16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

EARS OF THE ROBOT: NOISE REDUCTION USING FOUR-LINE ULTRA-MICRO OMNI-DIRECTIONAL MICROPHONES MOUNTED ON A ROBOT HEAD

Tetsuji Ogawa[1], *Hirofumi Takeuchi*[2], *Shintaro Takada*[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 type of noise reduction method suitable for autonomous mobile robots that require small and light-weighted devices and low-computational-cost algorithms. In the proposed method, four-line omni-directional micro electro mechanical system (MEMS) microphones are mounted on the robot head. The proposed method can reduce various kinds of noises simultaneously: directional noises are reduced by time-frequency masking using null beamformers and subtractive beamformers, and diffuse noises are reduced by multi-channel Wiener filtering using coherences. The effectiveness of the proposed method was shown in terms of speech recognition accuracies and speech qualities: the word error rate was reduced by 45% and 0.16 points of PESQ-MOS were improved compared with conventional time-frequency masking.

1. INTRODUCTION

We attempt to achieve high-performance hands-free speech recognition, which is a basis of robot audition, by performing noise reduction with a low computational cost using the compact and the light-weighted devices that can be mounted on an autonomous mobile robot.

For the case in which a robot makes a conversation with people in a real environment, only the target speaker's speech is required to be extracted from noisy speech that includes various kinds of noises, and then be precisely recognized. In order to achieve such a function for the autonomous mobile robot, both miniaturization and weight-saving are required for microphones and signal processing devices on which noise reduction and speech recognition are performed because the weight and the size of the devices are restricted so as to mount them on the robot. According to the miniaturization of the devices, low cost computations are also required.

Conventionally, noise reduction techniques using microphone arrays such as beamforming, blind source separation (BSS) and Wiener filtering are frequently applied to pre-processing of noisy speech recognition[1, 2]. These methods generally aim at reducing either only directional noises (e.g. BSS) or only diffuse noises (e.g. Wiener filtering). In addition, they require large number of microphones, large scale of microphone arrangements and high computational costs. Thus, they are not suitable for the noise reduction system implemented to the autonomous mobile robot in terms of the size and the computational costs. In addition, most of these methods assume the microphones are placed on a free-field. However, since effects of reflections and diffractions that occur around the robot head and the body cannot be ignored, they are difficult to give good performances of speech separation and recognition.

Aiming at coping with the reflections and the diffractions deriving from the robot, precise head related transfer functions (HRTFs) were measured in all possible areas around the robot[3]. However, the measurement of the HRTFs for each robot and each arrangement of the microphones is indeed troublesome work. The HRTFs were geometrically calculated by making an approximation on the shape of their robot head: they regard it as a simple sphere[4]. However, in most cases, the robot heads are far from spherical.

We proposed a new type of directional noise reduction method using four-line directional microphones mounted on the robot head that is free from strict HRTF measurements[5]. However, this method is difficult to apply to micro electro mechanical systems (MEMS) technologies because it requires the directional microphones. Thus, in this method, the microphones and the signal processing devices are difficult to be simply miniaturized. In addition, it could not explicitly cope with diffuse noises.

In the present paper, we propose a new type of noise reduction method using omni-directional microphones that are suitable for the MEMS technologies. In the proposed method, four-line MEMS omni-directional microphones are placed on the top of the robot head aiming at suppressing the influences of the HRTFs. The proposed method can simultaneously reduce both the directional noise and the diffuse noise using the low-computational-cost algorithms in which the directional noise is reduced on the basis of time-frequency masking[6] and the diffuse noise is reduced on the basis of coherences[7].

The rest of the present paper is organized as follows. The microphone system we used is described in section 2. In section 3, the algorithm of the proposed noise reduction method is described in detail. Section 4 gives conditions and results of noise reduction experiments in a real environment. Finally in section 5, we give conclusions.

2. MICROPHONE SYSTEM

We use the compact and the light-weighted microphones and signal processing devices, which are suitable for autonomous mobile robots.

2.1 MEMS microphone

We use four-line analog MEMS microphones, which are constructed on the basis of a semiconductor integrated technology and are significantly compact and light-weighted. We used SPM0208HD5 made by Knowles Co., Ltd. The width, the depth and the 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 which consists of a MEMS microphone and peripheral circuits including a pre-amplifier. These substrates are mounted on the robot head.

2.2 Microphone arrangement

As depicted in Fig. 1, the microphones are placed on the top of the robot head. This microphone arrangement aims at suppressing the influences of the reflections and the diffractions that occur around the robot. The microphones are arranged in a squared form in which each microphone spacing of neighboring microphones is 4 cm and the spacing of microphones in a diagonal position is 5.66 cm. Channels of the microphones are defined as described in Fig. 2. The front, the right and the left direction of the robot are defined as zero, positive and negative degrees, respectively. In the present paper, the target speech is assumed to arrive from the front of the robot.

2.3 A/D conversion system

Four channel analog signals received from the MEMS microphones are converted to digital signals using a compact embedded device.

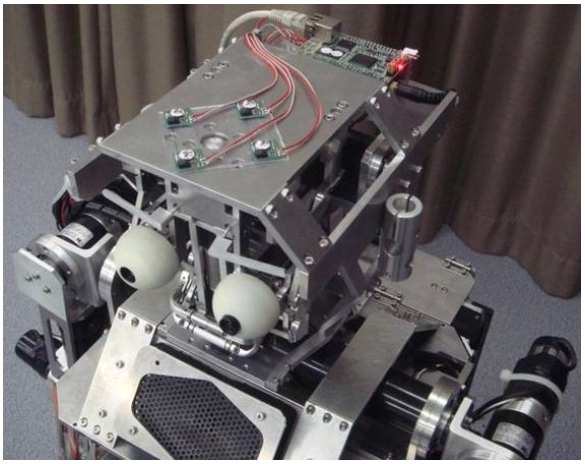

Figure 1: The robot and the microphones mounted on the robot head.

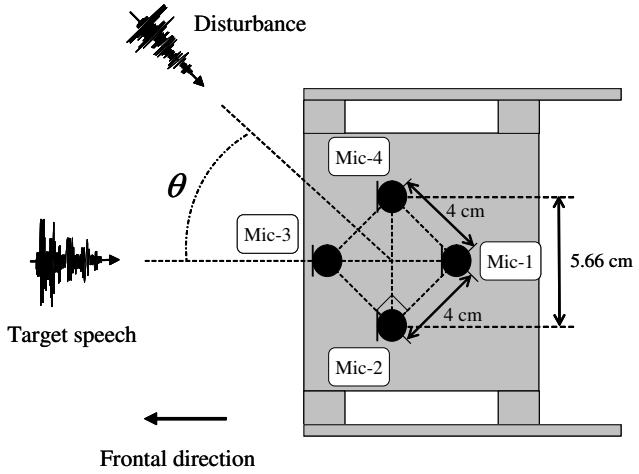

Figure 2: A microphone arrangement. This figure shows the top view of the robot.

The device consists of SUZAKU-V.SZ310 and SID00-U00, both of which are made by Atmark Techno Co., Ltd. SUZAKU-V.SZ310 is an universal embedded device platform, which is based on the combination of FPGA and Linux (with PowerPC405 CPU core) and mounts a 10BASE-T/100BASE-TX Ethernet connector. SID00-U00 is an eight-channel A/D conversion system used as an extension of the SUZAKU board. A resolution of the SID00-U00 is refined from original 12 bit to 16 bit. The digital signals are transfered to a laptop PC mounted on the robot through Ethernet. Then, noise reduction is performed on the laptop PC.

3. NOISE REDUCTION SYSTEM

Figure 3 illustrates a diagram of the proposed noise reduction method. The proposed method consists of three stage signal processing: 1) time-frequency masking for directional noise reduction, 2) multi-channel Wiener filtering for diffuse noise reduction and 3) single-channel Wiener filtering for residual noise reduction. Here, diffuse noise reduction and residual noise reduction are performed by the method that we have already proposed[7].

3.1 Directional noise reduction

Figure 4 illustrates a diagram of directional noise reduction. In the

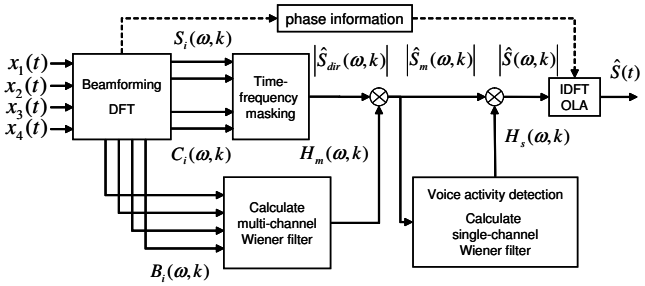

Figure 3: Block diagram of the proposed noise reduction system.

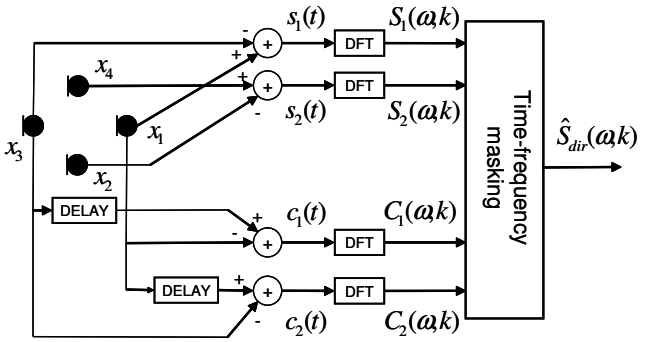

Figure 4: Directional noise reduction method.

present paper, $x_i(t)$ denotes a signal received from the microphone Mic-i at a discrete time of t, and $X_i(\omega,k)$ denotes a STFT coefficient of the x_i, where k and ω denotes a discrete frame and a discrete frequency, respectively.

In the proposed directional noise reduction system, outputs of multiple beamformers such as null beamformers and subtractive beamformers are positively utilized. $c_1(t)$ and $c_2(t)$ represent output signals of the null beamformers generated by performing delay addition and subtraction for the signals received from the Mic-1 and the Mic-3, $x_1(t)$ and $x_3(t)$. $c_1(t)$ and $c_2(t)$ are calculated as follows.

$$c_1(t) = x_3(t-\tau_d) - x_1(t) \quad (1)$$
$$c_2(t) = x_1(t-\tau_d) - x_3(t) \quad (2)$$

where τ_d denotes a delay corresponding to the spacing of the microphones arranged in a diagonal position. The directivity patterns of c_1 and c_2 are described in Fig. 5. Here, c_1 and c_2 forms the directivity pattern that has a null for the direction of 0° and 180°, respectively.

$s_1(t)$ and $s_2(t)$ represents the output signal of the subtractive beamformer formed by using $x_1(t)$ and $x_3(t)$ and that of the subtractive beamformer formed by using $x_2(t)$ and $x_4(t)$, respectively. $s_1(t)$ and $s_2(t)$ are obtained as follows.

$$s_1(t) = x_1(t) - x_3(t) \quad (3)$$
$$s_2(t) = x_4(t) - x_2(t) \quad (4)$$

The directivity patterns of s_1 and s_2 are described in Fig. 6. s_1 forms the directivity pattern that has maximum gains for the directions of 0° and 180° and nulls for the directions of 90° and -90°. s_2 forms the directivity pattern that has maximum gains for the directions of 90° and -90° and nulls for the directions of 0° and 180°.

The signals that come from the direction of the front of the

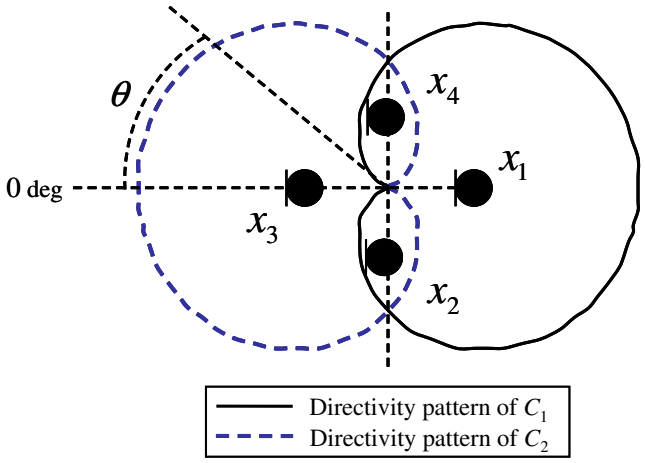

Figure 5: Directivity patterns of null beamformers.

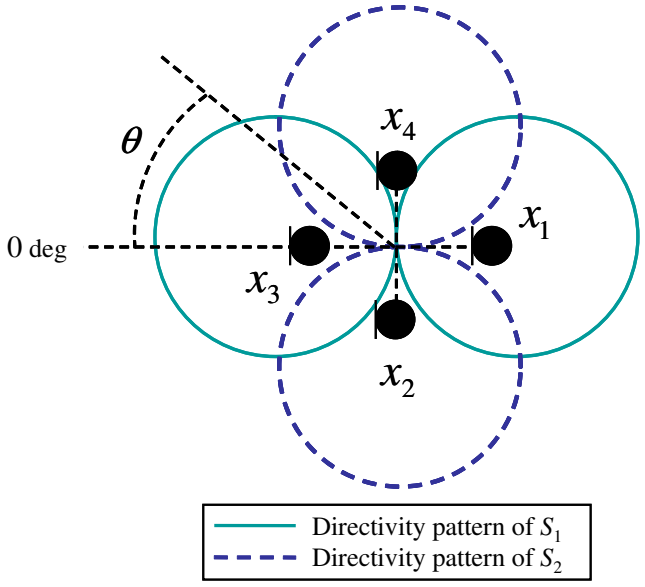

Figure 6: Directivity patterns of subtractive beamformers.

robot, $\hat{S}^{dir}$, are extracted by following time-frequency masking.

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

where $S_i(\omega,k)$ and $C_i(\omega,k)$ denotes a STFT coefficient of s_i and c_i, respectively, and β denotes a flooring constant.

In this time-frequency masking, the directional noise coming from the side of the robot is suppressed by selecting the time-frequency components in which the $S_1(\omega,k)$ is larger than the $S_2(\omega,k)$, and then the noise from the backward of the robot is suppressed by selecting the components in which the $C_2(\omega,k)$ is larger than the $C_1(\omega,k)$.

3.2 Diffuse noise reduction

Diffuse noises included in the $\hat{S}_{dir}$ are reduced by multi-channel Wiener filtering.

At first, four null beamformer outputs are calculated as follows.

$$b_1(t) = x_2(t-\tau_n) - x_1(t) \tag{6}$$
$$b_2(t) = x_3(t-\tau_n) - x_2(t) \tag{7}$$
$$b_3(t) = x_3(t) - x_4(t-\tau_n) \tag{8}$$
$$b_4(t) = x_4(t) - x_1(t-\tau_n) \tag{9}$$

where τ_n denotes a delay corresponding to the microphone spacing of neighboring microphones. While a conventional approach computes the multi-channel Wiener filter $H_m(\omega,k)$ using just the observations at omni-directional microphones[10], the proposed method computes it using the null beamformer outputs as follows.

$$H_m(\omega,k) = \frac{\frac{1}{2}\sum[\mathrm{abs}\{B_p(\omega,k)B_q^*(\omega,k)\}]}{\frac{1}{4}\sum_{r=1}^{4}[B_r(\omega,k)B_r^*(\omega,k)]} \tag{10}$$

where $B_r(\omega,k)$ denotes a STFT coefficient of b_r $(r=1,2,3,4)$ described in Eq.(6)-Eq.(9). For the case in which the microphone spacings are small, the proposed method, which uses the null beamformer outputs, can reduce theoretical magnitude-squared coherences in diffuse noise fields compared to the conventional multi-channel Wiener filtering, and thus is expected to improve the performance of diffuse noise reduction[7]. In Eq.(10), p and q are selected as $\{(p,q)\} = \{(1,2),(3,4)\}$ so that the null beamformers used in the calculation of the numerator can form line-symmetric directivity patterns to the axis containing the target source and the center of the microphones, where the difference between the directivity of B_p and that of B_q is just 90 degrees. Here, the correlation among the diffuse noise components is expected to be reduced.

By using the multi-channel Wiener filter described in Eq.(10), the amplitude spectrum of the signal in which the diffuse noise is suppressed is estimated as follows.

$$|\hat{S}_m(\omega,k)| = H_m(\omega,k)\cdot|\hat{S}_{dir}(\omega,k)| \tag{11}$$

3.3 Residual noise reduction

Residual stationary noises remaining in the signals $\hat{S}_m$, in which the directional and the diffuse noise are approximately removed, attempt to be suppressed by general single-channel Wiener filtering.

The residual noises are estimated as the signals in the non-speech parts that are detected using both the coherences calculated in the diffuse noise reduction stage and the signal powers. Then, the Wiener filter $H_s(\omega,k)$ is calculated using these residual noises. The target source $\hat{S}(\omega,k)$ can be estimated as follows.

$$|\hat{S}(\omega,k)| = H_s(\omega,k)\cdot|\hat{S}_m(\omega,k)| \tag{12}$$

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

4. NOISE REDUCTION EXPERIMENT

In order to evaluate the effectiveness of the proposed method, experimental comparisons were conducted under noisy conditions in which both directional noises and diffuse noises exist. Noise reduction systems were evaluated using the automatic speech recognition performance based on the word accuracy and the speech quality based on the PESQ-based MOS[8]. The word accuracy was calculated in a common manner as follows.

$$WA = \frac{N-D-S-I}{N} \times 100 \quad (\%) \tag{13}$$

where N, D, S and I represent the number of words included in correct word sequences, the number of deletion errors, the number of substitution errors and the number of insertion errors, respectively. PESQ-MOS were calculated using reference signals that were observed at the microphones for the case in which only the target source existed.

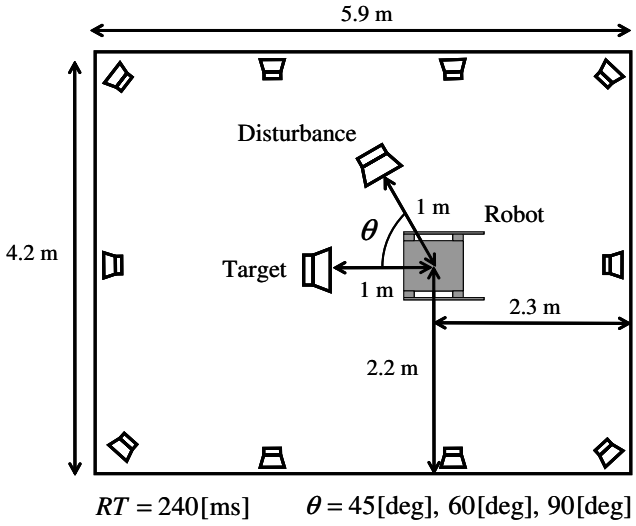

Figure 7: Recording environment.

4.1 Speech materials

Figure 7 illustrates the recording environment. 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 the direction of 0° and the disturbance was placed in the direction of 45°, 60° and 90°.

The target speeches consisted of 100 sentences, which were spoken by 23 male speakers, from the Japanese-Newspaper-Article continuous speech database[13]. As for the directional noises (disturbance speeches), 100 sentences that were different from the target speeches were also selected from the same database. Here, a disturbance utterance was selected so as to be approximately the same in the duration as the corresponding target utterance. In addition, the energy of the disturbance speech was adjusted so that the temporal-averaged energy of the disturbance speech would become the same as that of the corresponding target speech. Thus, a SNR of the target speech to the directional noise was approximately 0 dB. On the other hand, the diffuse noise was simulated by playing back the ambient noise of a large air-conditioning machine from ten loudspeakers placed in a square round the room. Then, the diffuse noise recorded at the microphones on the robot head was superimposed on the target speech with the directional noise so that a SNR of the target speech to the diffuse noise would be just 15 dB.

4.2 Evaluation items

The performances were investigated for a) the case of non-processing, and four noise reduction methods as follows: b) delay and sum (DS) method followed by Zelinski's post filtering, which was conventional multi-channel Wiener filtering[10], using four-channel signals (4ch-DS+MWF), c) generalized sidelobe canceller (GSC)[9] using four-channel signals (4ch-GSC), d) time-frequency masking based on phase differences between microphones using two-channel signals (2ch-TFmasking)[11], and e) the proposed method. Here, 2ch-TFmasking used two-channel signals received from the Mic-2 and the Mic-4. The range of the target source direction in 2ch-TFmasking was set to $\pm 20°$.

4.3 Experimental condition

Analysis setup for noise reduction is shown in Table 1.

Analysis setup for speech recognition is shown in Table 2. Acoustic models were trained with 20414 sentences spoken by 133 male speakers from the ASJ database, which consisted Japanese newspaper article sentences (ASJ-JNAS) and phoneme balanced sentences (ASJ-PB) recorded with close-talking microphones[13]. We adopted state-tied triphones in which the number of the states was 2000 and the distribution function in each state was represented by a 16-mixture Gaussian distribution with diagonal covariances. As for a language model, we used word trigrams that were constructed by using a lexicon of 20K vocabulary.

Table 1: Setup for noise reduction.

sampling frequency	16 kHz
frame length	32 ms
frame shift	8 ms
analysis window	Hamming window
analysis range of frequencies	300 - 5600 Hz

Table 2: Setup for speech recognition.

sampling frequency	16 kHz
frame length	25 ms
frame shift	10 ms
analysis window	Hamming window
feature parameters	12 MFCCs, 12ΔMFCCs, Δlog energy

4.4 Experimental results

The results based on the word accuracies and those based on the PESQ-MOS are shown in Fig.8 and Fig.9, respectively.

As depicted in Fig.8, the proposed method (e) achieved a word accuracy of better than 75% for the case in which the direction of arrival (DOA) of the disturbance was larger than 60°. Since the number of the insertion errors in Eq. (13) was significantly increased without any noise reduction (a), the word accuracy became below 0%. Thus, in this case, speech recognition did not work. The word accuracies were not improved even if four-channel DS method followed by conventional multi-channel Wiener filtering (b) and four-channel GSC (c) were applied. Time-frequency masking based on the phase differences between the microphones (d) gave good performances compared to the DS method and the GSC. It gave almost the same performance as the proposed method for the case in which the DOA of the disturbance was 45°. However, for the case in which the DOA of the disturbance was larger than 60°, the performance of this method was significantly degraded compared to the proposed method.

In addition, as depicted in Fig.9, the proposed method could improve also the speech qualities compared to the conventional methods with the similar tendency as the results based on the word accuracies.

For the case in which the performances were averaged for the DOAs of the disturbance, the proposed method reduced word errors of 45% and improved 0.16 points of PESQ-MOS compared to time-frequency masking, which gave the best performance in the compared conventional methods.

5. CONCLUSION

We proposed a new type of noise reduction method suitable for autonomous mobile robots, which used the compact and the light-weighted MEMS microphones and the low-computational-cost algorithms. The proposed method can cope with various kinds of noises such as the directional noise and the diffuse noise. The experimental results in the real environment including both the directional and the diffuse noise showed the effectiveness of the proposed method in terms of the word accuracy and the PESQ-MOS: 45% of word errors of the conventional method was reduced and 0.16 points of PESQ-MOS were improved compared to the conventional method.

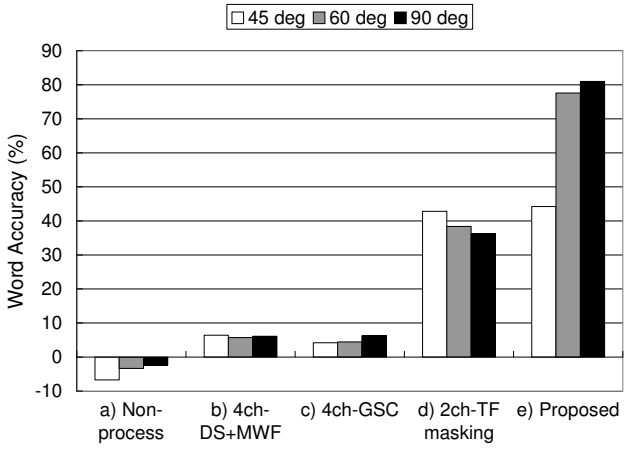

Figure 8: Word accuracy for evaluation items. Input signal consists of the target speech, the directional noise coming from the direction of 45°, 60° or 90° with a SNR of 0 dB and the diffuse noise with a SNR of 15 dB.

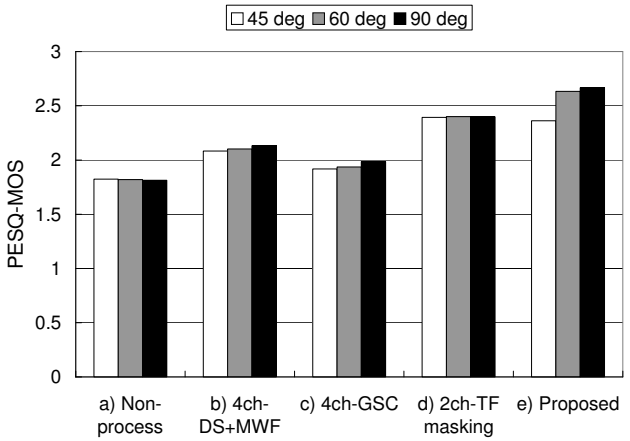

Figure 9: PESQ-MOS for evaluation items. Input signal consists of the target speech, the directional noise coming from the direction of 45°, 60° or 90° with a SNR of 0 dB and the diffuse noise with a SNR of 15 dB.

REFERENCES

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," IEEE ASSP Magazine, vol.5, pp.4-24, 1988.

[2] M. S. Brandstein and D. B. Ward, "Microphone arrays: signal processing techniques and applications," Springer-Verlag, Berlin, 2001.

[3] F. Asano, S. Hayamizu, T. Yamada and S. Nakamura, "Speech enhancement based on the subspace method," IEEE Trans. Speech Audio Processing, vol.SAP-8, no.5, pp.497-507, Sept. 2000.

[4] K. Nakadai, D. Matusura, H. G. Okuno and H. Kitano, "Applying scattering theory to robot audition system," Proc.IROS, pp.1147-1152, Oct. 2003.

[5] N. Mochiki, T. Sekiya, T. Ogawa and T. Kobayashi, "Recognition of three simultaneous utterance of speech by four-line directivity microphone mounted on head of robot," Proc. ICSLP, pp.821-824, Oct. 2004.

[6] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," J. Acoustic. Soc, vol.22, no.2, pp.149-157, March 2001.

[7] S. Takada, T. Ogawa, K. Akagiri and T. Kobayashi, "Speech enhancement using square microphone array for mobile devices," Proc. ICASSP, pp.313-316, March 2008.

[8] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Dec. 2001.

[9] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas and Propagation, vol. 30, no. 1, pp. 27-34, 1982.

[10] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," Proc. ICASSP, vol.5, pp.2578-2581, 1988.

[11] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. on Signal Processing, vol.52, no.7, July 2004.

[12] S. Fujie, T. Yamahata and T. Kobayashi, "Conversation robot with the function of gaze recognition," Proc. Humanoids, pp.364-369, Dec. 2006.

[13] K.Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, K. Shikano, T. Kobayashi and S. Itahashi, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," Proc. ICSLP, pp.3261-3264, Nov. 1998.