SPEECH ENHANCEMENT USING NULLSPACE-BASED SOUND FIELD CONTROL FOR BARGE-IN FREE SPOKEN DIALOGUE INTERFACE

Shigeki Miyabe,[1] *Hiroshi Saruwatari,*[1] *Kiyohiro Shikano,*[1] *and Yosuke Tatekura*[2]

[1] Nara Institute of Science and Technology, 8916-5, Takayama-cho, Ikoma-shi, Nara-ken, Japan
[2] Shizuoka University, 3-5-1, Johoku, Hamamastsu-shi, Shizuoka-ken, Japan

ABSTRACT

This paper describes a new small-scale interface for a barge-in free spoken dialogue system combining a multichannel sound field control and a microphone array, in which the response sound from the system can be canceled out at the microphone points. The conventional method inhibits the user from moving because the system forces the user to stay in the fixed position where the response sound is reproduced. However, since the proposed method doesn't arrange the control points for the reproduction of the response sound to the user, the user's move is allowed. Furthermore, relaxation of the strict reproduction for the response sound enables us to design a stable system with fewer loudspeakers than the conventional method. The proposed method shows higher performances in the speech recognition experiments.

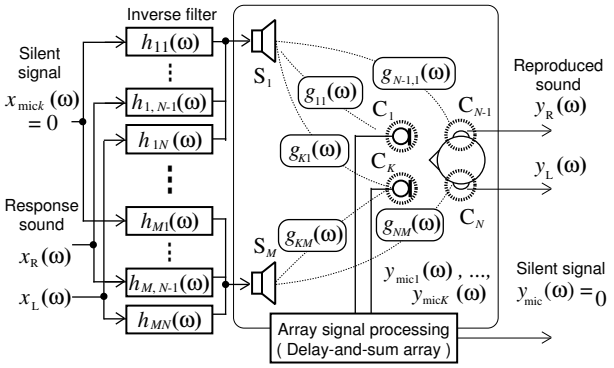

Fig. 1. Configuration of conventional MOMNI method.

1. INTRODUCTION

In human-machine communication based on a spoken dialogue system, it is desirable that the user can input his speech without wearing special equipments or being forced to stay in a particular position. In addition, the system should receive the user's speech even when the system speaks. However, when both the system and the user speak simultaneously, we cannot sufficiently reduce the response sound inputted into a microphone for recording user's speech. Therefore there occurs a problem that the speech recognition performance of the user's speech is degraded. This problem is referred to as *barge-in* [1].

In order to eliminate the response sound of the system, an acoustic echo canceller is commonly used. Many types of acoustic echo cancellers have been proposed, e.g., single channel, stereophonic, wave synthesis and integrated with a beamformer [3, 2, 5, 4] . However, the acoustic echo canceller has the inherent problem that the accurate adaptation is difficult in the barge-in situation (this is also called "double-talk problem"). Because of the problem, the conventional acoustic echo canceller should stop the adaptation in the barge-in duration; this implies that the elimination performance is likely to degrade when the change of transfer functions arises in the barge-in duration. In order to solve the problem of the acoustic echo canceller, one of the authors has proposed Multiple-Output and Multiple-Noise-Input (MOMNI) method [6] which combines sound field control and microphone array techniques. By increasing the number of the loudspeakers and the microphone elements, MOMNI method can make its control robust against the change of transfer functions, but huge number of loudspeakers are needed to achieve enough robustness for speech recognition. Furthermore, MOMNI method controls the sound field around user's ears and premises that the user doesn't move from the specific position.

To solve the problems of MOMNI method, we introduce a new method to stably realize a silent zone around the microphones with fewer loudspeakers and to remove the control points on the user's ears. The feasibility of the proposed algorithm can be shown by speech recognition experiments performed in a real acoustic environment. In addition, we discuss the efficacy of known noise superposition for further improvement.

This work was partly supported by CREST program "Advanced Media Technology for Everyday Living" of JST in Japan.

2. CONVENTIONAL MOMNI METHOD

We describe the MOMNI method shown in Fig. 1. The MOMNI method consists of two main parts, namely, sound field control and a microphone array.

2.1. Sound Field Control and Microphone Array

In Fig. 1, S_m $(m = 1, \ldots, M)$ are the loudspeakers which act as secondary sound sources, and C_n $(n = 1, \ldots, N)$ are the microphones which act as control points. $\mathrm{C}_1, \ldots, \mathrm{C}_K$ $(K = N - 2)$ are located in each of microphone elements for recording the user's speech, and C_{N-1} and C_N are placed in the vicinity of the two external auditory meatus of a user. Here the relation between the number of loudspeakers and that of the microphones must satisfy the condition $M > N = K + 2$. The intended signals to be reproduced at respective control points are represented by $\boldsymbol{x}(\omega)$=$[x_{\mathrm{mic}1}(\omega), \ldots, x_{\mathrm{mic}K}(\omega),\ x_{\mathrm{R}}(\omega),\ x_{\mathrm{L}}(\omega)]^{\mathrm{T}}$, where $x_{\mathrm{mic}k}(\omega)$ $(k = 1, \ldots, K)$, $x_{\mathrm{R}}(\omega)$ and $x_{\mathrm{L}}(\omega)$ are the signals to be reproduced at microphones C_k, the right and left ears of a user, respectively. Similarly, the observation signals at the control points are described as $\boldsymbol{y}(\omega) = [y_{\mathrm{mic}1}(\omega), \ldots, y_{\mathrm{mic}K}(\omega), y_{\mathrm{R}}(\omega), y_{\mathrm{L}}(\omega)]^{\mathrm{T}}$. We measure all the room transfer functions g_{nm} $(n = 1, \ldots, N, m = 1, \ldots, M)$ between S_m and C_n, and we denote them with an $N \times M$ matrix $\boldsymbol{G}(\omega)$. The $M \times N$ inverse filter matrix composed of the filter coefficients $h_{mn}(\omega)$ is expressed as $\boldsymbol{H}(\omega)$[7]. Then $\boldsymbol{y}(\omega)$ is denoted by

$$\boldsymbol{y}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{H}(\omega)\boldsymbol{x}(\omega), \tag{1}$$

where $\boldsymbol{G}(\omega)\boldsymbol{H}(\omega) = \boldsymbol{I}_N$, and $\boldsymbol{I}_N$ is the $N \times N$ identity matrix.

In Eq. (1), the response sounds of a dialogue system are reproduced at both ears of the user ($[y_{\mathrm{L}}, y_{\mathrm{R}}] = [x_{\mathrm{L}}, x_{\mathrm{R}}]$), and silent zones are materialized at each microphone elements ($[y_{\mathrm{mic}1}, \ldots, y_{\mathrm{mic}K}] = [0, \ldots, 0]$). Thereby, we can actualize the sound field which gives a user the response sound while preventing it from mixing into the observation signal at each microphone element.

Finally, in order to enhance the user's speech, delay-and-sum array signal processing is applied to the observed signals on the microphone elements.

2.2. Inverse Filter Design for Sound Field Control

In a multipoint control system based on loudspeakers, we must consider the influence of the room transfer functions. For this reason, we design the inverse filter $\boldsymbol{H}(\omega)$ by applying the least norm solution (LNS) in the frequency domain [8] so that the input signal $x_n(\omega)$ is observed only at C_n. In the case where the rank of $\boldsymbol{H}(\omega)$ is not decreased, since the solution of $\boldsymbol{H}(\omega)$ is indeterminate, we adopt the Moore-Penrose generalized inverse matrix as the inverse filter which provides the LNS [6].

2.3. Microphone Array Based on Delay-and-Sum Array Signal Processing

In multichannel speech enhancement, the delay-and-sum array is commonly used. To obtain the user's speech at the array output, we compensate for the time delay at each element and add the signals together to reinforce the target signal arriving from the look direction. The phase compensation filter $A_k(\omega)(k = 1, 2, \ldots, K)$ at the k-th element of a delay-and-sum array is designated as

$$A_k(\omega) = (1/K) \cdot e^{-j\omega\tau_k}, \tag{2}$$

where τ_k is the arrival time difference of the target signal between the source and the position of the k-th element. Thus, the array output $Y_{\mathrm{mic}}(\omega)$ is given by

$$Y_{\mathrm{mic}} = \sum_{k=1}^{K} A_k(\omega) Y_{\mathrm{mic}K}(\omega). \tag{3}$$

2.4. Response Sound Elimination Error When Changing Room Transfer Functions

In [6], it is shown that the elimination error of response sound is proportional to $1/\sqrt{M \cdot K}$. Therefore, if the number of transfer channels between loudspeakers and microphones is increased, the MOMNI method becomes more robust against the change of transfer functions than an acoustic echo canceller.

2.5. Problems in MOMNI Method

Since the MOMNI method must satisfy the condition $M > N = K + 2$, two additional loudspeakers are required for the control of the sound field at both of the user's ears. If we want to construct a small scale system with few loudspeakers, setting of these two control points at the user's ears is a barrier to secure enough number of microphone elements for robustness. Moreover, if we promise that the user can move around, the strict reproduction at these two control points becomes meaningless. Therefore, in order to realize the user's movement and reduction of the scale of the system, these two control points should be discarded. However, the MOMNI method cannot present the response sound to the user without these control points because each input signal must correspond to each of the control points.

3. PROPOSED METHOD: RESPONSE SOUND CANCELLATION

In this section, we propose a new filter design algorithm to provide silent zones on the microphone elements without setting representing points. Since no other control points than the microphone elements are settled, the sound field control can be performed stably with fewer loudspeakers.

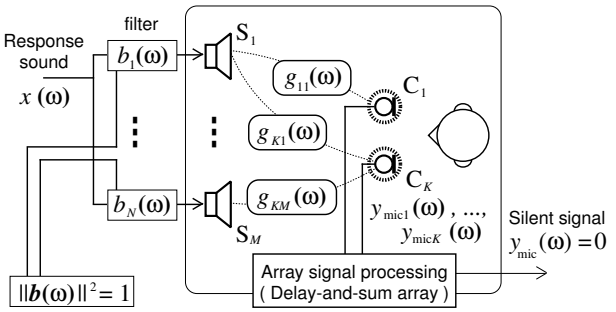

Fig. 2. Configuration of proposed method.

3.1. Sound Field Control to Cancel Out Response Sound

In Fig. 2, S_m ($m = 1, \ldots, M$) denote the loudspeakers, and C_k ($k = 1, \ldots, K$) represent the microphones. The numbers of loudspeakers and microphone elements must satisfy the condition $M > K$. The observed signals at the control points are designated as $\boldsymbol{y}(\omega) = [y_1(\omega), \ldots, y_K(\omega)]^{\mathrm{T}}$, where $y_k(\omega)$ ($k = 1, \ldots, K$) are the signals observed on the microphones C_k. A Response sound is monaural, and denoted by a scalar $x(\omega)$. The response sound is outputted from the loudspeakers after processed by filters. The filter coefficients are represented by $\boldsymbol{b}(\omega) = [b_1(\omega), \ldots, b_M(\omega)]^{\mathrm{T}}$, where $b_m(\omega)$ ($m = 1, \ldots, M$) are the filters which correspond to the loudspeakers S_m. The $M \times K$ matrix, composed of the room transfer functions $g_{km}(\omega)$ between the loudspeakers S_m and the control points C_k, which are measured in advance, is denoted by $\boldsymbol{G}(\omega)$. Then $\boldsymbol{y}(\omega)$ is denoted by

$$\boldsymbol{y}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{b}(\omega)x(\omega). \tag{4}$$

Therefore, the following condition must be satisfied when any response sounds are canceled out on the positions of microphone elements;

$$\boldsymbol{G}(\omega)\boldsymbol{b}(\omega) = \boldsymbol{0}, \tag{5}$$

$$\text{subject to} \quad \|\boldsymbol{b}(\omega)\| = C, \tag{6}$$

where $\boldsymbol{0}$ is a K-dimensional column zero vector and C is a constant to adjust the gain. The norm of $\boldsymbol{b}(\omega)$ is constrained to fix the total gain of the filters and to avoid the trivial filter coefficients which output no signal.

After this step, the recorded signals are applied to delay-and-sum array signal processing, and the user's speech is emphasized.

3.2. Extraction of Vectors Which Span Nullspace

Equation (5) shows that $\boldsymbol{b}(\omega)$ is orthogonal to all rows of $\boldsymbol{G}(\omega)$. The M-dimensional subspace which includes all orthogonal vectors to all rows of $\boldsymbol{G}(\omega)$ is called *nullspace* of $\boldsymbol{G}(\omega)$. Singular value decomposition provides the vectors which span the nullspace of $\boldsymbol{G}(\omega)$ in the form of eigenvectors which correspond to zero singular values. The filter coefficients $\boldsymbol{b}(\omega)$ can be designed by linear combination of these vectors.

Singular value decomposition of $\boldsymbol{G}(\omega)$ is denoted by

$$\boldsymbol{G}(\omega) = \boldsymbol{U}(\omega) \left[\begin{array}{c|c} \boldsymbol{\Lambda}_{R_\omega}(\omega) & \boldsymbol{O}_{R_\omega, M-R_\omega} \\ \hline \boldsymbol{O}_{K-R_\omega, R_\omega} & \boldsymbol{O}_{K-R_\omega, M-R_\omega} \end{array} \right] \boldsymbol{V}^{\mathrm{H}}(\omega), \tag{7}$$

where R_ω is the rank of $\boldsymbol{G}(\omega)$, $\boldsymbol{O}_{i,j}$ is an $i \times j$ zero matrix, and $\cdot^{\mathrm{H}}$ represents the conjugate transposed matrix. $\boldsymbol{\Lambda}_{R_\omega}(\omega)$ is an $R_\omega \times R_\omega$ diagonal matrix whose diagonal elements $\{\lambda_1(\omega), \ldots, \lambda_{R_\omega}(\omega)\}$ are singular values of $\boldsymbol{G}(\omega)$. $\boldsymbol{U}(\omega)$ and $\boldsymbol{V}(\omega)$ are $K \times K$ and $M \times M$ unitary matrices, whose column vectors $\{\boldsymbol{u}_1(\omega), \ldots, \boldsymbol{u}_{R_\omega}(\omega)\}$ and $\{\boldsymbol{v}_1(\omega), \ldots, \boldsymbol{v}_{R_\omega}(\omega)\}$ are eigenvectors corresponding to the singular values $\{\lambda_1(\omega), \ldots, \lambda_{R_\omega}(\omega)\}$, respectively, and the rest of vectors $\{\boldsymbol{u}_{R_\omega+1}(\omega), \ldots, \boldsymbol{u}_K(\omega)\}$ and $\{\boldsymbol{v}_{R_\omega+1}(\omega), \ldots, \boldsymbol{v}_M(\omega)\}$ are the eigenvectors corresponding to the zero singular value. In particular, $\{\boldsymbol{v}_{R_\omega+1}(\omega), \ldots, \boldsymbol{v}_M(\omega)\}$ are the nullspace vectors we need. These vectors certainly exist because of the condition $M - R_\omega \geq M - K > 0$.

Any $\boldsymbol{v}_r(\omega)$ ($r = R_\omega + 1, \ldots, M$) and any combination of them are orthogonal to all rows of $\boldsymbol{G}(\omega)$. We define $M \times (M - R_\omega)$ matrix $\boldsymbol{W}(\omega)$ as

$$\boldsymbol{W}(\omega) = [\underbrace{\boldsymbol{v}_{R_\omega+1}(\omega), \ldots, \boldsymbol{v}_M(\omega)}_{M-R_\omega}]. \tag{8}$$

Clearly, with any $M-R_\omega$ dimensional weight vector $\boldsymbol{\alpha}(\omega) = [\alpha_{R_\omega+1}(\omega), \ldots, \alpha_M(\omega)]$, M dimensional vector $\boldsymbol{b}'(\omega)$ which is given by

$$\boldsymbol{b}'(\omega) = \boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega) \tag{9}$$

satisfies the condition of Eq. (5). Thus the target filter $\boldsymbol{b}(\omega)$ can be obtained by normalizing $\boldsymbol{b}'(\omega)$ with its norm to satisfy the condition of Eq. (6), as

$$\boldsymbol{b}(\omega) = C\frac{\boldsymbol{b}'(\omega)}{\|\boldsymbol{b}'(\omega)\|} = C\frac{\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega)}{\sqrt{\boldsymbol{\alpha}^{\mathrm{H}}(\omega)\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega)}}. \tag{10}$$

Since $\boldsymbol{V}(\omega)$ is a unitary matrix,

$$\boldsymbol{v}_i^{\mathrm{H}}(\omega)\boldsymbol{v}_j(\omega) = \begin{cases} 1 & i = j, \\ 0 & i \neq j \end{cases}. \tag{11}$$

Accordingly,

$$\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{W}(\omega) = \boldsymbol{I}_{M-R_\omega}. \tag{12}$$

Substituting Eq. (12) in Eq. (10),

$$\boldsymbol{b}(\omega) = C\frac{\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega)}{\sqrt{\boldsymbol{\alpha}^{\mathrm{H}}(\omega)\boldsymbol{\alpha}(\omega)}}. \tag{13}$$

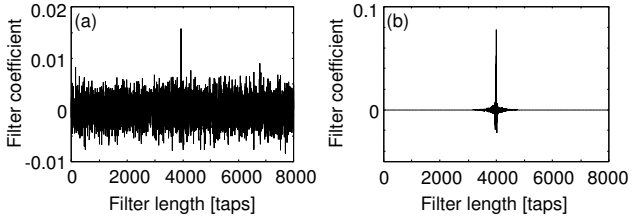

Fig. 3. An example of filter coefficients related to one loudspeaker designed by nullspace vectors selected randomly.

3.3. Optimal Filter Coefficients Closest to Impulses

In the previous section, we showed that the conditions of Eqs. (5) and (6) are satisfied with $\boldsymbol{b}(\omega)$ in Eq. (13), i.e., any appropriately normalized linear summation of nullspace vectors. However, the output sound becomes extremely distorted if the expansion coefficients $\boldsymbol{\alpha}(\omega)$ are selected randomly; indeed Fig. 3(a) shows an example of the filter designed with random combination of the nullspace vectors. In the following, we propose an algorithm for designing a filter with small distortion by utilizing the solution closest to impulses.

We define a following filter coefficients vector $\boldsymbol{l}(\omega)$, whose components are the filter coefficients of the impulses with the same amplitudes and the same latency τ;

$$\boldsymbol{l}(\omega) = e^{j\omega\tau}[\underbrace{1, \ldots, 1}_{M}]^{\mathrm{T}}. \tag{14}$$

Then we try to find the vector closest to the target vector $\boldsymbol{l}(\omega)$ in nullspace. We can obtain the optimal expanded coefficient vector $\boldsymbol{\alpha}(\omega)$ by solving the following least squares problem;

$$\min_{\boldsymbol{\alpha}(\omega)} ||\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega) - \boldsymbol{l}(\omega)||^2. \tag{15}$$

The output of each loudspeaker becomes less distorted because each filter coefficient becomes closest to the impulse, which has a property of full bandpass and linear phase. The solution of Eq. (15) satisfies the following condition:

$$\frac{\partial \|\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega) - \boldsymbol{l}(\omega)\|^2}{\partial \alpha_r^*(\omega)} = 0, \tag{16}$$

where $r = R_\omega + 1, \ldots, M$ and * shows complex conjugate. These simultaneous equations give the following solution

$$\boldsymbol{W}^{\mathrm{H}}(\omega)(\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega) - \boldsymbol{l}(\omega)) = 0. \tag{17}$$

Substituting Eq. (12), we have

$$\boldsymbol{\alpha}(\omega) = \boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega). \tag{18}$$

Finally, we can obtain the resultant filter coefficients $\boldsymbol{b}(\omega)$ by substituting Eq. (18) in Eq. (13), as

$$\boldsymbol{b}(\omega) = C\frac{\boldsymbol{W}(\omega)\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega)}{\sqrt{\boldsymbol{l}^{\mathrm{H}}(\omega)\boldsymbol{W}(\omega)\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega)}}. \tag{19}$$

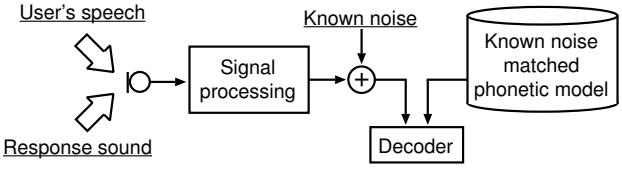

Fig. 4. Configuration of known noise superposition.

Figure 3(b) shows an example of the filter designed by the proposed method. We can find that its distortion is considerably lower than that of Fig. 3(a).

3.4. Known noise superposition [14]

In the previous section, we described response sound reduction procedures which utilized in the proposed system. There still, however, exists a residual component of the response sound caused by fluctuation of the transfer function. In order to achieve an optimum recognition performance, we generally need to create matched phonetic models for a speech decoder. However, without information of SNR in advance, an accurate construction of matched model is very difficult. To handle many different types of noise, known noise superposition has been proposed. We apply this technique in masking of the residual response sound as follows.

1. We superimpose known noise to speech database and make the corresponding matched model trained by EM algorithm in advance.
2. We superimpose known noise to the noise reduced output from the proposed system.
3. We perform speech recognition using known noise matched model for the system output.

Figure 4 shows a configuration of this process.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Conditions

In this section, we perform experiments of a large vocabulary continuous speech recognition task comparing the conventional methods (acoustic echo canceller and MOMNI method) with the proposed method. In order to verify the applicability of the proposed method, we simulate the change of room transfer functions and evaluate the performance of each method.

In the experiments, we premise that the fluctuation of transfer functions is caused by changes in the interference, i.e., a life-size mannequin. The interference is arranged under the assumption that another person approaches the user, which is a very common occurrence in real environments.

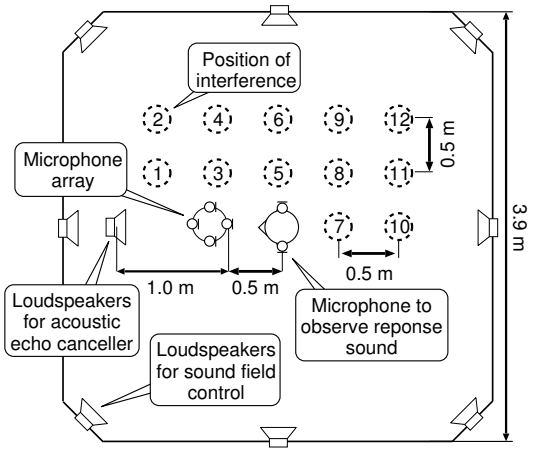

Fig. 5. Layout of acoustic experimental room.

Table 1. Experimental conditions for speech recognition

Training data	JNAS [10]
Frame length	25 ms (Hamming window)
Frame interval	8 ms
Feature vector	12 MFCCs, 12 ΔMFCCs, Δpower
Language model	Newspaper dictation[11]
Phoneme model	Phonetic Tied Mixture (PTM) [12] (clean or 25 dB noise superposed)
Decoder	Julius ver. 3.4.2 standard [13]
User's speech (test set)	200 sentences (23 males and 23 females) from JNAS database
Response sound of a dialogue system	1 sentence (female) from ASJ database [9]

We measured 13 kinds of impulse responses as follows: 12 patterns are for the states where the interference is allocated, and the other pattern is for the state where the interference does not exist. We used the impulse responses without the mannequin as the ones before fluctuation, and we evaluated the average performance in 12 kinds of fluctuation. Figure 5 shows the arrangement of the apparatuses. As shown in Fig. 5, we place a dummy head, which has an average human head and upper body, at the user's position.

The impulse responses used in this experiment are measured in an acoustic experimental room, where the reverberation time is approximately 160 ms, with 48 kHz sampling and 16 bit resolution. The loudspeakers used in the sound field control of MOMNI and proposed method are positioned on the outer circumference of the room. The primary sound source of MOMNI method is the loudspeaker used as the spoken dialogue system in the acoustic echo canceller.

The filters for sound field control, in which the number of loudspeakers is M ($M = 5$ or 8) and the number of control points on the microphone elements is K ($K = 1, 2, 3$ or 4) (hereafter we label the transfer system "M-K system"), are designed. The passband range is set to 150–4000 Hz. We use a circular microphone array with 12 elements, and we select the elements which are equally spaced. In this experiment, we assume that the echo canceller is adapted precisely without error before the fluctuation of the transfer functions, but after the fluctuation the adaptation cannot be performed because of the double talk.

4.2. Results

In order to valuate the speech recognition performance, we adopt the Word Accuracy (WA) as an evaluation score. Table 1 lists the experimental conditions for the speech recognition. We average each WA which is obtained from 200 speech in total. The speech signal, which is obtained by superposing the elimination error of response sound on the user's speech, is used for the speech recognition experiment. In the acoustic echo canceller, the power ratio of the response sound and the user's speech at the microphone is set to 0 dB. In MOMNI and proposed method, we arranged the power of the response sound observed at the user's ear to be equal to that of the acoustic echo canceller in 0 dB state. We use two PTM models. One is generated from clean speech, and the other is learned using the speech on which office noise of 25 dB is superposed (hereafter we call it "25 dB model"). In speech recognition with 25 dB model, the same noise of 25 dB is superposed on the recorded speech signal.

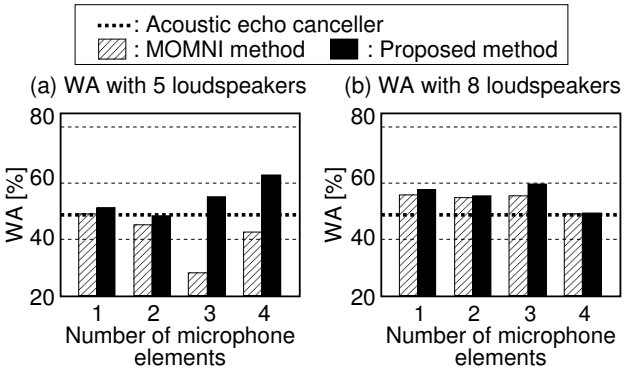

Fig. 6. Comparison of WA with clean model in the case of M loudspeakers: (a) $M = 5$ and (b) $M = 8$.

Figure 6 shows the speech recognition performances with clean model, and Fig. 7 is with that of noise superposed. In this experiment, since human speech is used as response sound, insertion errors are often caused even when elim-

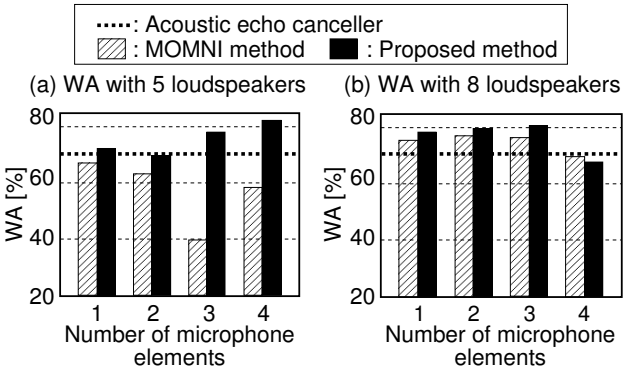

Fig. 7. Comparison of WA with 25 dB model in the case of M loudspeakers: (a) $M = 5$ and (b) $M = 8$.

ination performance is high. Known noise superposition can prevent this kind of error considerably, and all results of 25 dB model show higher performances than those with clean model. While MOMNI method provides lower performances in $M = 5$ than in $M = 8$ because of decrease of loudspeakers, the decrease doesn't influence the performance of the proposed method. The proposed method of 5-4 system marks the highest score in all of the results, e.g., that with 25 dB model shows a high performance of 82.4%. Using the proposed method together with known noise superposition, an improvement of 33.4% over the performance in the conventional acoustic echo canceller is achieved.

5. CONCLUSION

We proposed a small-size barge-in free interface using a response sound cancellation. As the results of the experiments, the robustness of sound elimination improved when the number of loudspeakers is relatively small. From these findings, the availability of the proposed method is ascertained.

6. REFERENCES

[1] B.H. Juang and F.K. Soong, "Hands-free telecommunications," *Proc. HSC,* pp.5–10, 2001.

[2] E. Hänsler, "Acoustic echo and noise control: where do we come from — where do we go?," *Proc. IWAENC,* pp.1–4, 2001.

[3] S. Makino and S. Shimauchi, "Stereophonic acoustic echo cancellation — an overview and recent solutions," *Proc. IWAENC,* pp.12–19, 1999.

[4] W. Herbordt, J. Ying, H. Buchner, and W. Kellermann, "A real-time acoustic human-machine front-end for multimedia applications integrating robust adaptive beamforming and stereophonic acoustic echo cancellation," *Proc. ICSLP,* vol.2, pp.773–776, 2002.

[5] H. Buchner, S. Spors, W. Kellermann, "Wave-domain adaptive filtering: acoustic echo cancellation for full-duplex system based on wave-field synthesis," *Proc. ICASSP,* vol.IV, pp.117–120, 2004.

[6] Y. Hinamoto, K. Mino, H. Saruwatari, and K. Shikano, "Interface for barge-in free spoken dialogue system based on sound field control and microphone array," *Proc. ICASSP,* vol.V, pp.505–508, 2003.

[7] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* vol.36, no.2, pp.145–152, 1988.

[8] Y. Tatekura, H. Saruwatari, and K. Shikano, "Sound reproduction system including adaptive compensation of temperature fluctuation effect for broad-band sound control," *IEICE Trans. Fundamentals*, vol.E85-A, no.8, pp.1851–1860, 2002.

[9] S. Hayamizu S. Itahashi, T. Kobayashi, and T. Takezawa, "Design and creation of speech and text cropora of dialogue," *IEICE Trans. Information and Systems,* vol.E76-D, no.1, pp.17–22, 1993.

[10] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese Speech Corpus for Large Vocabulary continuous speech recognition research," *The Journal of the Acoustical Society of Japan (E)*, vol.20, no.3, pp.199–206, 1999.

[11] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," *Proc. ICSLP,* vol.7, pp.3261–3264, 1998.

[12] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A new phonetic tied-mixture model for efficient decoding," *Proc. ICASSP,* vol.III, pp.1269–1272, 2000.

[13] A. Lee, T. Kawahara, and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine," *Proc. EUROSPEECH,* vol.3, pp.1691–1694, 2001.

[14] S. Yamade, A. Lee, H. Saruwatari, and K. Shikano, "Unsupervised speaker adaptation based on HMM sufficient statistics in various noisy environments," *Proc. EUROSPEECH,* pp.II-1493–1496, 2003.