Interface for Barge-in Free Spoken Dialogue System Using Adaptive Sound Field Control

Tatsunori Asai †, *Shigeki Miyabe* †, *Hiroshi Saruwatari* †, *Kiyohiro Shikano* †

† Nara Institute of Science and Technology, 630-0192, Nara, JAPAN (e-mail: sawatari@is.naist.jp)

Abstract

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 large change of transfer functions. To solve the problem, we introduce a new adaptive sound field control that follows the change of transfer functions. The experimental results reveal that the proposed method can improve the reduction accuracy of response sound in comparison with the conventional acoustic echo canceller as well as the previously proposed method which simply uses fixed sound field control system.

1. Introduction

In man-machine communication based on a spoken dialogue system, it is vital that user's speech reaches the dialogue system to communicate smoothly. However, the user usually utters before the dialogue system finishes responding. Such the situation in which a user and a system simultaneously utter 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 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 beamforming [2], [3], [4]. However, the acoustic echo canceller is inherently vulnerable to the change of transfer functions in the barge-in situation. To solve the problem of the acoustic echo canceller, one of the authors has proposed Multiple-Output and Multiple-No-Input (MOMNI) method [5] which combines sound field control and microphone array. Although MOMNI method is robust against the small change of transfer functions, there still exists the drawback that MOMNI method cannot adaptively follow the large change of transfer functions because the method consists of the fixed filters.

To improve the MOMNI method, in this paper, we introduce a new adaptive algorithm of sound field control, in which the large change of the room conditions can be adaptively detected and reflected in constructing the inverse filters used for the sound field control. The feasibility of the proposed algorithm can be shown by the experiment in the real room.

2. Conventional MOMNI method [5]

We describe MOMNI method shown in Fig.1. The MOMNI method consists of two main parts, i.e., 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 both external auditory meatus of a user, and $C_3, \cdots, C_{K+2}$ ($K = N - 2$) are placed in each microphone element for recording user's speech. The intended signals to be reproduced at each control point are represented by

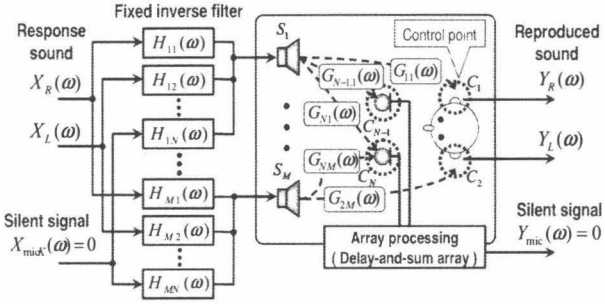

Figure 1: Configuration of conventional MOMNI method.

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

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 control points are described as

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

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 inverse filter matrix [6] is expressed as $\boldsymbol{H}(\omega)$, then $\boldsymbol{Y}(\omega)$ is given 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 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 and prevents the response sound from mixing into the observation signal at each microphone element.

2.2. Microphone array based on delay-and-sum array

In multi-channel speech enhancement, the delay-and-sum array is commonly used. To obtain the user's speech at array output, we compensate the delay for 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

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

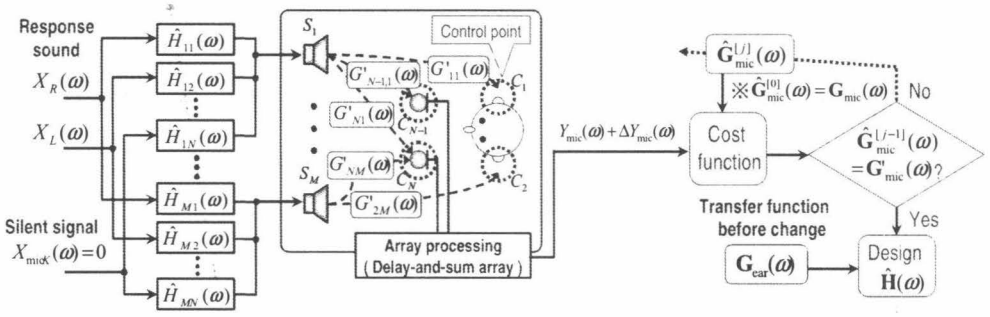

Figure 2: Configuration of proposed ASFC method.

where τ_k is the arriving 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

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

2.3. Inverse filter design for sound field control

In the multipoint control system based on loudspeakers, we need to consider the influence of the room transfer functions. For that 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 inverse matrix as the inverse filter which gives the LNS [5].

2.4. Response sound elimination error

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 observation signal $\boldsymbol{Y}'(\omega)$ denotes

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

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

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

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

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

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 a 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

$$\begin{aligned} \mathcal{E}(\omega) = \Delta Y_{\mathrm{mic}}(\omega) &\propto (1/M) \cdot (1/K) \cdot \sqrt{M \cdot K} \\ &= 1/\sqrt{M \cdot K}. \end{aligned} \tag{9}$$

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 increases, the MOMNI method becomes more robust against the change of transfer functions than an acoustic echo canceller.

3. Proposed adaptive method

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

3.1. Adaptive algorithm for transfer function estimation

The procedure to estimate the transfer functions using observed signals is as follows.

[*step* 0] The initial value $\hat{\boldsymbol{G}}^{[0]}(\omega)$ of estimated transfer function is set to $\boldsymbol{G}(\omega)$.

[*step* 1] In the case where the fluctuation of transfer function $\Delta G_{nm}(\omega)$ is added in 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

$$\boldsymbol{E}(\omega) = [E_{\mathrm{R}}(\omega), E_{\mathrm{L}}(\omega), E_{\mathrm{mic}1}(\omega), \cdots, E_{\mathrm{mic}K}(\omega)]^{\mathrm{T}}, \tag{13}$$

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

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

However, we cannot actually calculate the error in the neighborhood of both ears of a user because the microphones for observing 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 ($[E_L(\omega), E_R(\omega)] = [0, 0]$). 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

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

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

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

where α ($0 < \alpha < 2$) is a step-size parameter, and β is a minimal positive constant to be non-zero 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:

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

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

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

4. Experiment and result

4.1. Experimental condition

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 the other person except a user approaches the user. We measured the impulse responses thirteen times: twelve patterns are the states where the interference is allocated, and the other pattern is the state where the interference does not exist. Figure 3 shows the arrangement of the apparatuses.

The impulse responses used in this experiment are measured in an acoustic experiment room, where the reverberation time is about 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. We use a circular microphone array with six elements which are equally spaced. The inverse filters of transfer system, in which the number of secondary sound sources is M (M=12) and the number of control points is N (N=3, 4, 6, and 8) (hereafter, we label the transfer system as M-N system), are designed. Also, the passband range is 150--4000 Hz. As the response sound from the dialogue system, we use a female sound selected from the ASJ database.

First, the interference shifts to one of twelve positions, and we estimate the changed transfer functions and design inverse filters. In the estimation, we use every one-second sound cut

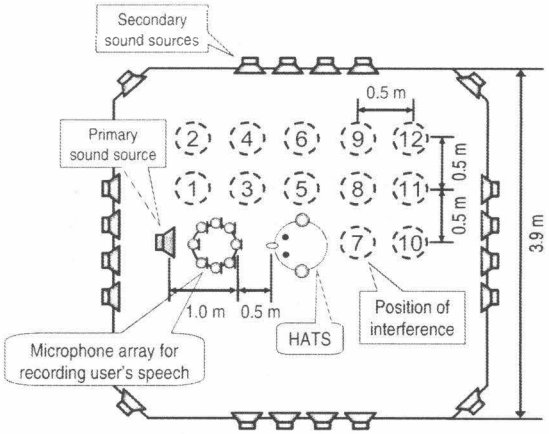

Figure 3: Layout of acoustic experiment room.

from the response sound which has adequate time length. The step-size parameter α in Eq. (16) is 0.1, β is 1.0×10^{-6}, and the number of iterations is 10.

Finally, it is assumed that the interference moves from one position to the other positions in the state of barge-in, we apply the MOMNI method and stop estimating.

The filter coefficient of acoustic echo canceller is constructed without a specific adaptive algorithm. In the case of no existence and after 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. Evaluation of response sound elimination

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

$$\mathrm{ERLE} = 10\log_{10}\sum_{\omega}\{Y_{\mathrm{micref}}(\omega)\}^2 / \sum_{\omega}\{\mathcal{E}(\omega)\}^2 \ [\mathrm{dB}], \quad (18)$$

where $Y_{\mathrm{micref}}(\omega)$ is the response sound reproduced at criterial microphone we assign, and $\mathcal{E}(\omega)$ is the error signal derived from Eq. (9). We average each ERLE which is obtained from twelve interference patterns.

Figure 4 shows the ERLEs which are with adaptation process and without adaptation in each M-N system. As compared with the results of the conventional acoustic echo canceller, by applying the proposed ASFC method, we can confirm that the improvement in ERLE of more than 5 dB is obtained when we cannot estimate the changed transfer functions, such as in the state of barge-in. The performance of response sound elimination is improved as well as MOMNI method if the number of transfer channels ($= M \cdot K$) increases.

In addition, as compared with the result of MOMNI method, it is shown that the proposed ASFC method can improve the ERLE by about 7 dB when the changed transfer functions are estimated. If we cannot estimate the changed transfer function, the performance is approximately equivalent to the MOMNI method. As shown in these results, the proposed ASFC method is more applicable to the spoken dialogue system than two conventional methods.

4.3. Evaluation of speech recognition performance

The effect of the elimination of response sound is evaluated with a large vocabulary continuous speech recognition task. To eval-

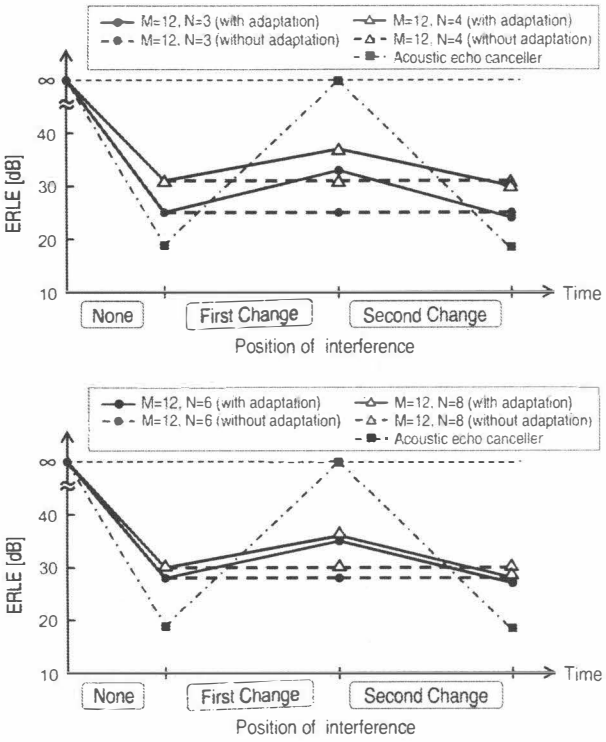

Figure 4: Result of ERLE in 12-3, 12-4, 12-6, and 12-8 systems.

uate the speech recognition performance, we adopt the Word Accuracy (WA) as an evaluation score; which is de□ned as

$$\mathrm{WA} = (N - S - D - I)/N \times 100\ [\%], \qquad (19)$$

where N is the total number of words in the test set, 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 use Phonetic Tied Mixture (PTM) model based on triphones [8] which is independent from speakers and is generated from clean speech.

Figure 5 shows the WA with the adaptation process and without adaptation in each M-N system. In the □gure, the vertical axis is expressed as the interference position with time passing, and the horizontal axis is designated as WA. As compared with the results of the conventional acoustic echo canceller, by applying the proposed ASFC method, we can con□rm that the improvement in WA 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 almost equivalent to that of the MOMNI method.

5. Conclusion

We proposed ASFC method which is a interface for barge-in free spoken dialogue system based on adaptive sound □eld control and a delay-and-sum microphone array. As the result of comparative experiment, the transfer functions after the change could be estimated, and the performance of response sound elimination prominently improved in comparison with an acoustic echo canceller. From these results, the availability of the proposed ASFC method is ascertained.

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
Decoder	Julius ver. 3.1 [10]
User's speech (test set)	200 sentences (23 males and 23 females) from JNAS database

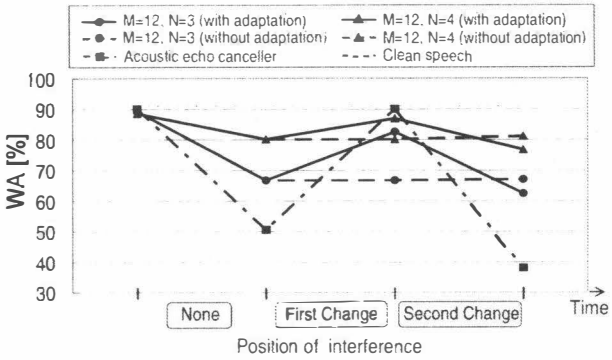

Figure 5: Result of WA in 12-3 and 12-4 systems.

6. Acknowledgement

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

7. References

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

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

[3] S. Makino and S. Shimauchi, "Stereophonic acoustic echo cancellation — an overview and recent solutions," *Proc. IWAENC 1999*, 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 2002*, vol.2, pp.773–776, 2002.

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

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

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

[8] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A new phonetic tied-mixture model for ef□cient decoding," *Proc. ICASSP2000*, vol.III, pp.1269–1272, 2000.

[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," *The Journal of the Acoustical Society of Japan (E)*, vol.20, no.3, pp.199–206, 1999.

[10] A. Lee, T. Kawahara, and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine," *Proc. EUROSPEECH2001*, vol.3, pp.1691–1694, 2001.