. Asano *et al.*, "Speech enhancement based on the subspace method," IEEE Trans. Speech Audio Process., vol.SAP-8, no.5, pp.497-507, Sept. 2000.
[2] K. Nakadai *et al.*, "Applying scattering theory to robot audition system," Proc. IROS2003, pp.1147-1152, Oct. 2003.
[3] K. Nakadai *et al.*, "An open source software system for robot audition HARK and its evaluation," Proc. Humanoids2008, pp.561-566, Dec. 2008.
[4] M. Sato *et al.*, "Auditory system in a personal robot, PaPeRo," Proc. ICCE2006, pp.19-20, Jan. 2006.
[5] T. Ohkubo *et al.*, "Two-channel-based noise reduction in a complex spectrum plane for hands-free communication system," Journal of VLSI Signal Processing System 2007, Springer, vol. 46, issue2-3, pp.123-131, March 2007.
[6] H.-D. Kim *et al.*, "Two-channel-based voice activity detection for humanoids robots in noisy home environments," Proc. ICRA2008, pp.3495-3501, May 2008.
[7] S. Takada *et al.*, "Speech enhancement using square microphone array for mobile devices," Proc. ICASSP2008, pp.313-316, March 2008.
[8] T. Ogawa *et al.*, "Ears of the robot: Noise reduction using four-line ultra-micro omni-directional microphones mounted on a robot head," Proc. EUSIPCO2008, Aug. 2008.
[9] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," Proc. ICASSP1988, vol.5, pp.2578-2581, April 1988.
[10] L. J. Griffiths *et al.*, "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas and Propagation, vol.30, no.1, pp.27-34, Jan. 1982.
[11] O. Yilmaz *et al.*, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process., vol.52, no.7, July 2004.
[12] Y. matsuyama *et al.*, "Designing Communication Activation System in Group Communication," Proc. Humanoids2008, pp.629-634, Dec. 2008.
[13] B. Lucas *et al.*, "An Iterative Image Registration Technique with an Application to Stereo Visio," Proc. IJCAI, pp.674-679, July 1981.
[14] P. Viola *et al.*, "Robust real-time object detection," Intl. J. Computer Vision, vol.57, no.2, pp.137-154, Dec. 2004.
[15] T. Nakano *et al.*, "MONEA: Message-oriented networked robot architecture," Proc. ICRA2006, pp.194-199, May 2006.