BARGE-IN FREE SPOKEN DIALOGUE INTERFACE USING NULLSPACE-BASED SOUND FIELD CONTROL AND BEAMFORMING

Shigeki MIYABE [1], *Hiroshi SARUWATARI* [1], *and Kiyohiro SHIKANO* [1], *Yosuke TATEKURA*[2]

[1]Nara Institute of Science and Technology,
8916-5, Takayama-Cho, 630-0192, Ikoma-City, Japan
phone: +81-0743-72-5287, fax: +82-0743-72-5289, email: shige-m@is.naist.jp
web: http://isw3.aist-nara.ac.jp/IS/Shikano-lab/e-home.html
[2] Shizuoka University, 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. Proposed method shows higher performances in the speech recognition experiments.

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 the system and the user speaks 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 issue 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 [2] [3] [4] [5]. 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-No-Input (MOMNI) method [6] which combines sound field control and microphone array techniques. By increasing the number of loudspeakers and microphone elements, MOMNI method can make its control robust against the change of transfer functions, but huge number of loudspeakers are needed to earn enough robustness for speech recognition. Furthermore, MOMNI method controls sound field around user's ears and premises that the user doesn't move from the assumed position.

In order 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 the speech recognition experiment.

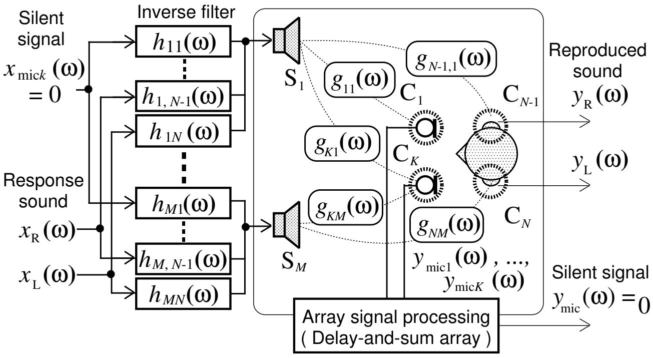

Figure 1: Configuration of conventional MOMNI method.

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 loudspeaker which acts as a secondary sound sources, and C_n ($n = 1$, ... , N) is the microphones which acts as control points. C_1, ..., 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 $\mathbf{X}(\omega) = [X_{\mathrm{mic}1}(\omega), \ldots, X_{\mathrm{mic}K}(\omega), X_{\mathrm{R}}(\omega), X_{\mathrm{L}}(\omega)]^{\mathrm{T}}$,

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

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 microphone C_k, the right and left ears of a user, respectively. Similarly, the observation signals at the control points are described as $\mathbf{Y}(\omega) = [Y_{\mathrm{mic}1}(\omega), \ldots, Y_{\mathrm{mic}K}(\omega), Y_{\mathrm{R}}(\omega), Y_{\mathrm{L}}(\omega)]^{\mathrm{T}}$. 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 $\mathbf{G}(\omega)$, and the $M \times N$ inverse filter matrix [7] is expressed as $\mathbf{H}(\omega)$. $\mathbf{Y}(\omega)$ is denoted by

$$\mathbf{Y}(\omega) = \mathbf{G}(\omega)\mathbf{H}(\omega)\mathbf{X}(\omega), \quad (2)$$

where $\mathbf{G}(\omega)\mathbf{H}(\omega) = \mathbf{I}_N$, and $\mathbf{I}_N$ 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 ($[Y_{\mathrm{L}}, Y_{\mathrm{R}}] = [X_{\mathrm{L}}, X_{\mathrm{R}}]$) and silent zones are materialized at each microphone elements ($[Y_{\mathrm{mic}1}, \ldots, Y_{\mathrm{mic}K}] =$

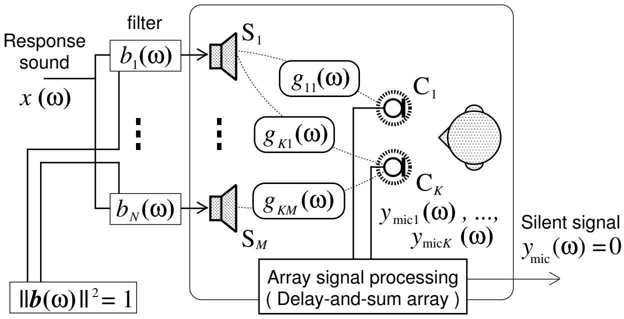

Figure 2: Configuration of proposed method.

$[0,\ldots,0]$). 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.

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 $A_k(\omega)(k = 1,2,\ldots,K)$ at the k-th element of a delay-and-sum array is designated as

$$A_k(\omega) = (1/K) \cdot e^{-j\omega\tau_k}, \tag{3}$$

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 $Y_{\mathrm{m}ic}(\omega)$ is given by

$$Y_{\mathrm{mic}} = \sum_{k=1}^{K} A_k(\omega) Y_{\mathrm{mic}K}(\omega). \tag{4}$$

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 $\mathbf{H}(\omega)$ 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 $\mathbf{H}(\omega)$ is not decreased, since the solution of $\mathbf{H}(\omega)$ is indeterminate, we adopt the Moore-Penrose generalized inverse matrix as the inverse filter which provides the LNS [6].

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.

3. PROPOSED METHOD: RESPONSE SOUND CANCELLATION

When we premise that the user can move around, control of the sound field around the user's ears in MOMNI method becomes meaningless. In order to satisfy the condition $M > N = K + 2$, control of the user's ears causes increase of loudspeakers or decrease of microphone elements. However, in an inverse filter used by MOMNI method, an output signal and a control point must be a pair. Thus MOMNI method cannot present user the response sound without setting the corresponding control points at the user's ears.

In this section, we propose a new filter design algorithm to provide silent zones on 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.

3.1 Sound field control to cancel out the response sound

In Fig. 2, $\mathrm{S}_m(m = 1,\ldots,M)$ denotes the loudspeaker, and C_k $(k = 1,\ldots,K)$ is the microphone. The numbers of loudspeakers and microphone elements must satisfy the condition $M > K$. The observation signals at each of the control points are described as $\mathbf{Y}(\omega) = [Y_1(\omega),\ldots,Y_K(\omega)]^{\mathrm{T}}$, where $Y_k(\omega)$ $(k = 1,\ldots,K)$ is the signal observed on the microphone C_k. Response sound is monaural, and denoted by a scalar $X(\omega)$. The response sound signal is outputted from the loudspeakers after processed by filters. The filter coefficients are represented by $\mathbf{B}(\omega) = [B_1(\omega),\ldots,B_M(\omega)]^{\mathrm{T}}$, where $B_m(\omega)$ $(m = 1,\ldots,M)$ is the filter corresponding to the loudspeaker S_m. The $M \times K$ matrix composed of the room transfer function $G_{km}(\omega)$ between the secondary sound source S_m and the control point C_k is denoted by $\mathbf{G}(\omega)$, and $\mathbf{Y}(\omega)$ is denoted by

$$\mathbf{Y}(\omega) = \mathbf{G}(\omega)\mathbf{B}(\omega)X(\omega). \tag{5}$$

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

$$\mathbf{G}(\omega)\mathbf{B}(\omega) = \mathbf{0} \quad \text{subject to} \quad \|\mathbf{B}(\omega)\| = C, \tag{6}$$

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

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

3.2 Producing vectors which span nullspace

Equation (6) shows that $\mathbf{B}(\omega)$ is orthogonal to all rows of $\mathbf{G}(\omega)$. The M-dimensional subspace which includes all orthogonal vectors to all rows of $\mathbf{G}(\omega)$ is called *n*ullspace of $\mathbf{G}(\omega)$. Singular value decomposition provides the vectors which span the nullspace of $\mathbf{G}(\omega)$ in the form of eigenvectors which correspond to zero singular values. The filter coefficients $\mathbf{B}(\omega)$ can be designed by linear summation of these vectors.

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

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

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

3.3 Filter coefficients closest to the impulses

Although any appropriately normalized linear summation of nullspace vectors satisfies the condition eq. (6), the output sound becomes extremely distorted if the expanded coefficients are selected

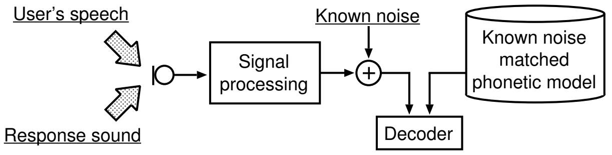

Figure 3: Layout of acoustic experiment room.

randomly. In order to eliminate the distortion, we utilize the solution whose filter coefficients are closest to impulses which has a property of full bandpass and linear phase.

We definite the matrix $\mathbf{W}(\omega)$ whose columns are nullspace vectors of $\mathbf{G}(\omega)$ as

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

and then solve the following least squares problem to obtain the expanded coefficient vector $\alpha(\omega)$;

$$\min_{\alpha(\omega)} ||\mathbf{W}(\omega)\alpha(\omega) - \mathbf{L}||^2, \tag{9}$$

where $\mathbf{L}$ is the filter coefficients of unit impulses and given by M-dimensional vector all of whose components are 1. The solution $\alpha(\omega)$ is the linear expansion coefficients of nullspace vectors to construct the filter coefficients closest to the unit impulses that have the same gain and the same timed peaks. Since each filter coefficient becomes close to the impulse which has a property of full bandpass and linear phase, the output of each loudspeaker becomes less distorted.

The solution of Eq. (9) is given by

$$\begin{aligned}\alpha(\omega) &= \left(\mathbf{W}^{\mathrm{H}}(\omega)\mathbf{W}(\omega)\right)^{-1}\mathbf{W}^{\mathrm{H}}(\omega)\mathbf{L}(\omega) \\ &= \mathbf{W}^{\mathrm{H}}(\omega)\mathbf{L}(\omega).\end{aligned} \tag{10}$$

With $\alpha(\omega)$ we can obtain the filter coefficients $\mathbf{B}'(\omega)$ which are closest to the impulses, as

$$\begin{aligned}\mathbf{B}'(\omega) &= \mathbf{W}(\omega)\alpha(\omega) \\ &= \mathbf{W}(\omega)\mathbf{W}^{\mathrm{H}}(\omega)\mathbf{L}(\omega).\end{aligned} \tag{11}$$

Finally, we can obtain the resultant filter coefficients $\mathbf{B}(\omega)$ by normalizing $\mathbf{B}'(\omega)$ with its norm to satisfy the condition of norm in Eq. (6), as

$$\begin{aligned}\mathbf{B}(\omega) &= C\frac{\mathbf{B}'(\omega)}{\|\mathbf{B}'(\omega)\|} \\ &= C\frac{\mathbf{W}(\omega)\mathbf{W}^{\mathrm{H}}(\omega)\mathbf{L}(\omega)}{\sqrt{\mathbf{L}^{\mathrm{H}}(\omega)\mathbf{W}(\omega)\mathbf{W}^{\mathrm{H}}(\omega)\mathbf{L}(\omega)}}.\end{aligned} \tag{12}$$

4. EXPERIMENTS AND RESULTS

In this section, we present two experiments comparing the conventional methods (acoustic echo canceller and MOMNI method) and the proposed method. In order to verify the applicability of the proposed method, we simulate 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 another person 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 the interference does not exist. Figure 3 shows the arrangement of the apparatuses. As shown in Fig. 3, we place the 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 experiment room, where the reverberation time is approximately 160 ms, with 48 kHz sampling and 16 bit resolution. The loudspeakers used by 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 secondary sound sources is M (M = 5, 8 or 12), and the number of control points on the microphone elements is N (N = 1, 2, 3 or 4) (hereafter we label the transfer system "M-N system"), are designed. Also, the passband range is 150–4000 Hz. We use a circular microphone array with twelve elements, and we select the elements which are equally spaced. It is worth nothing that the distance between the loudspeaker for acoustic echo canceller and the microphone array is shorter than those for the sound field control, and consequently the time casualty does not hold in these sound field control. Indeed, the 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.

The filters of the MOMNI and proposed method is designed with the room transfer functions without the interference. The evaluation is done with the average of 12 fluctuations. The filter coefficient of the acoustic echo canceller is constructed without a specific adaptive algorithm. In this experiment, we assume that the echo canceller can estimate the filter coefficient precisely under the ideal condition without error. The echo canceller is adapted precisely before the fluctuation, but after fluctuation the adaptation cannot be performed because of the double talk. The evaluation is also done with the average of 12 fluctuations.

4.2 Response sound elimination experiment

To evaluate both the performance of response sound elimination and the ability to present the response sound to the user, we calculate the barge-in reduction rate (BRR); which is defined by

$$\mathrm{BRR} = 10\log_{10}\frac{\sum_\omega |Y_{\mathrm{ear}}(\omega)|^2}{\sum_\omega |Y_{\mathrm{out}}(\omega)|^2} \quad [\mathrm{dB}], \tag{13}$$

where $Y_{\mathrm{ear}}(\omega)$ is the response sound signal observed at the ear of the user (dummy head), and $Y_{\mathrm{out}}(\omega)$ is the output in each method.

As the response sound from the dialogue system, we use a female utterance selected from the ASJ database [9]. Although the sampling frequency of the response sound is 16 kHz, we use the signal in which the frequency component over 4 kHz is eliminated.

Figure 5(a) and (b) shows the BRRs with all the combinations of the number of loudspeakers and microphone elements. As compared with the result of the acoustic echo canceller, the proposed method shows higher performance than that of the acoustic echo canceller in all combinations. In the case of the 5 loudspeakers, the proposed method shows higher performance than that of MOMNI method. Especially the proposed method of 5-3 system shows the highest performance of 28.8 dB in all of these results. With more loudspeakers, MOMNI method shows the improvements, but the increase of performance in the proposed method is not obvious. Thus the proposed method is highly beneficial for the application to the small number of loudspeakers.

4.3 Speech recognition experiment

The effect of the elimination of response sound is evaluated with a large vocabulary continuous speech recognition task. In order to

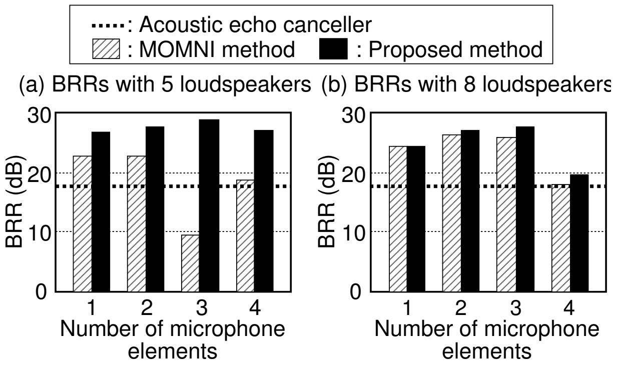

Figure 4: Comparison of BRRs.

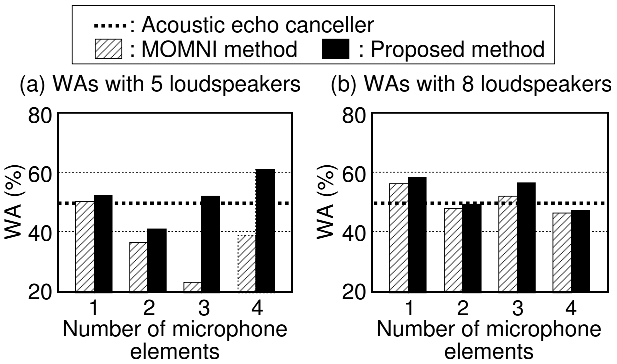

Figure 5: Comparison of WAs.

Table 1: Experimental conditions for speech recognition

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

valuate the speech recognition performance, we adopt the Word Accuracy (WA) as an evaluation score. WA is given by

$$\mathrm{WA}[\%] = \frac{W-S-D-I}{W} \times 100, \tag{14}$$

where W 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 the speech recognition. We average each WA which is obtained from 200 speech in total.

The speech signal, which is obtained by superimposing the elimination error of response sound, $E_{\mathrm{out}}(\omega)$, 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 PTM (Phonetic Tied Mixture model based on triphones) speaker-independent.

Figure 5 (c) and (d) shows the WAs with all the combinations. All the scores of the graph are similar to those of BRRs, 5-4 system shows the highest performance.

5. CONCLUSION

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

Acknowledgement

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

REFERENCES

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