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

Tatsunori ASAI, Yoichi HINAMOTO, Hiroshi SARUWATARI, Kiyohiro SHIKANO

† Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, JAPAN
† Graduate School of Informatics, Kyoto University
Yoshida-Hommachi, Sakyo-ku, Kyoto 606-8501, JAPAN
{sawatari, shikano}@is.aist-nara.ac.jp

Abstract

This paper describes a new interface for a barge-in free spoken dialogue system combined an adaptive sound field control and a microphone array. It is essential for an acoustic echo canceller to estimate transfer functions and update the adaptive filter coefficient, especially in the case of the change of transfer functions due to the various interferences. In addition, the estimation process of transfer functions prevents a user from uttering freely and simultaneously during the response of a dialogue system. In order to actualize the robustness against the change of transfer functions, the barge-in free spoken dialogue system which using sound field control and microphone array has been proposed. However, this method cannot follow the 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 reduction accuracy of response sound is improved in comparison with the MOMNI method.

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 [1], [2]. However, the acoustic echo canceller is inherently vulnerable to the change of transfer functions in the barge-in situation. 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 [3] which combines sound field control and microphone array techniques. Although MOMNI method is robust against the change of transfer functions, there still exists the drawback that MOMNI method cannot adaptively follow the 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 in the experiment performed in the real room.

The outline of this paper is as follows. In Sect.2, we review the construction of MOMNI method. In Sect.3, we describe the principle of proposed spoken dialogue system. In Sect.4, we present the performance compared the proposed method and an echo canceller. Following a discussion of the experimental results, we give the conclusion in Sect.5.

2. Conventional MOMNI method [3]

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, $S_m$ for $m = 1, \cdots, M$ is the loudspeaker which acts as a secondary sound source, and $C_n$ for $n = 1, \cdots, N$ 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 $C_{K+2}$ for $K = N - 2$ are placed in each microphone element for recording user's speech. The intended signals to be reproduced at each control point are represented by

$$\mathbf{X}(\omega) = [X_R(\omega), X_L(\omega), X_{mic1}(\omega), \cdots, X_{micK}(\omega)]^\top,$$

(1)
Reproduced sound

Figure 1: Configuration of conventional MOMNI method.

where \( X_L(\omega), X_R(\omega) \) and \( X_{\text{mic}}(\omega) \) for \( k = 1, \cdots, K \) 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

\[
Y(\omega) = [Y_R(\omega), Y_L(\omega), Y_{\text{mic1}}(\omega), \cdots, Y_{\text{micK}}(\omega)]^T
\]

(2)

If the \( N \times M \) matrix composed of the room transfer function \( G_{nm}(\omega) \) for \( N < M \) between the secondary sound source \( S_m \) and the control point \( C_n \) is denoted by \( G(\omega) \), and the \( M \times N \) inverse filter matrix [4] is expressed as \( H(\omega) \), \( Y(\omega) \) is denoted by

\[
Y(\omega) = G(\omega)H(\omega)X(\omega),
\]

(3)

where \( G(\omega)H(\omega) = I_N(\omega) \), and \( I_N(\omega) \) is the \( N \times N \) identity matrix.

In Eq. (2), the response sounds of a dialogue system are reproduced at both ears of the user (\( [Y_R(\omega), Y_L(\omega)] = [X_L(\omega), X_R(\omega)] \) ) and silent zones are materialized at each microphone element (\( [Y_{\text{mic1}}(\omega), \cdots, Y_{\text{micK}}(\omega)] = [0, \cdots, 0] \) ). 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 \( A_k(\omega) \) for \( k = 1, 2, \cdots, K \) at \( k \)-th element of a delay-and-sum array is designated as

\[
A_k(\omega) = \frac{1}{K} e^{-j\omega \tau_k},
\]

(4)

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

\[
Y_{\text{mic}}(\omega) = \sum_{k=1}^{K} A_k(\omega) Y_{\text{mick}}(\omega).
\]

(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 \( H(\omega) \) by applying the least norm solution (LNS) in the frequency domain [5] so that the input signal \( X_n(\omega) \) is observed only at \( C_n \). In the case where the rank of \( H(\omega) \) is not decreased, since the solution of \( H(\omega) \) is indeterminate, we adopt the Moore-Penrose generalized inverse matrix as the inverse filter which gives the LNS [3].

### 2.4. Response sound elimination error when changing room transfer functions

MOMNI method which uses fixed inverse filter coefficients is proved to be robust against the change of room transfer functions [3]. 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 \( Y'(\omega) \) denotes

\[
Y'(\omega) = (G(\omega) + \Delta G(\omega))H(\omega)X(\omega),
\]

(6)

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

\[
\Delta Y_{\text{mic}}(\omega) = \sum_{k=1}^{K} A_k(\omega) \left\{ \sum_{m=1}^{M} \Delta G_{(k+2)m}(\omega) \right\} (H_{m1}(\omega)X_{R}(\omega) + H_{m2}(\omega)X_{L}(\omega)).
\]

(7)

In the report in [3], the following relation holds in the elimination error of response sound, \( \varepsilon(\omega) \), as

\[
\varepsilon(\omega) = \Delta Y_{\text{mic}}(\omega) \propto \frac{1}{\sqrt{M \cdot K}}.
\]

(8)

Equation (8) 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 technique for response sound elimination

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

#### 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 \( G(\omega) \) of estimated transfer function is set to \( G(\omega) \).
[step 1] In the case where the fluctuation of transfer function $\Delta G_{\text{nm}}(\omega)$ is added in the transfer function $G_{\text{nm}}(\omega)$ because of the change of a transfer system, the changed transfer function $G'(\omega)$ becomes

$$G'(\omega) = G(\omega) + \Delta G(\omega),$$

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

$$Y'(\omega) = G'(\omega)H(\omega)X(\omega).$$

The estimated transfer function $\hat{G}(\omega)$ that minimizes the squared error between observation signal $Y'(\omega)$ and estimated signal $\hat{Y}'(\omega)$ at the control points is sought. $E(\omega)$ is defined as an error signal, and

$$E[i-1](\omega) = Y'(\omega) - \hat{Y}'[i-1](\omega) = (G'(\omega) - \hat{G}_{i-1}'(\omega))H(\omega)X(\omega),$$

where $i$ is the number of iterations. From Eq. (11), the partial differentiation of the squared error $||E[i-1](\omega)||^2$ with respect to $\hat{G}_{i-1}'(\omega)$ is given by

$$\frac{\partial}{\partial \hat{G}_{i-1}'(\omega)} ||E[i-1](\omega)||^2 = -E[i-1](\omega)(H(\omega)X(\omega))^H.$$

Thus, the modification amount of $\hat{G}_{i-1}'(\omega)$ using normalized least-mean-squares (NLMS) method is denoted as

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

where $\alpha$ for $0 < \alpha < 2$ is a step-size parameter, and $\beta$ is a minimal positive constant to be non-zero in the denominator term on the right-hand side of Eq. (13).

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

$$\hat{G}_i(\omega) = \hat{G}[i-1]'(\omega) + \Delta \hat{G}_{i-1}'(\omega).$$

[step 2] If $\hat{G}_{i-1}'(\omega)$ derived from Eq. (14) 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 $\hat{H}(\omega)$ based on $\hat{G}_{i-1}'(\omega)$ via LNS.

### 4. Experiment and result

In this section, we describe an experimental result comparing the acoustic echo canceller and proposed method. In order to verify the applicability of the proposed method, we simulate adaptation process based on the change of transfer functions and evaluate the convergence of a squared error and response sound elimination.

#### 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 response three times; two patterns are the state where the interference is allocated, and the other pattern is the state where the interference does not exist.

The impulse responses used in this experiment are measured in an acoustic experiment room, where the reverberation time is about 140 ms, with 48 kHz sampling and 16-bit resolution. Figure 3 shows the arrangement of
4000 Hz. As the response sound of a dialogue system,
we estimate the changed transfer functions and design in­
verse Dlters. In the estimation, we use the response sound
that the apparatuses. The primary sound source is the loud­
speaker used as the spoken dialogue system in the acous­
tic echo canceller. We use a circular microphone array
with three elements which are equally spaced. In the de­
sign condition of the inverse Dlters, the number of sec­
ondary sound sources, M, is 8, the number of control
points, N, is 3 or 4, and the passband range is 150–
4000 Hz. As the response sound of a dialogue system,
we use a female sound selected from the ASJ database.

In the case where the interference shifts to Position 1,
we estimate the changed transfer functions and design in­
verse Dlters. In the estimation, we use the response sound
cut every one second which has adequate time length.
The step-size parameter \( \alpha \) in Eq. (13) is 0.1, which is
optimized experimentally, and \( \beta \) is \( 1.0 \times 10^{-6} \).

Finally, it is assumed that the interference moves from
Position 1 to Position 2 in the state of barge-in, we apply
the MOMNI method and stop estimating.

4.2. Evaluation of convergence in adaptation process
Figure 4 illustrates the relationship between the number of
iterations and the squared error of estimation in the case
where \( \alpha = 0.1 \), \( N = 4 \) and the number of iterations
is 1000. The squared error decreases as the number of
iterations increases, and approximately converges at 100
iterations. Since ERLE is 105 dB in the instance of 100
iterations, the proposed adaptation process can estimate
transfer functions completely.

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

\[
ERLE = 10 \log_{10} \left( \frac{\sum \{ Y_{\text{micref}}(\omega) \}^2}{\sum \{ E(\omega) \}^2} \right),
\]

where \( Y_{\text{micref}}(\omega) \) is the response sound reproduced at a
critical microphone we assign, and \( E(\omega) \) is the error signal
derived from Eq. (8).

Figure 5 shows the ERLEs which are with adaptation
process and without adaptation under the condition in
which \( \alpha = 0.1 \) and \( N = 3 \) or 4. We estimate the changed
transfer functions when the interference enters Position

1. From Fig. 5, we can see that the proposed method has
two advantages. The first is to enable us to design inverse
Dlters in the case where the changed transfer functions are
unknown. The second is to sustain the better performance
of response sound elimination than a conventional
acoustic echo canceller as well as MOMNI method which
does not adapt the change of transfer functions.

5. Conclusion
We proposed the spoken dialogue system based on adaptive
sound field 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
method is ascertained.

6. Acknowledgement
This research is partly supported by Core Research for
Evolutive Science and Technology (CREST) Program
"Advanced Media Technology for Everyday Living" of
Japan Science and Technology Agency (JST).

7. References
Communication, pp.5–10, Kyoto, Japan, Apr. 2001
time Acoustic Human-Machine Front-end for Multimedia Appli­
cations Integrating Robust Adaptive Beamforming and Stere­
776, Colorado, USA, Sep. 2002
for Barge-in Free Spoken Dialogue System Based on Sound Field
508, Apr. 2003
[4] M. Miyoshi, Yutaka, Kaneda, "Inverse Filtering of Room Acous­
ing, vol.36, no.2, pp.145–152, 1988
Design Method for the Multichannel Sound Field Reproduction