IEICE TRANS. FUNDAMENTALS, VOL.E89–A, NO.3 MARCH 2006

PAPER Special Section on Multidimensional Signal Processing and Its Application

Interface for Barge-in Free Spoken Dialogue System Using Nullspace Based Sound Field Control and Beamforming

Shigeki MIYABE[†a)], *Nonmember*, **Hiroshi SARUWATARI**[†], *Member*, **Kiyohiro SHIKANO**[†], *Fellow*, *and* **Yosuke TATEKURA**[††], *Member*

SUMMARY In this paper, we describe a new interface for a barge-in free spoken dialogue system combining multichannel sound field control and beamforming, in which the response sound from the system can be canceled out at the microphone points. The conventional method inhibits a user from moving because the system forces the user to stay at a fixed position where the response sound is reproduced. However, since the proposed method does not set control points for the reproduction of the response sound to the user, the user is allowed to move. Furthermore, the relaxation of strict reproduction for the response sound enables us to design a stable system with fewer loudspeakers than those used in the conventional method. The proposed method shows a higher performance in speech recognition experiments.

key words: *spoken dialogue system, barge-in, sound field control, beamforming, nullspace*

1. Introduction

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

To eliminate the response sound from the system, an acoustic echo canceller is commonly used. Many types of acoustic echo canceller have been proposed, e.g., single channel, stereophonic, wave synthesis, and beamformer-integrated types [2]–[5]. However, the acoustic echo canceller has an inherent problem in which accurate adaptation is difficult in a barge-in situation (this is also called "double-talk problem"). Because of this problem, the conventional acoustic echo canceller should stop adaptation during barge-in; this implies that the elimination performance is likely to degrade when a change in room transfer functions arises during barge-in.

To solve the problem of the acoustic echo canceller, one of the authors has proposed the Multiple-Output and Multiple-No-Input (MOMNI) method [6], which combines sound field control and beamforming. By increasing the number of loudspeakers and microphone elements, the MOMNI method can make its control robust against the fluctuation of room transfer functions, but a large number of loudspeakers are needed to achieve sufficient robustness for speech recognition. Furthermore, the MOMNI method controls the sound field only around the user's ears, and premises that the user does not move from the assumed specific position.

To address the problems of the MOMNI method, we propose a new filter design method for realizing silence at positions of the microphone elements without reproducing the response sound at the user's particular position. First, singular value decomposition is utilized to provide vectors that span the nullspace of the matrix of room transfer functions among the loudspeakers and microphones. Nullspace vectors are assumed to be the filter candidates that can realize silence at the microphone positions. Second, the linear summation of the vectors closest to the delayed impulse yields the resultant filter coefficients corresponding to the nullspace, while maintaining better sound qualities. The relaxation of the strict reproduction of the response sound can reduce the number of loudspeakers while maintaining stable control and allowing the user to move.

A computer simulation using impulse responses measured in a real acoustic environment reveals that the proposed method is more robust against fluctuation of the room transfer function than the conventional methods even when there are few loudspeakers. However, although the proposed method with many microphone elements improves its speech recognition performance, the quality of response sound is speculated to slightly degrade. We discuss the trade-off between speech recognition performance and sound quality in our sound quality assessment experiment.

The outline of this paper is as follows: The algorithm of the MOMNI method is reviewed in Sect. 2. The proposed filter design method is described in Sect. 3. The performance of the proposed method is discussed in Sect. 4. Finally, the conclusion of this study is provided in Sect. 5.

Manuscript received June 27, 2005.
Manuscript revised October 3, 2005.
Final manuscript received November 17, 2005.

[†]The authors are with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630-0192 Japan.
[††]The author is with the Faculty of Engineering, Shizuoka University, Hamamatsu-shi, 432-8561 Japan.
a) E-mail: shige-m@is.naist.jp

DOI: 10.1093/ietfec/e89–a.3.716

2. Conventional MOMNI Method [6]

We describe the MOMNI method shown in Fig. 1. The MOMNI method consists of two main steps, namely, sound field control and beamforming.

2.1 Sound Field Control

We first design an inverse filter of room transfer functions for sound field reproduction [7]. With this inverse filter, any signal can be reproduced at the control points. In Fig. 1, S_m ($m = 1, \ldots, M$) are the loudspeakers that act as secondary sound sources, and C_n ($n = 1, \ldots, N$) are the microphones that act as control points. $C_1, \ldots, C_K$ ($K = N-2$) are located in each microphone element for recording the user's speech, and C_{N-1} and C_N are placed in the vicinity of the two external auditory meatuses of the user. The relationship between the number of loudspeakers and that of microphones must satisfy the condition

$$M > N = K + 2. \tag{1}$$

The intended signals to be reproduced at respective control points are represented as

$$\boldsymbol{x}(\omega) = [x_{\mathrm{mic}1}(\omega), \ldots, x_{\mathrm{mic}K}(\omega), x_{\mathrm{R}}(\omega), x_{\mathrm{L}}(\omega)]^{\mathrm{T}}, \tag{2}$$

where $[\cdot]^{\mathrm{T}}$ describes transposition, ω denotes angular frequency, $x_{\mathrm{mic}k}(\omega)$ ($k = 1, \ldots, K$) are the signals to be reproduced at microphone C_k, and $x_{\mathrm{R}}(\omega)$ and $x_{\mathrm{L}}(\omega)$ are response sounds to be reproduced at the right and left ears of the user, respectively. Similarly, the observation signals at the control points are given by

$$\boldsymbol{y}(\omega) = [y_{\mathrm{mic}1}(\omega), \ldots, y_{\mathrm{mic}K}(\omega), y_{\mathrm{R}}(\omega), y_{\mathrm{L}}(\omega)]^{\mathrm{T}}. \tag{3}$$

The $N \times M$ matrix, which is composed of the room transfer functions $g_{nm}(\omega)$ between the secondary sound sources S_m and the control points C_n, is denoted by $\boldsymbol{G}(\omega)$ as

$$\boldsymbol{G}(\omega) = \begin{bmatrix} g_{11}(\omega) & \cdots & g_{1M}(\omega) \\ \vdots & \ddots & \vdots \\ g_{N1}(\omega) & \cdots & g_{NM}(\omega) \end{bmatrix}, \tag{4}$$

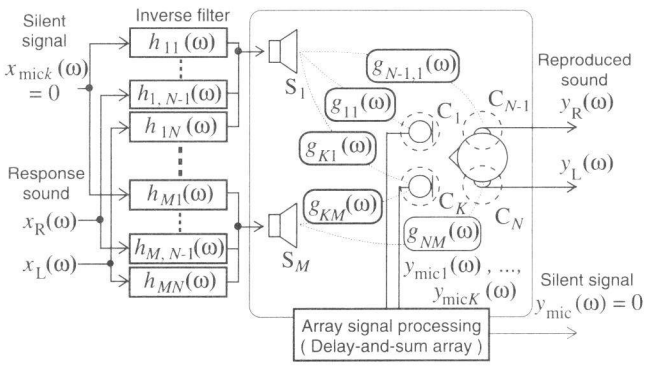


Fig. 1 Configuration of the conventional MOMNI method.

and the $M \times N$ inverse filter matrix $\boldsymbol{H}(\omega)$ is expressed as

$$\boldsymbol{H}(\omega) = \begin{bmatrix} h_{11}(\omega) & \cdots & h_{1N}(\omega) \\ \vdots & \ddots & \vdots \\ h_{M1}(\omega) & \cdots & h_{MN}(\omega) \end{bmatrix}. \tag{5}$$

Here, $\boldsymbol{y}(\omega)$ is denoted by

$$\boldsymbol{y}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{H}(\omega) = \boldsymbol{x}(\omega), \tag{6}$$

where $\boldsymbol{G}(\omega)\boldsymbol{H}(\omega) = \boldsymbol{I}_N$, and $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. We design the inverse filter $\boldsymbol{H}(\omega)$ by applying the least norm solution (LNS) in the frequency domain [9]. In the case where the rank of $\boldsymbol{H}(\omega)$ is not decreased, the solution of $\boldsymbol{H}(\omega)$ is indeterminate. We adopt the Moore-Penrose generalized inverse matrix as the inverse filter matrix that provides the LNS.

In Eq. (3), the response sounds of a dialogue system are reproduced at both ears of the user (i.e., $[y_{\mathrm{L}}, y_{\mathrm{R}}] = [x_{\mathrm{L}}, x_{\mathrm{R}}]$), and silent zones are materialized at each microphone element (i.e., $[y_{\mathrm{mic}1}, \ldots, y_{\mathrm{mic}K}] = [0, \ldots, 0]$). Thereby, we can actualize the sound field that gives the user the response sound while preventing it from mixing with the observation signal at each microphone element.

2.2 Beamforming 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 speech 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

$$A_k(\omega) = (1/K) \cdot e^{-j\omega\tau_k}, \tag{7}$$

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

$$y_{\mathrm{mic}} = \sum_{k=1}^{K} A_k(\omega) y_{\mathrm{mic}K}(\omega). \tag{8}$$

2.3 Response Sound Elimination Error When Changing Room Transfer Functions

By approximating the fluctuation of the room transfer function to random variables, it is proved that the MOMNI method is robust against the fluctuation of the room transfer functions [6]. Assume that the fluctuation Δg_{nm} caused by the fluctuation of transfer functions is added to the original transfer function $g_{nm}(\omega)$ as

$$g'_{nm}(\omega) = g_{nm}(\omega) + \Delta g_{nm}(\omega) \tag{9}$$

where $g'_{nm}(\omega)$ describes the room transfer function after fluctuation. After fluctuation, observation signal $\boldsymbol{y}'(\omega)$ differs

from $\boldsymbol{y}(\omega)$ in Eq. (6) and can be denoted by

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

where $\boldsymbol{G}'(\omega)$ and $\Delta\boldsymbol{G}(\omega)$ are $N \times M$ matrices composed of $g'_{nm}(\omega)$ and $\Delta g_{nm}(\omega)$. Then the elimination error of the response sound at the array output is represented as

$$\begin{aligned}\Delta y_{\mathrm{mic}}(\omega) = \sum_{k=1}^{K} A_k(\omega)\Big\{ \sum_{m=1}^{M} \Delta g_{km}(\omega) \\ \cdot\big(h_{m(N-1)}(\omega)x_{\mathrm{R}}(\omega) \\ + h_{mN}(\omega)x_{\mathrm{L}}(\omega)\big)\Big\}.\end{aligned} \tag{11}$$

Denoting the matrix norm of $\boldsymbol{H}(\omega)$ as $\|\boldsymbol{H}(\omega)\|$, we can rewrite Eq. (11) as

$$\begin{aligned}\Delta y_{\mathrm{mic}}(\omega) = \|\boldsymbol{H}(\omega)\| \cdot \frac{1}{K} \cdot \Big\{ \sum_{k=1}^{K} \sum_{m=1}^{M} \Delta g_{km}(\omega) \\ \cdot\big(\mathcal{H}_{m(N-1)}(\omega)x_{\mathrm{R}}(\omega) \\ + \mathcal{H}_{mN}(\omega)x_{\mathrm{L}}(\omega)\big)e^{-j\omega\tau_k}\Big\},\end{aligned} \tag{12}$$

where $\mathcal{H}_{mn}(\omega) = h_{mn}(\omega)/\|\boldsymbol{H}(\omega)\|$. It is assumed that $\Delta g_{nm}(\omega)$ is a Gaussian random variable with the variance σ^2. Furthermore, since $\mathcal{H}_{mn}(\omega)$ is normalized by $\|\mathcal{H}(\omega)\|$ and is independent of the change in M, the variance in $\{\cdot\}$ of Eq. (12) can be expressed as $\eta\sqrt{M \cdot K}\sigma$, where η is an appropriate constant. In addition, $\|\boldsymbol{H}\|$ is proportional to $1/M$ because $\|\boldsymbol{H}(\omega)\| \simeq 1/\|\boldsymbol{G}(\omega)\| \propto 1/M$. Therefore, the following relation holds in the elimination error of response sound $\mathcal{E}(\omega)$:

$$\begin{aligned}\mathcal{E}(\omega) = \Delta y_{\mathrm{mic}}(\omega) &\propto (1/M) \cdot (1/K) \cdot \sqrt{M \cdot K} \\ &= 1/\sqrt{M \cdot K}.\end{aligned} \tag{13}$$

Equation (13) shows that the elimination error of response sound is inversely proportional to $\sqrt{M \cdot K}$. Therefore, if the number of transfer channels between the loudspeakers and microphones is increased, the MOMNI method becomes more robust against the change in transfer functions than an acoustic echo canceller.

2.4 Problems in MOMNI Method

Since the MOMNI method must satisfy the condition $M > N = K + 2$, two additional loudspeakers are required to control the sound field at both of the user's ears. If we want to construct a small-scale system with fewer loudspeakers, setting these two control points at the user's ears is a barrier to securing sufficient microphone elements for robustness. Moreover, if we premise that the user can move around, the strict reproduction at these two control points becomes meaningless. To allow movement of the user and reduce 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 control points but reproduction of input signal. Silent zones are realized by cancellation of input signal but reproduction of zero signals. Since no control points other than the microphone elements are set, sound field control can be performed stably with fewer loudspeakers.

3.1 Sound Field Control for Cancelling Out Response Sound

In Fig. 2, $\mathrm{S}_m (m = 1, \ldots, M)$ denote the loudspeakers and C_k $(k = 1, \ldots, K)$ represent the control points where microphones are set. The numbers of loudspeakers and microphone elements must satisfy the condition

$$M > K. \tag{14}$$

The observed signals at the control points are designated as

$$\boldsymbol{y}(\omega) = [y_1(\omega), \ldots, y_K(\omega)]^{\mathrm{T}}, \tag{15}$$

where $y_k(\omega)$ $(k = 1, \ldots, K)$ are the signals observed at the microphones C_k. The response sound is monaural and denoted by a scalar $x(\omega)$. The response sound is outputted from the loudspeakers after being processed by filters. The filter coefficients are represented by

$$\boldsymbol{b}(\omega) = [b_1(\omega), \ldots, b_M(\omega)]^{\mathrm{T}}, \tag{16}$$

where $b_m(\omega)$ $(m = 1, \ldots, M)$ are the filter coefficients corresponding to the loudspeakers S_m. The $M \times K$ matrix, which is composed of the room transfer functions $g_{km}(\omega)$ between the loudspeakers S_m and the control points C_k, is denoted by $\boldsymbol{G}(\omega)$ as

$$\boldsymbol{G}(\omega) = \begin{bmatrix} g_{11}(\omega) & \cdots & g_{1M}(\omega) \\ \vdots & \ddots & \vdots \\ g_{K1}(\omega) & \cdots & g_{KM}(\omega) \end{bmatrix}, \tag{17}$$

and $\boldsymbol{y}(\omega)$ is denoted by

$$\boldsymbol{y}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{b}(\omega)x(\omega). \tag{18}$$

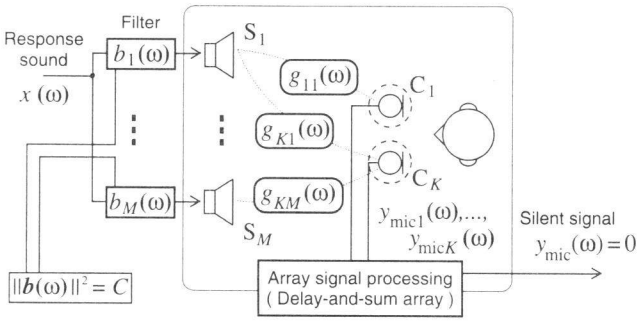


Fig. 2 Configuration of proposed method.

The following condition must be satisfied when any response sounds are canceled out at the positions of microphone elements.

$$\boldsymbol{G}(\omega)\boldsymbol{b}(\omega) = \boldsymbol{0} \tag{19}$$

$$\text{subject to} \quad \|\boldsymbol{b}(\omega)\| = C \tag{20}$$

where, $\boldsymbol{0}$ is a K-dimensional column zero vector and C is a constant for adjusting 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 that output no signal.

3.2 Extracting Vectors That Span Nullspace

Equation (19) shows that $\boldsymbol{b}(\omega)$ is orthogonal to the row space of $\boldsymbol{G}(\omega)$. The subspace that includes all orthogonal vectors in all rows of $\boldsymbol{G}(\omega)$ is called the *nullspace* of $\boldsymbol{G}(\omega)$. Singular value decomposition provides the vectors that span the nullspace of $\boldsymbol{G}(\omega)$ in the form of eigenvectors that correspond to zero singular values. The filter coefficients $\boldsymbol{b}(\omega)$ can be designed by the linear summation of these vectors.

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

$$\boldsymbol{G}(\omega) = \boldsymbol{U}(\omega)\left[\, \boldsymbol{\Lambda}(\omega) \mid \boldsymbol{O}_{K,M-K} \,\right]\boldsymbol{V}^{\mathrm{H}}(\omega), \tag{21}$$

where $\boldsymbol{O}_{i,j}$ is an $i \times j$ zero matrix, and $[\cdot]^{\mathrm{H}}$ represents the conjugate transpose. $\boldsymbol{\Lambda}(\omega)$ is a $K \times K$ diagonal matrix:

$$\boldsymbol{\Lambda}(\omega) = \begin{bmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_{R_\omega} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}. \tag{22}$$

R_ω is the rank of $\boldsymbol{G}(\omega)$, and its diagonal elements $\{\lambda_1(\omega), \ldots, \lambda_{R_\omega(\omega)}, 0, \ldots, 0\}$ 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 singular values $\{\lambda_1(\omega), \ldots, \lambda_{R_\omega(\omega)}\}$, respectively, and the rest of the 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 zero singular values. In particular, $\{\boldsymbol{v}_{R_\omega+1}(\omega), \ldots, \boldsymbol{v}_M(\omega)\}$ are the nullspace vectors we need. These vectors exist because of the condition $M - R_\omega \geq M - K > 0$.

The nullspace vectors $\boldsymbol{v}_r(\omega)$ $(r = R_\omega + 1, \ldots, M)$ satisfy Eq. (19) as shown by

$$\begin{aligned} \boldsymbol{G}(\omega)\boldsymbol{v}_r(\omega) &= \boldsymbol{U}(\omega)\left[\, \boldsymbol{\Lambda}(\omega) \mid \boldsymbol{O}_{K,M-K} \,\right] \\ &\quad \cdot \boldsymbol{V}^{\mathrm{H}}(\omega)\boldsymbol{v}_r(\omega) \\ &= \boldsymbol{U}(\omega)\left[\, \boldsymbol{\Lambda}(\omega) \mid \boldsymbol{O}_{K,M-K} \,\right] \\ &\quad \cdot [\underbrace{0, \ldots, 0}_{r-1}, 1, \underbrace{0, \ldots, 0}_{M-r}]^{\mathrm{T}} \\ &= \boldsymbol{U}(\omega)\boldsymbol{0} \\ &= \boldsymbol{0}. \end{aligned} \tag{23}$$

Therefore, any summation of the nullspace vectors $\boldsymbol{b}'(\omega)$, which is a model of the filter $\boldsymbol{b}(\omega)$, can be denoted by

$$\boldsymbol{b}'(\omega) = \sum_{r=R_\omega+1}^{M} \alpha_r(\omega)\boldsymbol{v}_r(\omega), \tag{24}$$

where α_r $(r = R_\omega + 1, \ldots, M)$ are arbitrary complex values. This vector $\boldsymbol{b}'(\omega)$ satisfies Eq. (19) as shown by

$$\begin{aligned} \boldsymbol{G}(\omega)\boldsymbol{b}'(\omega) &= \sum_{r=R_\omega+1}^{M} \alpha_r(\omega)\boldsymbol{G}(\omega)\boldsymbol{v}_r(\omega) \\ &= \sum_{r=R_\omega+1}^{M} \alpha_r(\omega)\boldsymbol{0} \\ &= \boldsymbol{0}. \end{aligned} \tag{25}$$

In another denotation,

$$\boldsymbol{b}'(\omega) = \boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega), \tag{26}$$

where

$$\boldsymbol{W}(\omega) = \left[\boldsymbol{v}_{R_\omega+1}(\omega), \ldots, \boldsymbol{v}_M(\omega)\right], \tag{27}$$

and $\boldsymbol{\alpha}(\omega)$ is an $(M - R_\omega)$-dimensional complex vector given by

$$\boldsymbol{\alpha}(\omega) = \left[\alpha_{R_\omega+1}, \ldots, \alpha_M\right]^{\mathrm{T}}. \tag{28}$$

Since $\boldsymbol{b}'(\omega)$ does not satisfy the norm condition of Eq. (20), the resultant filter $\boldsymbol{b}(\omega)$ is obtained by normalizing $\boldsymbol{b}'(\omega)$ with its norm, as

$$\begin{aligned} \boldsymbol{b}(\omega) &= C\frac{\boldsymbol{b}'(\omega)}{\sqrt{\|\boldsymbol{b}'(\omega)\|^2}} \\ &= C\frac{\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega)}{\sqrt{\boldsymbol{\alpha}^{\mathrm{H}}(\omega)\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega)}}. \end{aligned} \tag{29}$$

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

$$\boldsymbol{v}_i^{\mathrm{H}}(\omega)\boldsymbol{v}_j(\omega) = \begin{cases} 1 & \text{for} \quad i = j; \\ 0 & \text{for} \quad i \neq j. \end{cases} \tag{30}$$

Accordingly,

$$\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{W}(\omega) = \boldsymbol{I}_{M-R_\omega}, \tag{31}$$

where $\boldsymbol{I}_{M-R_\omega}$ denotes $(M - R_\omega) \times (M - R_\omega)$ identity matrix. The substitution of Eq. (31) in Eq. (29) leads to

$$\boldsymbol{b}(\omega) = C\frac{\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega)}{\sqrt{\boldsymbol{\alpha}^{\mathrm{H}}(\omega)\boldsymbol{\alpha}(\omega)}}. \tag{32}$$

3.3 Filter Coefficients Closest to Impulses

In the previous section, we showed that the conditions of Eqs. (19) and (20) are satisfied by $\boldsymbol{b}(\omega)$ in Eq. (32), i.e., any appropriately normalized linear summations of the nullspace vectors. However, the output sound becomes extremely distorted if the expansion coefficients $\boldsymbol{\alpha}(\omega)$ are selected randomly. Indeed, Fig. 3 shows an example of the

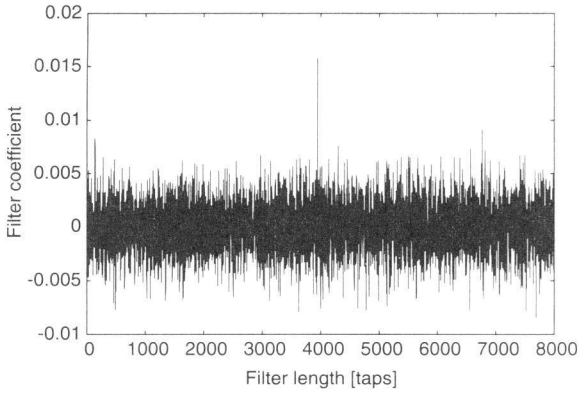


Fig. 3 Example of waveform of filter coefficients designed by random summation of the nullspace vectors. The filter is designed with eight loudspeakers and two microphones, and corresponds to one loudspeaker. The filter is designed with FFT points of 16384, and cut by a rectangle window of 8000 points. The sampling frequency is 16 kHz and the bandwidth is 150–4000 Hz.

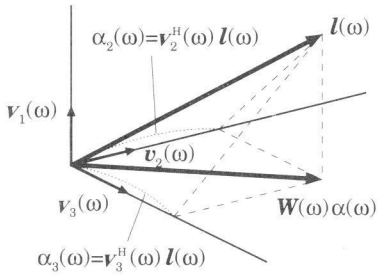


Fig. 4 Example of the projection of $\boldsymbol{l}(\omega)$ to the nullspace of $\boldsymbol{G}(\omega)$ when the row space of $\boldsymbol{G}(\omega)$ is spanned by $\boldsymbol{v}_1(\omega)$, and $\boldsymbol{W}(\omega) = [\boldsymbol{v}_2(\omega)\ \boldsymbol{v}_3(\omega)]$.

filter designed by random summation of the nullspace vectors, where we see a large undesired pre/post-echo. In the following, we propose an algorithm for designing a filter with a small distortion by utilizing the solution closest to impulses.

We define the following filter coefficient vector $\boldsymbol{l}(\omega)$ whose components are the filter coefficients of the impulses with the same amplitude and the same latency τ.

$$\boldsymbol{l}(\omega) = e^{j\omega\tau}[\underbrace{1, \ldots, 1}_{M}]^{\mathrm{T}} \tag{33}$$

Then we try to find a vector closest to the target vector $\boldsymbol{l}(\omega)$ within the nullspace. The output of each loudspeaker becomes less distorted because each filter coefficient closely approaches the impulse that has a full bandpass property and a linear phase. We can obtain the optimal expanded coefficient vector $\boldsymbol{\alpha}(\omega)$ by solving the following least squares problem:

$$\min_{\boldsymbol{\alpha}(\omega)} \|\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{\alpha}(\omega) - \boldsymbol{l}(\omega)\|^2. \tag{34}$$

Such a vector $\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega)$ can be obtained by projection of $\boldsymbol{l}(\omega)$ to the nullspace of $\boldsymbol{G}(\omega)$, or in other words, the column space of $\boldsymbol{W}(\omega)$, as an example shown in Fig. 4. Such expanded coefficients $\alpha_r(\omega)$ $(r = R_\omega + 1, \ldots, M)$ that satisfy

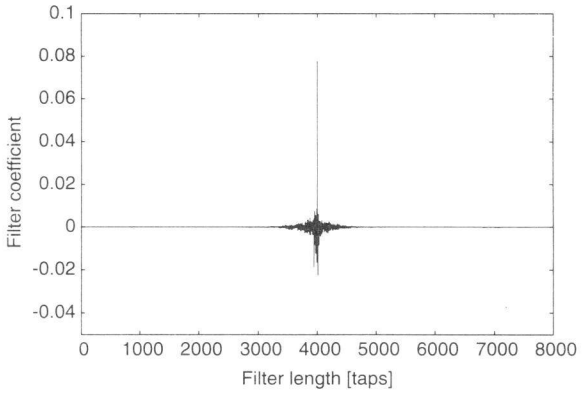


Fig. 5 Example of waveform of filter coefficients designed by proposed method. The filter is designed with eight loudspeakers and two microphones, and corresponds to one loudspeaker. The filter is designed with FFT points of 16384, and cut by a rectangle window of 8000 points. The sampling frequency is 16 kHz and the bandwidth is 150–4000 Hz.

Eq. (34) can be given by the inner product of $\boldsymbol{v}_r(\omega)$ and $\boldsymbol{l}(\omega)$ as

$$\alpha_r(\omega) = \frac{\boldsymbol{v}_r^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega)}{\boldsymbol{v}_r^{\mathrm{H}}(\omega)\boldsymbol{v}_r(\omega)} = \boldsymbol{v}_r^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega). \tag{35}$$

Therefore, $\boldsymbol{\alpha}(\omega)$ can be given by

$$\boldsymbol{\alpha}(\omega) = \begin{bmatrix} \boldsymbol{v}_{R_\omega+1}^{\mathrm{H}}(\omega) \\ \vdots \\ \boldsymbol{v}_M^{\mathrm{H}}(\omega) \end{bmatrix} \boldsymbol{l}(\omega) = \boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega). \tag{36}$$

Then the resultant filter coefficients $\boldsymbol{b}(\omega)$ is obtained by substituting Eq. (36) in Eq. (32) as

$$\boldsymbol{b}(\omega) = C\frac{\boldsymbol{W}(\omega)\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega)}{\sqrt{\boldsymbol{l}^{\mathrm{H}}(\omega)\boldsymbol{W}(\omega)\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega)}}. \tag{37}$$

Figure 5 shows an example of a filter designed by the proposed method. We can find that its distortion is considerably lower than that in Fig. 3.

3.4 Known-Noise Imposition [15]

In the previous section, we described the response sound reduction procedures. However, there still exists a residual component of the response sound caused by the fluctuation of the transfer function. To obtain optimum recognition performance, we generally need to develop *matched* phoneme models for a speech decoder. However, without a priori information on signal-to-noise ratio, the accurate construction of such matched models is very difficult. To handle many different types of noise, known-noise superposition [15] has been proposed. We apply this technique in the masking of the residual response sound as follows.

1. We impose known-noise to a speech database and train the corresponding matched model using an EM algorithm in advance.

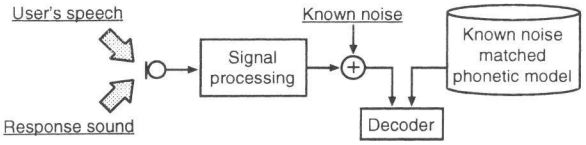


Fig. 6 Configuration of known-noise superposition.

2. We impose known-noise to the noise-reduced output from the delay-and-sum array in the proposed system.
3. We perform speech recognition using a known-noise-matched model for the system output.

Figure 6 shows a configuration of this process.

4. Experiments and Results

In this section, we present two experiments in which the conventional methods (an acoustic echo canceller and the MOMNI method) and the proposed method are compared. To validate the robustness of the proposed method against the fluctuation of the room transfer functions, we perform a response sound elimination experiment in which changes in the transfer functions are simulated. Then we evaluate the sound quality of the response sound. Subsequently, we evaluate the performance of each method on the basis of a speech recognition experiment to verify the applicability of the proposed method. Finally, we assess the sound quality of these methods.

4.1 Experimental Conditions

In the experiments, we premise that the fluctuation of transfer functions is caused by changes in positions of an interfarence, i.e., a life-size mannequin. The interference is arranged under the assumption that another person (the mannequin) approaches the user, which is a very common occurrence in real environments. We measured 13 types of impulse responses: 12 patterns are for the states in which the interference is allocated, and the remaining pattern is for the state in which no interference exists. We used impulse responses without the mannequin as those before fluctuation, and we evaluated the average performance in 12 types of fluctuation. Figure 7 shows the arrangement of the apparatuses. As shown in Fig. 7, we place a dummy head, which has an average human head and an upper body, at the user's position. We designed the filters used in MOMNI and the proposed method with room transfer functions before fluctuation. We gave the acoustic echo canceller the room transfer function before fluctuation as its filter coefficients, assuming that its adaptation was performed accurately without errors before the fluctuation of the transfer functions; however, after the fluctuation, the adaptation could not be performed because of double talk. We evaluated the performances with the average of 12 kinds of impulse responses with the mannequin.

The impulse responses used in this experiment are

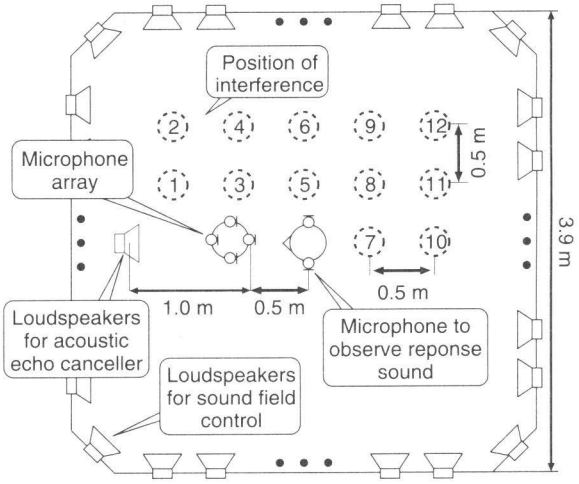


Fig. 7 Layout of acoustic experiment room.

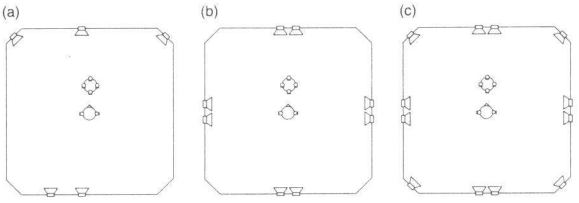


Fig. 8 Exact locations of loudspeakers when (a) five, (b) eight and (c) twelve loudspeakers are used.

measured in an acoustic experimental room. The reverberation time is approximately 160 ms. The sampling rate is with a 48 kHz and resolution is 16-bit. The loudspeakers used in the sound field control of MOMNI and the proposed method are positioned in the outer circumference of the room. The primary sound source of the 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, 8 or 12) 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 exact locations of the loudspeakers are shown in Fig. 8. The passband range is 150–4000 Hz. We use a circular microphone array with 12 elements and select the elements that are spaced equally.

4.2 Response Sound Elimination Experiment

To evaluate the performance of response sound elimination, we calculate barge-in reduction rate (BRR), which is defined by

$$\mathrm{BRR} = 10\log_{10}\frac{\sum_{\omega}|y_{\mathrm{ear}}(\omega)|^2}{\sum_{\omega}|y_{\mathrm{out}}(\omega)|^2} \quad [\mathrm{dB}], \tag{38}$$

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. For instance, a large BRR score indicates a

Table 1 BRRs [dB] of MOMNI method before fluctuation.

	$K=1$	$K=2$	$K=3$	$K=4$
$M=5$	82.1	50.1	15.1	21.8
$M=8$	92.3	90.3	85.4	67.5
$M=12$	98.6	98.0	93.4	74.5

Table 2 BRRs [dB] of the proposed method before fluctuation.

	$K=1$	$K=2$	$K=3$	$K=4$
$M=5$	119.4	101.1	94.5	66.4
$M=8$	117.4	115.0	111.6	88.2
$M=12$	119.3	113.6	112.4	91.1

desirable situation in which the barge-in sound can be removed from the array output while maintaining the presentation of the response sound to the user.

As the response sound from the dialogue system, we use a female utterance selected from the ASJ database [10]. Although the sampling frequency of the response sound is 16 kHz, we use the signal in which the frequency components beyond 4 kHz are eliminated.

We show the BRRs of the MOMNI and proposed method before fluctuation in Tables 1 and 2. The BRR of the acoustic echo canceller before fluctuation was almost infinity. Theoretically, the performances are infinity except for the 5-4 and 5-3 system of the MOMNI method, which do not satisfy the condition (1). However, their performances are not infinity because of the computational error. The effect of the error is not so large that their speech recognition performance does not degrade.

Though the performances of all the method are very high in Tables 1 and 2, they degrade after fluctuation. Figure 9 shows the BRRs for all the combinations of the number of loudspeakers and microphone elements. In this figure, (a), (b) and (c) show the results of 5, 8 and 12 loudspeakers, respectively. The horizontal axis represents the number of microphone elements, and the vertical axis represents the BRR.

The proposed method shows a higher performance than the acoustic echo canceller in all combinations. For the 5 loudspeakers, the proposed method shows a higher performance than the MOMNI method. In particular, the proposed method of the 5-4 system shows the highest performance of 28.8 dB among all of these results. With more loudspeakers, the MOMNI method shows an improvement while the increase in the performance of the proposed method can not be seen. This is because the performance of the proposed method is independent of the number of loudspeakers as discussed in the appendix. Thus, the proposed method is highly beneficial for application to a few loudspeakers.

4.3 Speech Recognition Experiment

The effect of response sound elimination is evaluated using a large vocabulary continuous speech recognition task. To evaluate the speech recognition performance, we adopt word accuracy (WA) as an evaluation score. WA is given by

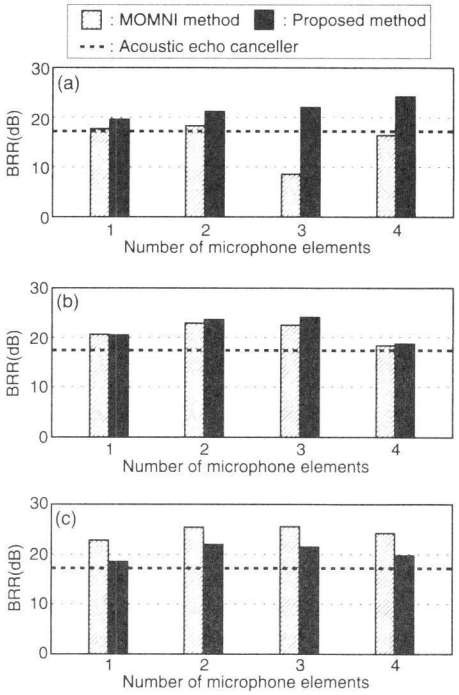


Fig. 9 Comparison of BRR for M loudspeakers: (a) $M=5$, (b) $M=8$ and (c) $M=12$.

Table 3 Experimental conditions for speech recognition.

Speech database	JNAS [11]
Frame length	25 msec (Hamming window)
Frame interval	8 msec
Feature vector	12 MFCCs, 12 ΔMFCCs, Δpower
Language model	Newspaper dictation [12]
Phoneme model	Phonetic Tied Mixture (PTM) [13] (clean or 25 dB office noise imposed)
Decoder	Julius ver. 3.4.2 standard [14]
User's speech (test set)	200 sentences (23 males and 23 females) from JNAS database
Response sound of dialogue system	1 sentence (female) from ASJ database [10]

$$\mathrm{WA}[\%] = \frac{W - S - D - I}{W} \times 100, \tag{39}$$

where W is the total number of words in a 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 3 lists the experimental conditions for the speech recognition. We average each WA obtained from 200 utterances.

We show the configuration of the condition of the speech recognition in Fig. 10. The speech signal obtained by superimposing the elimination error of response sound, $E_{\mathrm{out}}(\omega)$, on the user's speech is used in 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 the MOMNI and proposed methods, 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 the 0 dB state. We use two speaker-independent

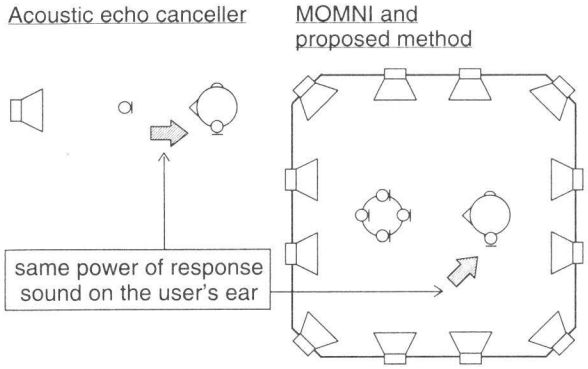


Fig. 10 Configuration of the speech recognition.

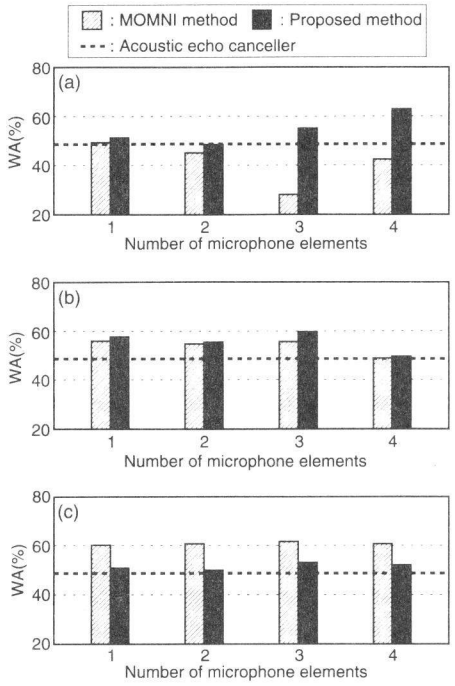


Fig. 11 Comparison of WA with clean model for M loudspeakers: (a) $M = 5$, (b) $M = 8$ and (c) $M = 12$.

phonetic tied mixture (PTM) models based on triphones. One is generated from clean speech, and the other is learned using the speech imposed office noise of 25 dB. In speech recognition with a 25 dB model, the same noise of 30 dB is imposed on the recorded speech signal.

Figures 11 and 12 show the WA for all the combinations. Figure 11 shows the speech recognition performance with a clean model, and Fig. 12 shows that with known-noise imposition. All the scores are similar to those in Fig. 9, e.g., the 5-4 system shows the highest performance in both clean and 25 dB models.

Because of the property of this experiment in which the interference noise is the speech signal, it is difficult to sufficiently prevent insertion error in the user's silent period even when the noise reduction performance is high. Therefore the known-noise imposition technique improves

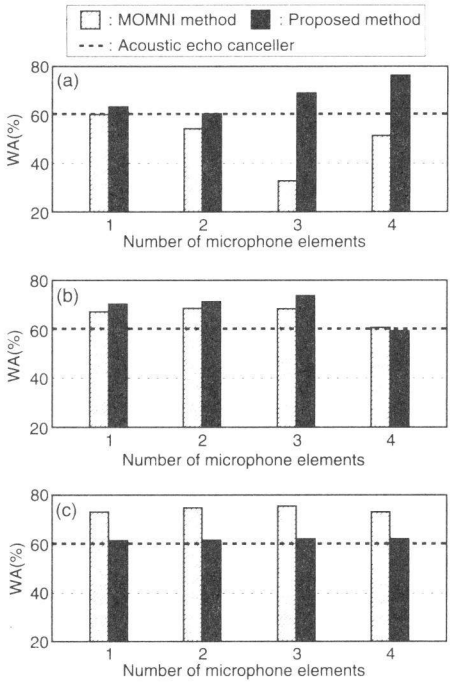


Fig. 12 Comparison of WA with model imposed 25 dB known-noise in the case of M loudspeakers: (a) $M = 5$, (b) $M = 8$ and (c) $M = 12$.

the speech recognition performance considerably. The proposed method for the 5-4 system with known-noise imposition has the highest score of 76.1%. Using the proposed method together with known-noise superposition, an improvement of 27.3% over the performance of the conventional acoustic echo canceller with clean speech is achieved.

4.4 Sound Quality Assessment

Although the proposed method shows high speech recognition performance, the quality of its output sound is not guaranteed because the proposed filter design method maintains only the total gain (see Eq. (20)). In this section, we assess the quality of the response sound produced by the proposed method. We evaluate the sound quality against the user's movement. We show the index of the user's position in Fig. 13. We apply cepstral distance (CD [16]) as an evaluation score. CD is given by

$$\mathrm{CD} = \frac{1}{F}\sum_{f=1}^{F}\frac{20}{\log 10}\sqrt{\sum_{l=1}^{20}2(C_{\mathrm{out}}(l;f)-C_{\mathrm{ref}}(l;f))^2}, \tag{40}$$

where F denotes the number of speech frames, $C_{\mathrm{out}}(l;f)$ the l-th FFT-based cepstrum of the output signal at the f-th frame, and $C_{\mathrm{ref}}(l;f)$ a reference signal for evaluating the distance. The number of liftering points is 20. We average the CDs at both ears. The less CD shows the better sound quality.

Figure 14 shows the CDs between the observed signals at the user's position and original response sound signal

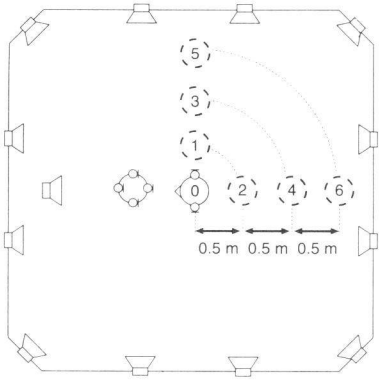


Fig. 13 Configuration of user's movement. The symbol "0" indexes the position where the MOMNI method presents the response sound. Symbols 1, 3 and 5 index the positions 0.5 m, 1.0 m and 1.5 m rightside from position 0, respectively. Similarly, symbols 2, 4 and 6 index the positions 0.5 m, 1.0 m and 1.5 m behind position 0, respectively.

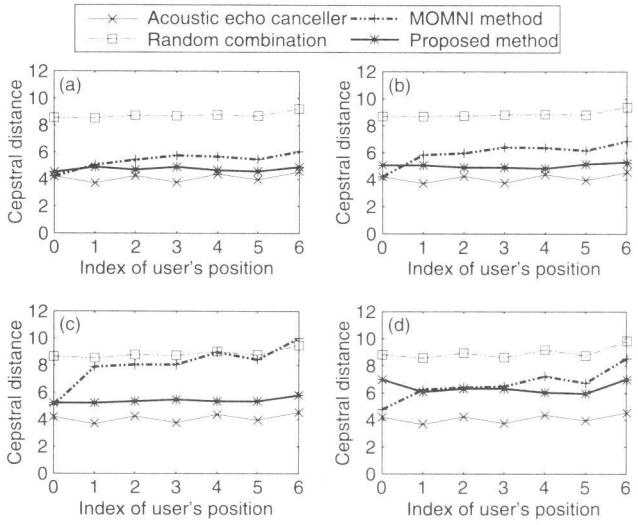


Fig. 14 Comparison of cepstral distance from original response sound signal. The experiments are performed with 5 loudspeakers and (a) 1, (b) 2, (c) 3, and (d) 4 microphone elements except for the acoustic echo canceller.

(dry source). The proposed method is evaluated in comparison with the acoustic echo canceller, the MOMNI method and the filter designed by the random summation of the nullspace vectors explained in Sect. 3.3.

The distortion of the acoustic echo canceller is caused only by the reverberation of the room. Therefore, its CD increases when the user moves away from the loudspeaker, but its influence is very small.

Since the MOMNI method reproduces the output sound of the acoustic echo canceller at position 0, its CDs are similar to those of the acoustic echo canceller. However, when the user moves away from position 0, the increase in the CDs for the MOMNI method is larger than that for the acoustic echo canceller because the observed signals at the user's ears are influenced not only by the reverberation of the room but also by that of the inverse filter.

The sound quality of random summation is very poor, as shown by all the results. The CDs of the proposed method are considerably lower than those of the random-summation nullspace filter. The efficacy of distance minimization on impulses is obvious.

In the proposed method, the degradation of sound quality does not occur regardless of the user's position. The strict reproduction at control points by the MOMNI method distorts the response sound at positions other than the control points. On the other hand, the mitigated presentation of the proposed method has no specific control points where sound quality is high, the distortion of the output signal from filter coefficients close to impulses is not very high throughout the room. The sound quality of the proposed method is almost the same as that of the MOMNI method when the user moves 0.5 m from position 0, and better when the distance is longer than 1.0 m. The sound quality of the proposed method is slightly lower than that of the acoustic echo canceller.

On one hand, increasing the number of microphone elements improves speech recognition performance. However, the quality of the response sound degrades when the number of microphone elements is large. This is because the dimensions of nullspace decrease and the distance between the filter $\boldsymbol{b}(\omega)$ and the pulses $\boldsymbol{l}(\omega)$ increases. The proposed method has a trade-off between sound quality and speech recognition performance. We list the summary of the results as follows.

With one or two microphone elements: On the average, the difference in the sound quality of the proposed method from that of the acoustic echo canceller is small and within 1 dB. However, the improvement in speech recognition performance is small. Therefore, the merit of using the proposed method is not significant.

With three microphone elements: The deterioration of the CDs is slightly more than 1 dB, but this presents no problem in hearing. Since the improvement in speech recognition performance is large, we can say that the proposed method is beneficial for a spoken dialogue interface.

With four microphone elements: Although the improvement in speech recognition performance is significant, the deterioration of the CDs is nearly 2 dB and the user can perceive the degradation of sound quality.

5. Conclusion

We proposed a new small-scale barge-in free interface using sound field control to realize response sound cancellation. Since the proposed method does not control the sound around the user's position, the user is allowed to move. In addition, relaxation of the reproduction improves the robustness of the control. As results of the experiment, the robustness of sound elimination and the performance of speech recognition are improved when the number of loudspeakers is relatively small. It is also validated that the use of the proposed method together with an appropriate number of microphone elements does not degrade the quality of the

response sound. From these findings, the availability of the proposed method for spoken dialogue interface is ascertained.

Acknowledgement

This work was partly supported by the 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, April 2001.

[2] E. Hänsler, "Acoustic echo and noise control: Where do we come from—where do we go?," Proc. 7th International Workshop on Acoustic Echo and Noise Control, pp.1–4, Sept. 2001.

[3] S. Makino and S. Shimauchi, "Stereophonic acoustic echo cancellation—An overview and recent solutions," Proc. 1999 IEEE Workshop on Acoustic Echo and Noise Control, pp.12–19, 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. 7th International Conf. on Spoken Language Processing, vol.2, pp.773–776, Sept. 2002.

[5] H. Buchner, S. Spors, and W. Kellermann, "Wave-domain adaptive filtering: Acoustic echo cancellation for full-duplex system based on wave-field synthesis," Proc. International Conf. on Acoustics, Speech, and Signal Processing, vol.IV, pp.117–120, May 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. 2003 IEEE International Conf. on Acoustics, Speech, and Signal Processing, vol.V, pp.505–508, April 2003.

[7] 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.

[8] G. Strang, Linear Algebra and its Applications, Academic Press, New York, 1976.

[9] 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.

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

[11] 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.

[12] 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. 5th International Conf. on Spoken Language Processing, vol.7, pp.3261–3264, Dec. 1998.

[13] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A new phonetic tied-mixture model for efficient decoding," Proc. 2000 IEEE International Conf. on Acoustics, Speech, Signal Processing, vol.III, pp.1269–1272, June 2000.

[14] A. Lee, T. Kawahara, and K. Shikano, "Julius—An open source real-time large vocabulary recognition engine," Proc. 7th European Conf. on Speech Communication and Technology, vol.3, pp.1691–1694, Sept. 2001.

[15] S. Yamade, A. Lee, H. Saruwatari, and K. Shikano, "Unsupervised speaker adaptation based on HMM sufficient statistics in various noisy environments," Proc. 8th European Conf. on Speech Communication and Technology, pp.II-1493–1496, Sept. 2003.

[16] L. Rabiner and B. Juang, Fundamentals of speech recognition, Prentice Hall, Englewood Cliffs, NJ, 1993.

Appendix: Response Sound Elimination Error When Changing Room Transfer Functions

In this section we discuss the theoretical estimation of error caused by fluctuation in the proposed method, as we showed for the MOMNI method in Sect. 2.3. We assume that the fluctuation of the room transfer functions between the m-th loudspeaker and the k-th microphone, denoted by $\Delta g_{km}(\omega)$, are Gaussian random variables with the variance σ^2. Here, we define an M-dimensional orthonormal basis $\boldsymbol{b}_1(\omega), \ldots, \boldsymbol{b}_M(\omega)$ with its first vector $\boldsymbol{b}_1(\omega) = \boldsymbol{b}(\omega)/C$. Then $\Delta \boldsymbol{g}_k(\omega) = [g_{k1} \ldots g_{kM}]$ can be written as

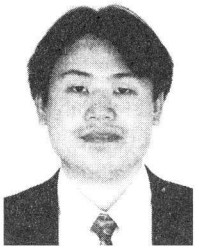
$$\Delta \boldsymbol{g}_k(\omega) = \sum_{m=1}^{M} \phi_{km}(\omega) \boldsymbol{b}_m^{\mathrm{H}}(\omega), \tag{A·1}$$

where $\phi_{km}(k = 1, \ldots, K, M = 1, \ldots, M)$ are random variables with variance σ^2. Then ϵ can be expressed as

$$\begin{aligned} \epsilon(\omega) &= \frac{1}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} A_k(\omega) \phi_{km}(\omega) \boldsymbol{b}_m^{\mathrm{H}}(\omega) \boldsymbol{b}(\omega) \\ &= \frac{C}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} \phi_{km}(\omega) \boldsymbol{b}_m^{\mathrm{H}}(\omega) \boldsymbol{b}_1(\omega) e^{-j\omega\tau_k} \\ &= \frac{C}{K} \sum_{k=1}^{K} \phi_{k1}(\omega) e^{-j\omega\tau_k} \\ &\propto \frac{1}{\sqrt{K}}. \end{aligned} \tag{A·2}$$

This reveals that its performance is influenced only by the number of microphones and no improvement can be obtained by increasing the number of loudspeakers. Therefore the MOMNI method performs better than the proposed method if many loudspeakers are available. However, if the number of loudspeakers is small, the control of the MOMNI method is inferior because of two reasons. The first reason is that MOMNI method can control 2 less microphones than the proposed method, as can be seen in Eqs. (1) and (14), because of the sound field reproduction at the user's ears. The second reason is that the condition in Eq. (13) fails to hold and its performance degrades. Since that condition is based on an assumption that the condition number of $\boldsymbol{H}(\omega)$ approaches 1, this assumption can hold only when the number of loudspeakers is much larger than the number of control points. Therefore, even a control of few microphones is unstable with a small number of loudspeakers because of the large condition number. From these findings, the proposed method performs better than the MOMNI method when the number of loudspeakers is small.

Shigeki Miyabe was born in Nara, Japan, on July 1, 1978. He received the B.E. degrees in electrical and electronics engineering from Kobe University in 2003, and received the M.E. degrees in information and science from Nara Institute of Science and Technology (NAIST) in 2005. He is now a Ph.D. student at Graduate School of Information Science, NAIST. His research interests include sound field control and array signal processing. He is a member of the the Acoustical Society of Japan.



Hiroshi Saruwatari was born in Nagoya, Japan, on July 27, 1967. He received the B.E., M.E. and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991, 1993 and 2000, respectively. He joined Intelligent Systems Laboratory, SECOM CO., LTD., Mitaka, Tokyo, Japan, in 1993, where he engaged in the research and development on the ultrasonic array system for the acoustic imaging. He is currently an associate professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include array signal processing, blind source separation, and sound field reproduction. He received the Paper Award from IEICE in 2000, and from TAF in 2004. He is a member of the IEEE, the VR Society of Japan, and the Acoustical Society of Japan.

Kiyohiro Shikano received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972, and 1980, respectively. He is currently a professor of Nara Institute of Science and Technology (NAIST), where he is directing speech and acoustics laboratory. His major research areas are speech recognition, multi-modal dialog system, speech enhancement, adaptive microphone array, Non-Audible Murmur recognition/synthesis, and acoustic field reproduction. From 1972, he had been working at NTT Laboratories, where he had been engaged in speech recognition research. During 1990–1993, he was the executive research scientist at NTT Human Interface Laboratories, where he supervised the research of speech recognition and speech coding. During 1986–1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories, where he was directing speech recognition and speech synthesis research. During 1984–1986, he was a visiting scientist in Carnegie Mellon University, where he was working on distance measures, speaker adaptation, and statistical language modeling. He received the Yonezawa Prize from IEICE in 1975, the Signal Processing Society 1990 Senior Award from IEEE in 1991, the Technical Development Award from ASJ in 1994, IPSJ Yamashita SIG Research Award in 2000, Paper Award from the Virtual Reality Society of Japan in 2001, and Paper Award and Inose Best Paper Award from IEICE in 2005. He is a fellow member of Information Processing Society of Japan. He is a member of the Acoustical Society of Japan (ASJ), Japan VR Society, the Institute of Electrical and Electronics, Engineers (IEEE), and International Speech Communication Association.

Yosuke Tatekura was born in Kyoto, Japan on May 17, 1975. He received the B.E. degrees in precision engineering from Osaka University in 1998, and received the M.E. and Ph.D. degrees in information science from Nara Institute of Science and Technology (NAIST) in 2000 and 2002, respectively. He is currently a research associate of Shizuoka University. His research interests include sound field control and virtual sound source synthesis. He is a member of the Acoustical Society of Japan, and the VR Society of Japan.