LETTER

Interface for Barge-in Free Spoken Dialogue System Combining Adaptive Sound Field Control and Microphone Array

Tatsunori ASAI[†], *Nonmember*, **Hiroshi SARUWATARI**[†a)], *and* **Kiyohiro SHIKANO**[†], *Members*

SUMMARY This paper describes a new interface for a barge-in free spoken dialogue system combining an adaptive sound field control and a microphone array. In order to actualize robustness against the change of transfer functions due to the various interferences, the barge-in free spoken dialogue system which uses sound field control and a microphone array has been proposed by one of the authors. However, this method cannot follow the change of transfer functions because the method consists of fixed filters. To solve the problem, we introduce a new adaptive sound field control that follows the change of transfer functions.

key words: *spoken dialogue system, barge-in, adaptive sound field control, microphone array, speech recognition*

1. Introduction

In human-machine communication based on a spoken dialogue system, it is vital that user's speech reaches the dialogue system to enable smooth communication. However, the user usually makes an utterance before the dialogue system finishes responding. Such a situation, in which a user and a system utter simultaneously is referred to as *barge-in* [1]. In the state of barge-in, the recognition performance of the user's speech is degraded because the response sound of the dialogue system is inputted into the microphone for recording the user's speech.

In order to eliminate the response sound, an acoustic echo canceller is commonly used. Many types of acoustic echo cancellers have been proposed, e.g., single-channel, stereophonic, and integrated with a beamformer [2]–[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 the Multiple-Output and Multiple-No-Input (MOMNI) method [5] which combines sound field control and microphone array techniques. Although the MOMNI method is robust against the change of transfer functions, there still exists the drawback that the MOMNI method cannot adaptively follow the change of transfer functions because the method consists of fixed filters.

To improve the MOMNI method, in this paper, we introduce a new adaptive algorithm of sound field control, in which the changes in the room conditions can be adaptively detected and reflected in the construction of the inverse filters used for the sound field control. The feasibility of the proposed algorithm is demonstrated in an experiment performed in a real room.

Manuscript received April 6, 2004.
Manuscript revised September 27, 2004.
Final manuscript received February 16, 2005.

[†]The authors are with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630-0192 Japan.
a) E-mail: sawatari@is.naist.jp

DOI: 10.1093/ietfec/e88–a.6.1613

2. Conventional MOMNI Method [5]

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

In Fig. 1, S_m $(m = 1, \cdots, M)$ is the loudspeaker which acts as a secondary sound source, and C_n $(n = 1, \cdots, N)$ is the microphone which acts as a control point. C_1 and C_2 are located in the vicinity of the two external auditory meatus of a user, and $C_3, \cdots, C_{K+2}$ $(K = N - 2)$ are placed in each microphone element for recording the user's speech. The intended signals to be reproduced at each control point are represented by

$$\boldsymbol{X}(\omega) = [X_{\mathrm{R}}(\omega), X_{\mathrm{L}}(\omega), X_{\mathrm{mic}1}(\omega), \cdots, X_{\mathrm{mic}K}(\omega)]^{\mathrm{T}}, \quad (1)$$

where $X_{\mathrm{L}}(\omega)$, $X_{\mathrm{R}}(\omega)$ and $X_{\mathrm{mic}k}(\omega)$ $(k = 1, \cdots, K)$ are the signals to be reproduced at the left and right ears of a user, and at microphone C_{k+2}, respectively. Similarly, the observation signals at each of the control points are described as

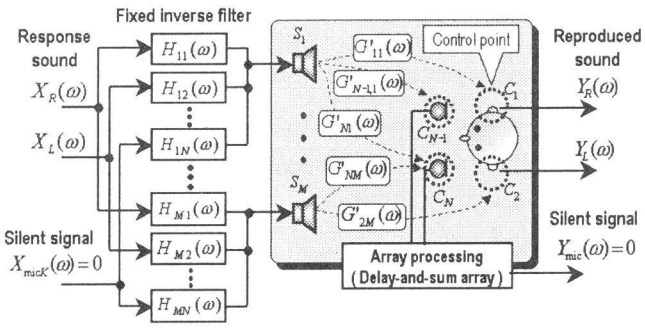

Fig. 1 Configuration of conventional MOMNI method.

$$\boldsymbol{Y}(\omega) = [Y_{\mathrm{R}}(\omega), Y_{\mathrm{L}}(\omega), Y_{\mathrm{mic}1}(\omega), \cdots, Y_{\mathrm{mic}K}(\omega)]^{\mathrm{T}}. \quad (2)$$

If the $N\times M$ matrix composed of the room transfer function $G_{nm}(\omega)$ $(N < M)$ between the secondary sound source S_m and the control point C_n is denoted by $\boldsymbol{G}(\omega)$, and the $M\times N$ inverse filter matrix [6] is expressed as $\boldsymbol{H}(\omega)$, $\boldsymbol{Y}(\omega)$ is denoted by

$$\boldsymbol{Y}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{H}(\omega)\boldsymbol{X}(\omega), \quad (3)$$

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

In Eq. (2), the response sounds of a dialogue system are reproduced at both ears of the user ($[Y_{\mathrm{L}}(\omega), Y_{\mathrm{R}}(\omega)] = [X_{\mathrm{L}}(\omega), X_{\mathrm{R}}(\omega)]$) and silent zones are materialized at each microphone element ($[Y_{\mathrm{mic}1}(\omega), \cdots, Y_{\mathrm{mic}K}(\omega)] = [0, \cdots, 0]$). 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.

2.2 Microphone Array Based on Delay-and-Sum Array

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

$$A_k(\omega) = (1/K)\cdot e^{-j\omega\tau_k}, \quad (4)$$

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

$$Y_{\mathrm{mic}}(\omega) = \sum_{k=1}^{K} A_k(\omega) Y_{\mathrm{mic}k}(\omega). \quad (5)$$

2.3 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 [7] 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 [5].

2.4 Response Sound Elimination Error When Changing Room Transfer Functions

The MOMNI method which uses fixed inverse filter coefficients is proved to be robust against the change of room transfer functions [5]. Assume that the fluctuation $\Delta G_{nm}(\omega)$ caused by the change of transfer functions is added to the transfer function $G_{nm}(\omega)$. Since the observation signal $\boldsymbol{Y}'(\omega)$ denotes

$$\boldsymbol{Y}'(\omega) = (\boldsymbol{G}(\omega) + \Delta\boldsymbol{G}(\omega))\boldsymbol{H}(\omega)\boldsymbol{X}(\omega), \quad (6)$$

the elimination error of response sound at the array output is represented as

$$\Delta Y_{\mathrm{mic}}(\omega) = \sum_{k=1}^{K} A_k(\omega)\left\{\sum_{m=1}^{M}\Delta G_{(k+2)m}(\omega) \cdot\left(H_{m1}(\omega)X_{\mathrm{R}}(\omega) + H_{m2}(\omega)X_{\mathrm{L}}(\omega)\right)\right\}. \quad (7)$$

Denoting the matrix norm of $\boldsymbol{H}(\omega)$ by $\|\boldsymbol{H}(\omega)\|$, Eq. (7) can be rewritten as

$$\Delta Y_{\mathrm{mic}}(\omega) = \|\boldsymbol{H}(\omega)\| \cdot \frac{1}{K} \cdot \left\{\sum_{k=1}^{K}\sum_{m=1}^{M}\Delta G_{(k+2)m}(\omega) \cdot\left(\mathcal{H}_{m1}(\omega)X_{\mathrm{R}}(\omega) + \mathcal{H}_{m2}(\omega)X_{\mathrm{L}}(\omega)\right)e^{-j\omega\tau_k}\right\}, \quad (8)$$

where $\mathcal{H}_{mn}(\omega) = H_{mn}(\omega)/\|\boldsymbol{H}(\omega)\|$. It is assumed that $\Delta G_{nm}(\omega)$ is the Gaussian random variable with the variance σ^2. Furthermore, since $\mathcal{H}_{mn}(\omega)$ is normalized by $\|\boldsymbol{H}(\omega)\|$, and is independent from the change of M, the variance in $\{\cdot\}$ of Eq. (8) can be expressed as $\eta\sqrt{M\cdot K}\sigma$, where η is an appropriate constant. Additionally, $\|\boldsymbol{H}(\omega)\|$ is proportional to $1/M$ because $\|\boldsymbol{H}(\omega)\| \simeq 1/\|\boldsymbol{G}(\omega)\| \propto 1/M$. Therefore, in the report in [5], the following relation holds in the elimination error of response sound, $\mathcal{E}(\omega)$, as

$$\mathcal{E}(\omega) = \Delta Y_{\mathrm{mic}}(\omega) \propto (1/M)\cdot(1/K)\cdot\sqrt{M\cdot K} = 1/\sqrt{M\cdot K}. \quad (9)$$

Equation (9) shows that the elimination error of response sound is inversely proportional to $\sqrt{M\cdot K}$. 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.

3. Proposed Technique for Response Sound Elimination

Although the MOMNI method is robust against the change of transfer functions, we cannot estimate the changed transfer functions. Therefore, we propose the Adaptive Sound Field Control (ASFC) method. The ASFC method is a new interface for a barge-in free spoken dialogue system which follows the changed transfer functions. Figure 2 depicts the configuration of the proposed ASFC method.

3.1 Adaptive Algorithm for Transfer Function Estimation

The procedure for estimating the transfer functions using observed signals is as follows.

[*step* 0] The initial value $\hat{\boldsymbol{G}}^{[0]}(\omega)$ of the estimated transfer

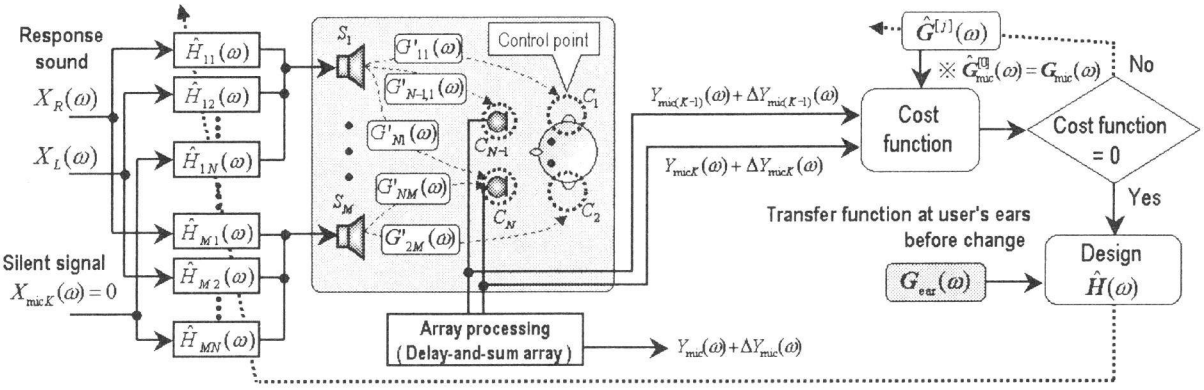

Fig. 2 Configuration of proposed ASFC method.

function is set to $\boldsymbol{G}(\omega)$.

[*step* 1] In the case where the fluctuation of transfer function $\Delta G_{nm}(\omega)$ is added into the transfer function $G_{nm}(\omega)$ because of the change of a transfer system, the changed transfer function $\boldsymbol{G}'(\omega)$ becomes

$$\boldsymbol{G}'(\omega) = \boldsymbol{G}(\omega) + \Delta\boldsymbol{G}(\omega), \tag{10}$$

and the observation signal $\boldsymbol{Y}'(\omega)$ at the control points is expressed as

$$\boldsymbol{Y}'(\omega) = \boldsymbol{G}'(\omega)\boldsymbol{H}(\omega)\boldsymbol{X}(\omega). \tag{11}$$

Similarly, the estimated signal $\hat{\boldsymbol{Y}}^{[i-1]}(\omega)$ is depicted in

$$\hat{\boldsymbol{Y}}^{[i-1]}(\omega) = \hat{\boldsymbol{G}}^{[i-1]}(\omega)\boldsymbol{H}(\omega)\boldsymbol{X}(\omega), \tag{12}$$

where i is the number of iterations, and $\hat{\boldsymbol{G}}^{[i-1]}(\omega)$ is the estimated transfer function of $\boldsymbol{G}'(\omega)$. In the estimation process, we derive $\hat{\boldsymbol{G}}^{[i-1]}(\omega)$ that minimizes the squared error between $\boldsymbol{Y}'(\omega)$ and $\hat{\boldsymbol{Y}}^{[i-1]}(\omega)$. When we define the error signal vector $\boldsymbol{E}(\omega)$ as

$$\boldsymbol{E}(\omega) = [E_{\mathrm{R}}(\omega), E_{\mathrm{L}}(\omega), E_{\mathrm{mic1}}(\omega), \cdots, E_{\mathrm{mic}K}(\omega)]^{\mathrm{T}}, \tag{13}$$

where $E_{\mathrm{L}}(\omega)$, $E_{\mathrm{R}}(\omega)$ and $E_{\mathrm{mic}k}(\omega)$ are the error signals at each of the control points, $\boldsymbol{E}^{[i-1]}(\omega)$ can be given by

$$\begin{aligned}\boldsymbol{E}^{[i-1]}(\omega) &= \boldsymbol{Y}'(\omega) - \hat{\boldsymbol{Y}}^{[i-1]}(\omega) \\ &= (\boldsymbol{G}'(\omega) - \hat{\boldsymbol{G}}^{[i-1]}(\omega))\boldsymbol{H}(\omega)\boldsymbol{X}(\omega).\end{aligned} \tag{14}$$

However, we cannot actually calculate the error in the neighborhood of both ears of a user because the microphones for observing the changed transfer functions are not placed at C_1 and C_2. Hence, the error signals, $E_L(\omega)$ and $E_R(\omega)$, are set to be zero ($[E_L(\omega), E_R(\omega)] = [0, 0]$). We regard $\|\boldsymbol{E}^{[i-1]}(\omega)\|^2$ as the cost function to be minimized in this algorithm. From Eq. (14), the partial differentiation of the squared error $\|\boldsymbol{E}^{[i-1]}(\omega)\|^2$ with respect to $\hat{\boldsymbol{G}}^{[i-1]}(\omega)$ is given by

$$\frac{\partial \|\boldsymbol{E}^{[i-1]}(\omega)\|^2}{\partial \hat{\boldsymbol{G}}^{*[i-1]}(\omega)} = -\boldsymbol{E}^{[i-1]}(\omega)(\boldsymbol{H}(\omega)\boldsymbol{X}(\omega))^{\mathrm{H}}. \tag{15}$$

Thus, the modification amount of $\hat{\boldsymbol{G}}^{[i-1]}(\omega)$ in a normalized least-mean-squares (NLMS) method is denoted by

$$\begin{aligned}\Delta\hat{\boldsymbol{G}}^{[i-1]}(\omega) = &\frac{\alpha}{\|\boldsymbol{H}(\omega)\boldsymbol{X}(\omega)\|^2 + \beta} \\ &\cdot\boldsymbol{E}^{[i-1]}(\omega)(\boldsymbol{H}(\omega)\boldsymbol{X}(\omega))^{\mathrm{H}},\end{aligned} \tag{16}$$

where α ($0 < \alpha < 2$) is a step-size parameter, and β is a minimal positive constant that is nonzero in the denominator term on the right-hand side of Eq. (16).

The i-th estimated transfer function $\hat{\boldsymbol{G}}^{[i]}(\omega)$ can be updated, as shown below:

$$\hat{\boldsymbol{G}}^{[i]}(\omega) = \hat{\boldsymbol{G}}^{[i-1]}(\omega) + \Delta\hat{\boldsymbol{G}}^{[i-1]}(\omega). \tag{17}$$

[*step* 2] If $\hat{\boldsymbol{G}}^{[i]}(\omega)$ derived from Eq. (17) is converged, we return to *step* 1 and update the estimated transfer function in the next frame repeatedly.

[*step* 3] We design the new inverse filter $\hat{\boldsymbol{H}}(\omega)$ based on $\hat{\boldsymbol{G}}^{[i]}(\omega)$ via LNS.

4. Experiments and Results

In this section, we present two experiments comparing the conventional method (acoustic echo canceller and MOMNI method) and the proposed method (ASFC method). In order to verify the applicability of the proposed method, we simulate an adaptation process based on the change of transfer functions and evaluate the performance of each method, on the basis of the response sound elimination experiment and the speech recognition experiment.

4.1 Experimental Conditions

In this experiment, 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 an other person who is not a user approaches the user, which is a very common occurrence in real environments.

We measured thirteen kinds of impulse responses as follows: twelve patterns are for the states where the interference is allocated, and the other pattern is for the state where

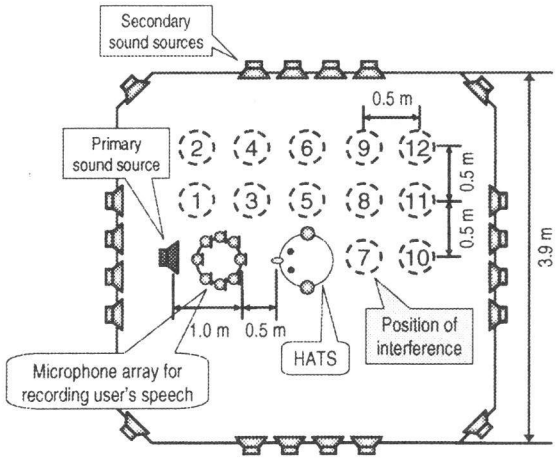

Fig. 3 Layout of acoustic experiment room.

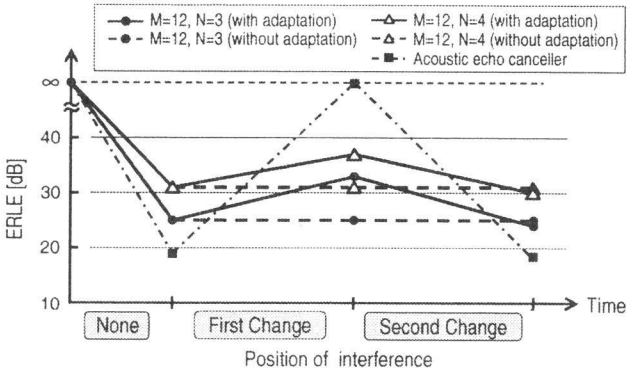

Fig. 4 ERLE in the 12-3 and 12-4 systems.

the interference does not exist. Figure 3 shows the arrangement of the apparatuses. As shown in Fig. 3, we place the HATS (Head And Torso Simulator), 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 experiment room, where the reverberation time is approximately 200 ms, with 48 kHz sampling and 16 bit resolution. The primary sound source is the loudspeaker used as the spoken dialogue system in the acoustic echo canceller.

The inverse filters of the transfer system, in which the number of secondary sound sources is M (M = 12 or 16) and the number of control points is N (N = 3, 4, 6, or 8) (hereafter, we label the transfer system, the M-N system), are designed. Also, the passband range is 150–4000 Hz. We use a circular microphone array with eight elements, and we select the elements which are equally spaced. It is worth noting that the distance between the primary source and the microphone array is shorter than those of the secondly sources and the microphone array, and consequently the time causality does not hold in this sound reproduction. Indeed, the inverse filters used in this experiment contains an appropriate time delay, and this causes a slight latency in the reproduced sound. However, such kind of latency is not so harmful and can be acceptable, especially in the spoken dialogue interface.

First, the interference shifts to one of twelve positions (hereafter we designate this change as "First Change"), and we estimate the changed transfer functions and design inverse filters. In the estimation, we use every one-second sound cut from the response sound, which is an adequate time length for the adaptation. The step-size parameter α in Eq. (16) is 0.1, which is optimized experimentally, β is 1.0×10^{-6}, and the number of iterations, i, is 10.

Finally, it is assumed that the interference moves from one position to the other positions in the state of barge-in (hereafter we designate this change as "Second Change"), and therefore we apply the MOMNI method and stop estimating.

The filter coefficient of the acoustic echo canceller is constructed without a specific adaptive algorithm. In this experiment, in the case of no existence and after the first movement of the interference, we assume that the echo canceller can estimate the filter coefficient precisely under the ideal condition without error.

4.2 Response Sound Elimination Experiment

4.2.1 Evaluation of Response Sound Elimination

To evaluate the response sound elimination performance, we calculate the echo return loss enhancement (ERLE); this is given by

$$\mathrm{ERLE} = 10\log_{10}\frac{\sum_{\omega}\{Y_{\mathrm{micref}}(\omega)\}^2}{\sum_{\omega}\{\mathcal{E}(\omega)\}^2}, \tag{18}$$

where $Y_{\mathrm{micref}}(\omega)$ is the response sound reproduced at a criterial microphone assigned by us, and $\mathcal{E}(\omega)$ is the error signal derived from Eq. (9). In "First Change" situation, we average each ERLE which is obtained from the twelve interference patterns. In "Second Change" scenario, we should consider $12 \times 11 (= 132)$ patterns which correspond to the total possible movements from "First Change" (12 patterns) to "Second Change" ($12-1$ patterns). Thus, we average 132 ERLEs as the resultant ERLE score in "Second Change."

As the response sound from the dialogue system, we use a female utterance selected from the ASJ database [8]. Although the sampling frequency of the response sound is 16 kHz, we use the signal from which the frequency component above 4 kHz is eliminated.

4.2.2 Experimental Result and Consideration

Figures 4–7 show the ERLEs with the adaptation process and without adaptation in each M-N system. In each figure, the vertical axis is expressed as the interference position with time passing, and the horizontal axis is designated as ERLE.

As compared with the results of the conventional acoustic echo canceller, the proposed ASFC method can achieve the improvement of the ERLE by more than 6 dB

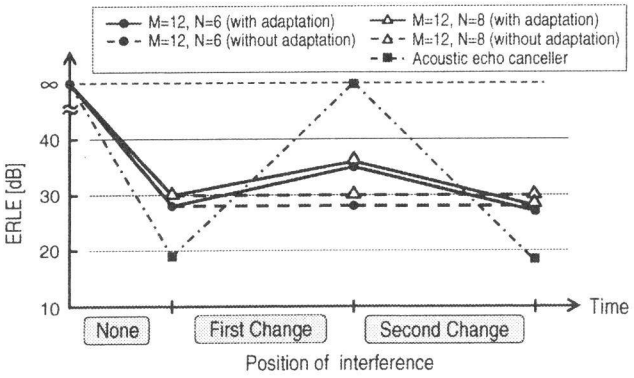

Fig. 5 ERLE in the 12-6 and 12-8 systems.

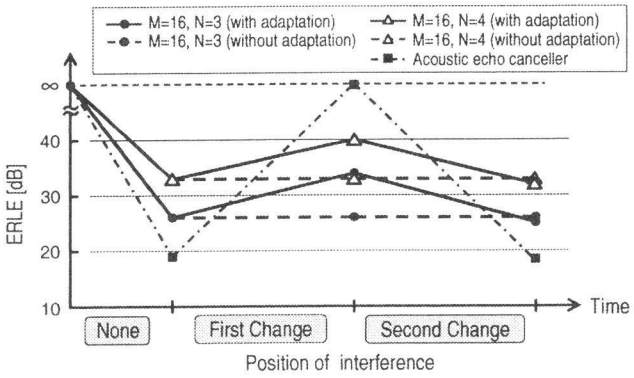

Fig. 6 ERLE in the 16-3 and 16-4 systems.

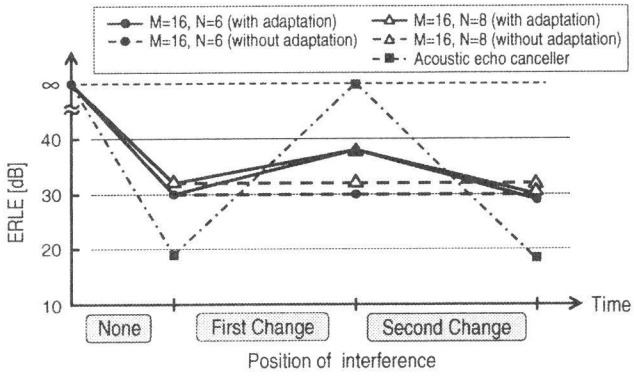

Fig. 7 ERLE in the 16-6 and 16-8 systems.

in all *M*-*N* systems when we cannot estimate the changed transfer functions, such as in the barge-in situation. The response sound elimination performance is also improved in the MOMNI method if the number of transfer channels ($= M \cdot K$) increases. The theoretical performance of ERLE is expressed as [5]

$$\begin{aligned} \mathrm{ERLE}_{\mathrm{theory}} &\propto \xi + 10\log_{10}(1/\{1/(M \cdot K)\}) \\ &= \xi + 10\log_{10}(M \cdot K), \end{aligned} \tag{19}$$

where ξ is a suitable constant. We can see the good agreement of the results with the theory in Eq. (19). In addition, as compared with the result of the MOMNI method, it is shown that the proposed ASFC method can improve the ERLE by more than 6 dB when the changed transfer functions are estimated. If we cannot estimate the changed transfer function, the response sound elimination performance is approximately equivalent to that of the MOMNI method. In summary, the performance of the proposed ASFC method is superior to or equal to that of the MOMNI method throughout the changes of the interference.

Table 1 Experimental conditions for speech recognition.

Speech database	JNAS [9]
Frame length	25 msec (Hamming window)
Frame interval	8 msec
Feature vector	12 MFCCs, 12 ΔMFCCs, Δpower
Language model	Newspaper dictation [10]
Phoneme model	Phonetic Tied Mixture (PTM) [11]
Decoder	Julius ver. 3.1 [12]
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 [8]

4.3 Speech Recognition Experiment

4.3.1 Evaluation of Speech Recognition Performance

The effect of the elimination of response sound is evaluated with a large vocabulary continuous speech recognition task. In order to evaluate the speech recognition performance, we adopt the Word Accuracy (WA) as an evaluation score. WA is defined as follows:

$$\mathrm{WA}[\%] = \frac{N - S - D - I}{N} \times 100, \tag{20}$$

where N is the total number of words in the test speech, S is the number of substitution errors, D is the number of deletion errors, and I is the number of insertion errors. Table 1 lists the experimental conditions for speech recognition. We average each WA which is obtained from 200 people in total.

The speech signal which is obtained by superimposing the elimination error of response sound, $\mathcal{E}(\omega)$, on the user's speech is used for the speech recognition experiment, where the power ratio of the response sound and the user's speech at the microphone is set to 0 dB. We use the PTM (Phonetic Tied Mixture model based on triphones) which is independent from speakers and is generated from clean speech.

4.3.2 Experimental Result and Consideration

Figures 8 and 9 show the WA with the adaptation process and without adaptation in each *M*-*N* system. In each figure, the vertical axis is expressed as the interference position with time passing, and the horizontal axis is designated as WA.

For example, as compared with the results of the conventional acoustic echo canceller, by applying the proposed ASFC method, we can confirm that the improvement in WA

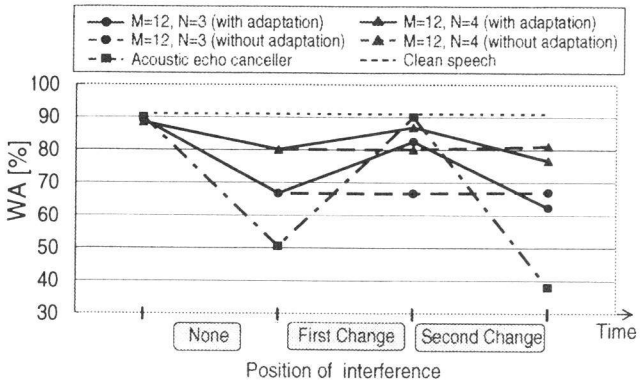

Fig. 8 Word Accuracy in the 12-3 and 12-4 systems.

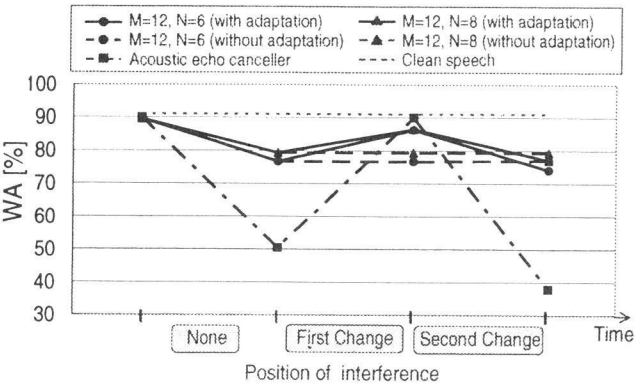

Fig. 9 Word Accuracy in the 12-6 and 12-8 systems.

of 24.4% (in the 12-3 system) is obtained when we cannot estimate the changed transfer functions, such as in the state of barge-in. In addition, as compared with the result of the MOMNI method, it is shown that the proposed ASFC method can improve the WA by 15.8% (in the 12-3 system) when the changed transfer functions are estimated. If we cannot estimate the changed transfer function, the performance is approximately equivalent to that of the MOMNI method.

5. Conclusion

We proposed an ASFC method which is an interface for the barge-in free spoken dialogue system based on adaptive sound field control and a delay-and-sum microphone array. As the result of two comparative experiments on response sound elimination and speech recognition, the performance of the ASFC method was enhanced further than that of the MOMNI method when the transfer functions after change could be estimated. Also, in the barge-in situation, the performance of the ASFC method was prominently improved in comparison with that of an acoustic echo canceller. From these results, the applicability of the proposed ASFC method is ascertained.

References

[1] B.H. Juang and F.K. Soong, "Hands-free telecommunications," Proc. International Workshop on Hands-Free Speech Communication 2001, pp.5–10, Kyoto, Japan, April 2001.

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

[3] S. Makino and S. Shimauchi, "Stereophonic acoustic echo cancellation—An overview and recent solutions," Proc. IWAENC 1999, pp.12–19, Pennsylvania, USA, Sept. 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 2002, vol.2, pp.773–776, Colorado, USA, Sept. 2002.

[5] 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 2003, vol.V, pp.505–508, Hong Kong, China, April 2003.

[6] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," IEEE Trans. Acoust. Speech Signal Process., vol.36, no.2, pp.145–152, Feb. 1988.

[7] Y. Tatekura, H. Saruwatari, and K. Shikano, "An iterative inverse filter design method for the multichannel sound field reproduction system," IEICE Trans. Fundamentals, vol.E84-A, no.4, pp.991–998, April 2001.

[8] S. Hayamizu, S. Itahashi, T. Kobayashi, and T. Takezawa, "Design and creation of speech and text corpora of dialogue," IEICE Trans. Inf. & Syst., vol.E76-D, no.1, pp.17–22, Jan. 1993.

[9] 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," J. Acoust. Soc. Jpn (E), vol.20, no.3, pp.199–206, May 1999.

[10] 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 1998, vol.7, pp.3261–3264, Sydney, Australia, Dec. 1998.

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

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