INTERFACE FOR BARGE-IN FREE SPOKEN DIALOGUE SYSTEM BASED ON SOUND FIELD CONTROL AND MICROPHONE ARRAY
Yoichi Hinamoto † Kouichi Mino † Hiroshi Saruwatari † Kiyohiro Shikano †
† Nara Institute of Science and Technology 8916-5 Takayama-cho, Ikoma, Nara 630-0101, Japan
Email: {yoichi-h, sawatari, shikano}@is.aist-nara.ac.jp

ABSTRACT
In this paper, a barge-in free spoken dialogue system using sound field control and microphone array is proposed. In the conventional spoken dialogue system using an acoustic echo canceller, it is indispensable to estimate and update the room transfer function, especially when the transfer function is changed by various interferences. However, the estimation process for the transfer function prevents the user from speaking freely and simultaneously with speech responses from the system. In order to resolve the problem, we have already proposed a barge-in free spoken dialogue system that controls a sound field using multiple loudspeakers. In this paper, a microphone array for acquisition of user’s speech is newly introduced in the previously proposed system. By introducing the microphone array, we can reduce the number of loudspeakers to be required in the system, and make the interface for spoken dialogue system more robust against the change of room transfer functions.

1. INTRODUCTION
In a spoken dialogue system between man and machine, it is important to clearly input speech, a way for a user to communicate information, into the system for the sake of maintaining the speech recognition accuracy. However, a user might start speaking before the termination of sound responses from the system. Such a situation is called barge-in [1]. In the state of barge-in, if the sound and music to be given from the system to the user are inputted into a microphone for acquisition of user’s speech as an acoustic echo, the performance of the system to recognize user’s speech is degraded. Hereafter, the acoustic echo is called barge-in sound.

In general, we can reduce the barge-in sound by using an acoustic echo canceller. However, when the room transfer function is changed by various interferences in the state of barge-in, the speech recognition accuracy is degraded due to the errors between the estimated filter-coefficients in the echo canceller and the transfer function.

On the other hand, several techniques have been reported for eliminating the barge-in sound, such as the acoustic echo cancellers embedded into a sidetone canceller [2], a stereophonic acoustic echo canceller using an affine projection algorithm with consideration of human’s auditory characteristics [3]. However, there is no argument in [2],[3] on the robustness for the change of transfer functions.

In order to achieve the robustness, we have proposed a barge-in free spoken dialogue system which controls a sound field by using multiple loudspeakers [4]. However, this technique requires too many loudspeakers to improve the robustness for the change of transfer functions. In the technique reported in [4], sound responses from the system are outputted by multiple loudspeakers, and the system controls a sound field without inputting the responses into a single microphone for acquisition of user’s speech. Hereafter, the technique in [4] is called the MONI (Multiple Output and No Input) method.

To resolve the above problem, we newly propose the interface for a barge-in free spoken dialogue system, which is more robust for the change of room transfer functions. This is achieved by using sound field control and microphone array to reduce the number of loudspeakers to be required in the MONI method.

2. PROPOSED TECHNIQUE FOR BARGE-IN SOUND ELIMINATION
The configuration of the proposed system is shown in Fig. 1. In this system, if the number of elements in a microphone array is set to be one, that is, if microphone array processing is not performed, this system coincides with the MONI method. In this section, we discuss the principle of the proposed method by dividing two parts: sound field control and signal processing using a microphone array.

2.1. Sound field control
First, the principle of an interface for a spoken dialogue system using sound field control is discussed. Let $M$ be the number of secondary sound sources $(S_1, S_2, \ldots, S_M)$ and let $N$ be the number of control points $(C_1, C_2, \ldots, C_N)$. In the control points, $C_1$ and $C_2$ are arranged to both ears of a user, and $C_3, C_4, \ldots, C_{K+2}$ ($K = N - 2$) are arranged to the microphone array elements for acquisition of user’s speech. The signals to be reproduced to each control point are described by...
filter

Inverse system design for sound field control

Therefore, we design an inverse frequency domain [5]. The method has advantages that the amount of calculations are small, and the designed system is

2.2. Signal processing using microphone array

where \( Y_R(\omega) \) and \( Y_L(\omega) \) are signals reproduced at the right and left ears of a user, respectively. The transfer function and the observed signals at each control point are represented by

\[
\mathbf{Y}(\omega) = [\mathbf{Y}_n(\omega), \mathbf{Y}_l(\omega), \mathbf{Y}_{mic1}(\omega), \ldots, \mathbf{Y}_{micK}(\omega)]^T,
\]

and the observed signals at each control point are represented by

\[
\mathbf{\hat{Y}}(\omega) = [\mathbf{\hat{Y}}_n(\omega), \mathbf{\hat{Y}}_l(\omega), \mathbf{\hat{Y}}_{mic1}(\omega), \ldots, \mathbf{\hat{Y}}_{micK}(\omega)]^T,
\]

where \( Y_n(\omega) \) and \( Y_l(\omega) \) are signals reproduced at the right and left ears of a user, respectively. The transfer function from the secondary sound source \( S_m \) to the control point \( C_n \) is denoted by \( G_{nm}(\omega) \) where \( n = 1, 2, \ldots, N \), \( m = 1, 2, \ldots, M \) and \( N < M \). Let \( G(\omega) \) be an \( N \times M \) matrix consisting of \( G_{nm}(\omega) \), and let \( H(\omega) \) be its \( M \times N \) inverse filter. Then, the relation between \( \mathbf{\hat{Y}}(\omega) \) and \( \mathbf{Y}(\omega) \) is expressed as

\[
\mathbf{\hat{Y}}(\omega) = G(\omega)H(\omega)\mathbf{Y}(\omega),
\]

where \( G(\omega) \) is an \( N \times N \) identity matrix.

In Eq. (3), if the response sounds \( Y_n(\omega) \) and \( Y_l(\omega) \) from the system are reproduced at both ears of the user, and if \([\mathbf{Y}_{mic1}(\omega), \ldots, \mathbf{Y}_{micK}(\omega)] = [0, 0] \) is reproduced at each position of the microphone array elements, then the sound field can be realized without barge-in sound at each of the microphone array elements, while the dialogue system gives the response sounds to the user.

2.2. Signal processing using microphone array

In Subsection 2.1, sound field control is performed at each of \( K \) elements in the microphone array for acquisition of user's speech as a part of control points. In this subsection, we pay an attention to array signal processing. In this study, we adopt the delay-and-sum type for array signal processing. This is a general way for array signal processing with small amounts of operations. The filter of \( k \)-th element in a delay-and-sum array is denoted by \( W_k(\omega) \) for \( k = 1, 2, \ldots, K \). Then \( W_k(\omega) \) can be expressed as

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

where \( \tau_k \) stands for the arrival time difference of the target signal between a suitable array origin and \( k \)-th element position. Suppose that the signal added through the array filters is a signal for the speech recognition. Then the eliminated barge-in sound in the speech recognition signal is expressed as

\[
\mathbf{\hat{Y}}_{mic}(\omega) = \sum_{k=1}^{K} W_k(\omega) \mathbf{Y}_{mic}(\omega).
\]

When this delay-and-sum array is used, the system response sounds which come from other than the target direction are out of phase at each element, and only the user's speech which comes from the target direction is in phase at each element and added. In result, only the user's speech can be emphasized.

2.3. Inverse system design for sound field control

In a multi-point control system which controls plural control points by many sound sources, a large amount of calculations and memories are needed to design an inverse filter in the time domain. Therefore, we design an inverse filter \( H(\omega) \) by using the least-norm solution (LNS) in the frequency domain [5]. The method has advantages that the amount of calculations are small, and the designed system is

stable because the output from each sound source is reduced to be minimum.

Since the solution of an inverse matrix of \( G(\omega) \) is not unique if the rank of \( H(\omega) \) is not reduced, the Moore-Penrose type generalized inverse matrix is used as an inverse matrix which gives the least-norm solution. For this reason, the singular value decomposition of \( G(\omega) \) is performed.

\[
G(\omega) = U(\omega) \cdot \big[ \Gamma_N(\omega), \mathbf{0}_{N \times (M-N)} \big] \cdot V^H(\omega),
\]

where \( U(\omega) \) and \( V(\omega) \) are \( N \times N \) and \( M \times M \) orthogonal matrices, respectively, \( \Gamma_m(\omega) \) for \( m = 1, 2, \ldots, N \) are the singular values of \( G(\omega) \) and are arranged so that \( \Gamma_m(\omega) \geq \Gamma_{m+1}(\omega) \) in matrix \( \Gamma_m(\omega) \). \( O_{N \times (M-N)} \) denotes an \( N \times (M-N) \) null matrix. \( V^H(\omega) \) denotes the Hermitian transposed matrix of \( V(\omega) \).

Then the Moore-Penrose type generalized inverse matrix \( G^\dagger(\omega) \) is expressed as

\[
G^\dagger(\omega) = V(\omega) \cdot \left[ \frac{\Lambda_N(\omega)}{O_{M-N \times N}} \right] \cdot U^H(\omega),
\]

where

\[
\Lambda_N(\omega) = \text{diag} \left[ \frac{1}{\mu_1(\omega)}, \frac{1}{\mu_2(\omega)}, \ldots, \frac{1}{\mu_N(\omega)} \right].
\]

Therefore, the inverse filter \( H(\omega) \) can be designed by calculating an inverse matrix which gives the least-norm solution for every frequency.

2.4. Barge-in sound elimination error when changing the room transfer functions

The following arguments will be valid when the change of the transfer function is independent for every channel between loudspeaker and microphone. Suppose that the variation \( \Delta G_{nm}(\omega) \) caused by the change of room transfer functions is added to the original transfer function \( G_{nm}(\omega) \). The elimination error of barge-in sound, \( \Delta \mathbf{Y}_{mic}(\omega) \) observed in \( \mathbf{Y}_{mic}(\omega) \), is then expressed as

\[
\Delta \mathbf{Y}_{mic}(\omega) = \sum_{k=1}^{K} W_k(\omega) \left\{ \sum_{m=1}^{M} \Delta G(k+2)m(\omega) \right\} \left\{ H_{m1}(\omega)Y_n(\omega) + H_{m2}(\omega)Y_l(\omega) \right\}. \tag{10}
\]

Denoting the matrix norm of \( H(\omega) \) by \( \|H(\omega)\| \), we can write Eq. (10) as

\[
\Delta \mathbf{Y}_{mic}(\omega) = \|H(\omega)\| \frac{1}{K} \left\{ \sum_{k=1}^{K} \sum_{m=1}^{M} \Delta G(k+2)m(\omega) \right\} \left\{ H_{m1}(\omega)Y_n(\omega) + H_{m2}(\omega)Y_l(\omega) \right\} e^{-j\omega \tau_k}. \tag{11}
\]

where \( H_{m1}(\omega) = H_{mn}(\omega)/\|H(\omega)\| \). It is assumed that \( \Delta G_{nm}(\omega) \) is the Gaussian random variable with the variance \( \sigma^2 \). Furthermore, since \( H_{m1}(\omega) \) is normalized by \( \|H(\omega)\| \), and is independent on the change of \( M \), the deviation in \( \{\} \) of Eq. (11) can be expressed as \( \eta M \cdot K \sigma \), where \( \eta \) is a suitable constant. Also, \( \|H(\omega)\| \) is proportional to \( 1/M \) because the followings hold in the case of
\[ \mu_1 \approx \mu_N \quad [4], \quad \|H(\omega)\| = \|G^*(\omega)\| = 1/\|\mu_N(\omega)\| = 1/\|\mu_1(\omega)\| = 1/\|G(\omega)\| = 1/\sqrt{MK}. \]

Therefore, the elimination error \( E(\omega) \) of barge-in sound is expressed as

\[ E(\omega) = \Delta Y_{\text{mic}}(\omega) \propto 1/M \cdot 1/K \cdot \sqrt{MK}. \] (12)

In other words, Eq. (12) shows that the elimination error of barge-in sound at the change of transfer functions is inversely proportional to \( \sqrt{MK} \). Thus, if the number of transfer channels from loudspeakers to microphones increases, the proposed barge-in sound elimination method is more robust against the change of transfer functions.

It is remarked that it is difficult in the real environment to prove whether or not the variation \( G_{nm}(\omega) \) caused by the change of room transfer functions is independent for every channel between loudspeakers and microphones. However, the simulation results using real environmental data described in the next section show that the error estimation computed using Eq. (12) is acceptable.

3. SIMULATIONS AND DISCUSSIONS

3.1. Experimental comparison of barge-in sound elimination performance

The barge-in sound elimination performance of the proposed method is evaluated through the simulations which are carried out by using the measured room impulse responses in order to investigate the robustness against the change of room transfer functions. For comparison, the performance of the conventional acoustic echo canceller is also computed. In the experiment, the interference, i.e., a life-size manikin is arranged near a user under the assumption that a person obtains access to the user so as to cause the change of the room transfer functions. In addition, since this experiment is interested in the robustness against the change of room transfer functions, the processing of eliminating the barge-in sound is performed by using the filter coefficients before updating after the room transfer functions change. The barge-in sound elimination performance is evaluated by using ERLE (Echo Return Loss Enhancement) in Eq. (13).

\[ \text{ERLE} = 10 \log_{10} \left( \sum \frac{\{Y_{\text{mic}}(\omega)\}^2}{\sum \{E(\omega)\}^2} \right), \] (13)

where \( Y_{\text{mic}}(\omega) \) is the signal reproduced the system response sound at a standard microphone and \( E(\omega) \) is the error signal in the barge-in sound elimination.

The arrangement of the acoustic experiment room is shown in Fig. 2 where reverberation time is about 200 ms. The room impulse responses are measured in the experiment room where the primary sound source is a loudspeaker used for a spoken dialogue system in the acoustic echo canceller. The room impulse responses are sampled with frequency 48 kHz and the magnitudes are quantized to 16 bit. In the experiment, we use a circle array with 6 elements and one microphone to six microphones by our choice. As for the spoken dialogue system sounds, we use the sound from 6 men and 6 women, totally 12, through the ASJ database. ERLE is computed by the average value of the results from the above 12 people.

In Fig. 3, ERLE is shown for different number of the room transfer channels from loudspeakers to microphone array elements.

a. Conventional acoustic echo canceller

In the experiment, an acoustic echo canceller with fixed filter coefficients is constructed without using a specific adaptive algorithm in the echo canceller. The experiment is carried out under the assumption that when the transfer functions are in time-invariant, and the filter coefficients are estimated accurately.

b. Proposed method

The inverse filter in the proposed method is calculated only using the impulse responses in case there are no transfer functions change. The inverse filter are designed by the technique of using the least-norm solution (LNS) [5]. The design conditions of the inverse filters are such as the number of secondary sound sources \( M = 4 \) to 36, the number of control points \( N = 3 \) to 8, filter length 32768, and passband range 150 to 4000 Hz.

In Fig. 3, ERLE is shown for the number of transfer channels \( (\approx M \cdot K) \) from loudspeakers to microphone. The theoretical curve in the figure is obtained by plotting the ERLE derived from Eq. (12), which is given by

\[ \text{ERLE}_{\text{theory}} = 10 \log_{10} \left( \sum \frac{\{Y_{\text{mic}}(\omega)\}^2}{\sum \{E(\omega)\}^2} \right) \]
The results of word accuracy are shown in Table 1. The recognition results show that when changing the transfer functions, the degradation of speech recognition accuracy can be suppressed by increasing the number of transfer channels from loudspeaker to microphone.

4. CONCLUSION

It has been shown that in the system with sound field control and microphone array, the performance to the change of room transfer functions is dependent on the number of transfer channels from loudspeaker to microphone and the stability of the inverse filter. The proposed method has shown the better performance than the conventional acoustic echo canceller. Moreover, by increasing the number of array elements, the number of loudspeakers to be required in MONI method has been reduced, and the interface for a barge-in free spoken dialogue system has become more robust against the change of room transfer functions.

ACKNOWLEDGEMENTS

We would like to thank Dr. Shoji Makino for his detailed discussions.

REFERENCES