NAIST-IS-DD0561031

Doctoral Dissertation

Barge-in Robust Spoken Dialogue Interface Using Multichannel Sound Field Control and Array Signal Processing

Shigeki Miyabe

September 30, 2007

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Shigeki Miyabe

Thesis Committee:

Professor Kiyohiro Shikano	(Supervisor)
Professor Hirokazu Nishitani	(Co-supervisor)
Associate Professor Hiroshi Saruwatari	(Co-superviosr)

Barge-in Robust Spoken Dialogue Interface Using Multichannel Sound Field Control and Array Signal Processing*

Shigeki Miyabe

Abstract

A spoken dialogue system is demanded as a user-friendly human-machine interface that does not require any special skills in its manipulation. Speech has advantageous features: they are hands-free and eyes-free, i.e., one can use speech while doing other tasks. For effective utilization of the features, it is desirable that the system can be used even when the user stands away from the microphone or the user's speech is uttered interrupting the output sound of the system (response sound). The problem in satisfying such demands is the degradation of automatic speech recognition (ASR) because of feedback of response sound and observation of interfering noise due to other sound than the user's speech. Since current ASR systems are sensitive to noise, a noise reduction method is indispensable.

In elimination of the response sound and the interfering noise, an acoustic echo canceller (AEC) and an adaptive beamformer (ABF) are generally used, respectively. In each of the methods, a filter is adapted to eliminate its target noise based on the minimum-mean-squared-error criterion. Thus, when their filters are trained using signals containing sources other than their target noise, their performances degrade severely. To prevent such degradation, the system should detect the times when the observed signals contain sounds other than the target noise, denoted as double-talk detection (DTD). However, accurate DTD is difficult, particularly in such a situation that both response sound and interfering

*Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0561031, September 30, 2007.

noise exist. In this dissertation, I propose a new framework for eliminating both the response sound and the interfering noise without DTD by (1) response sound elimination based on fixed-filter-based control with robust structure against the fluctuation of the acoustical transfer system, and (2) filter adaptation for the elimination of both the response sound and the interfering noise without DTD.

The robustness of the response sound elimination against the fluctuation of the transfer system can be improved by increasing the number of transfer channels in the structure of the elimination. To realize this, I propose a new mechanism to eliminate the response sound using sound field control with multiple loudspeakers which cancels the response sound at microphones. With this structure, the effect of changes in the transfer system from the state of the measurement is dispersed to multiple channels. As a means of sound field control for the elimination, I propose two methods with and without high-quality reproduction of the response sound. While the former can improve the robustness by increasing the number of loudspeakers, the latter can achieve high robustness using fewer loudspeakers than the former. The former using 24 loudspeakers and the latter using five loudspeakers performed 20% and 15% better than the conventional AEC in word accuracy in a speech recognition experiment involving a dictation task.

Blind source separation (BSS) has been developed over a last few decades as an unsupervised method of training a beamformer to separate unknown sound sources. In our problem, the source of the response sound is known. In this dissertation, I propose a new semiblind source separation by modifying the structure of BSS to deal with a semisupervised problem, and combine the semiblind source separation with the sound field control. With this combination, semiblind source separation enforces the speech enhancement of the sound field control by eliminating both the residual response sound caused by the fluctuation and the interfering noise. As a result of a comparison between the performance of the proposed method and the performance limit of the combination of AEC and ABF, the proposed method was found to perform about 10% better in word accuracy.

Keywords:

Hands-free speech recognition, multichannel sound field control, blind source separation, independent component analysis, acoustic echo canceller.

マルチチャネル音場制御とアレー信号処理を用いた割り込み発話に頑健な音声対話インタフェース*

宮部 滋樹

内容梗概

音声対話システムは，操作に特別な技術を必要としないマンマシン・インタフェースとして期待されている．このシステムでは，ほかの作業をしながらでも利用できるハンズフリー・アイズフリーという音声の長所を生かすため，ユーザがマイクロホンから離れている場合や，ユーザ音声がシステム自身が発する応答音に割り込んで入力された場合でも入力音声を受理できるのが望ましい．このような要求を満たすにあたって問題となるのは，システム出力音 (応答音) のマイクロホンへの回り込みと，システムを使用する環境のユーザ音声以外の音に起因する外部雑音がユーザ音声に混入して観測されることによる，音声認識精度の劣化である．現状の音声認識技術は雑音に対して脆弱であり，雑音除去の仕組みが必要不可欠である．

応答音と外部雑音の除去には，それぞれ音響エコーキャンセラとマイクロホンアレーによる適応ビームフォーマを用いるのが一般的である．これらはそれぞれの対象となる雑音を除去するために誤差最小化基準を用いてフィルタの適応を行う．そのため，除去対象以外の音を含んだ観測信号で学習すると，その精度が著しく低下してしまう．これを防ぐため，対象雑音以外の音が鳴っている時間区間を検出するダブルトーク検出という仕組みが必要となる．しかしダブルトークの完全な検出は難しく，特に応答音と外部雑音の両方が存在するような騒音下では高い精度の検出は期待できない．そこで本論文では，(1) 音響伝達系変動に対して頑健な構造を持つ固定フィルタを用いた応答音除去と，(2) 応答音と外部雑音を除去する雑音除去のダブルトーク検出不要な適応に基づく，ハンズフリー音声対話のための新しい枠組みを提案する．

*奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文，NAIST-IS-DD0561031, 2007年9月30日.

伝達系変動に対する頑健性の向上は，応答音除去の仕組みを複数の伝達経路を持つ構造にすることで実現できる．これを実現するために，複数のラウドスピーカを用いてマイクロホンの位置で応答音を相殺させる音場制御を用いた新しい応答音除去の仕組みを提案する．このようなマルチチャネル構造により，伝達系の測定時からの変化による誤差は複数パスに分散される．応答音を消すための音場制御としては，応答音の高品質再現を行うものと，再現精度を緩めた 2 つの手法を提案する．前者はラウドスピーカ数を増やすことにより高い頑健性を得ることができ，ディクテーションタスクの音声認識実験では，24 個のラウドスピーカを用いることで従来型音響エコーキャンセラよりも 20%高い単語正解精度が得られた．また，後者は前者に比べてやや音質が劣るものの，少ないスピーカ数でも安定な制御を実現することができ，実験では 5 個のラウドスピーカを用いて従来型音響エコーキャンセラよりも 15%高い単語正解精度が得られた．

未知の音源に対する教師なしのビームフォーマの学習則として，近年ブラインド音源分離 (BSS) の研究が進んでいる．本研究で扱う問題では，応答音の音源はシステムにとって既知である．そこで，BSS を既知である応答音の情報を利用した半教師なし構造に拡張することにより，BSS よりも効率的に音源分離を行うセミブラインド音源分離を提案し，音場制御による応答音除去と組み合わせる．このような組み合わせにより，セミブラインド音源分離は応答音の消し残りと外部雑音を除去し，音場制御による音声強調を強化する．提案手法の性能を音響エコーキャンセラと適応ビームフォーマの併用の性能限界と比較する実験の結果，単語正解精度の約 10%の向上が得られた．

キーワード

ハンズフリー音声認識, マルチチャネル音場制御, ブラインド音源分離, 独立成分分析, 音響エコーキャンセラ

Contents

List of Figures

List of Tables

1. Introduction

1.1 Background

Speech is the most basic means of human communication. As a user-friendly human-machine interface, there has been an increasing demand for a spoken dialogue system, where speech is used as the means of both command input and message output. As a result of the marked improvement of automatic speech recognition (ASR) in recent decades, speech interaction with machines is no longer a fantasy. In fact, there have been several implementations, e.g., speech dialogue systems with robots [KYHS97], flight reservation systems [SP00], guidance systems at community centres [CNLS07], and call systems [SHL+98, ZSG+00]. However, in all such implementations, the vocabulary, grammar, and topics are limited, and they do not represent speech interaction in the true sense. There are many problems in realizing an intuitive, unconstrained, and stress-free speech interaction between human and machine.

The problems concerning the spoken dialogue system are listed below.

(1) Hands-free speech input: For the sake of eyes-free manipulation, the system should receive the user's speech uttered far from the microphone. In addition, to free the user from physical constraint, it is desirable that the user be able to input their commands without special equipment such as head-set microphones.

(2) linguistic phenomena inherent to free spoken language: Current ASR systems cannot deal sufficiently with the unnecessary words that appear only in spoken language, such as hesitation or rephrasing.

(3) Dialogue control: Control of the flow of conversation should be improved to accept user's demands smoothly.

(4) Quality of synthesized speech: For the generation of the sound response to the user, text-to-speech (TTS) is desirable because of its flexibility. However, TTS is generally inferior to recorded speech in its quality and sometimes hardly understandable.

(5) Real-time processing: To realize smooth interaction, all processes should be completed in real time.

(6) Free-timing input: To remove inconvenience, it is desirable that the user's speech utterance be accepted any time, even when the system is outputting its response.

1.1.1 Research purpose

In this research, I focus on (1) hands-free speech input and (6) Free-timing input.

Hands-free speech input is an important issue related to the significance of spoken dialogue systems. In the conventional manipulation of machines, input is generally achieved by manipulating buttons or keyboards, while the output is displayed on LEDs or monitor screens. Spoken-dialogue-based manipulation has the advantage of freeing the user's hands and eyes. Accordingly, the user is allowed to do other tasks than device manipulation. Thus close-talking or headset microphones that constrain the user's pose or position are undesirable. However, in such a situation where the microphone is set away from the user, the microphone often picks up interfering noise with relatively high energy, caused by, for example, air-conditioners or speech of people other than the user. Since ASR is weak against noise, a noise reduction method to enhance the speech is indispensable for realizing robust ASR in the form of hands-free input. In this paper, such noise in a room is denoted as *interfering noise*.

Free-timing input is important to develop natural usability. In conversation among people, a person often interrupts another person's speech. This is a necessary action for a person to process the conversation rapidly or to acquire initiative in the conversation. In a spoken dialogue system, such interruption corresponds to the user's speech commands neglecting the system's responses, or an unintentional start of the speech command in the end of the system's response. The system should be able to accept such speech utterances. Such an interruptive speech utterance is refered to as *barge-in* [JS01], as depicted in Fig. 1. However, to accept barge-in in a hands-free system, the sound message of the system is fed back to the microphone and acts as noise in ASR. In the following, such feedback is denoted as *response sound*.

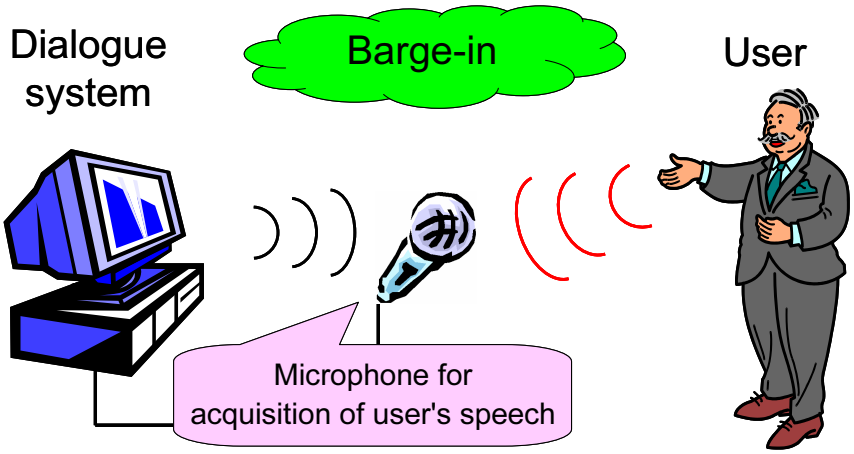

Figure 1. Situation of barge-in in a spoken dialogue system.

In this study, the goal is to realize a barge-in- and interfering-noise-robust interface for a hands-free spoken dialogue system by enhancing the user's speech and eliminating the response sound and the interfering noise.

1.1.2 Related works

To eliminate the response sound from the system, an acoustic echo canceller (AEC) is commonly used. Many types of AECs have been proposed, such as single channel, stereophonic [Han01, SJH01], wave-synthesis [BBK05, BSK04], and beamformer-integrated types [HBK03, HK02, Her04]. Typical echo cancellers are listed in Table 1, and their adaptation methods are listed in Table 2. However, the AEC has the inherent problem that accurate adaptation is difficult to work accurately and sometimes diverges in duration when both the user and the system emit sound simultaneously (this is also referred to as *double-talk*). Because of this problem, the conventional AEC should detect the double-talk duration and stop adaptation to optimize filter coefficients only in the single-talk of the system; this implies that the elimination performance is likely to degrade when a change in a room transfer function arises during double-talk. Although there is much research on the double-talk detection (DTD) [YW91, GB01, GB06], in general, DTD is quite difficult in noisy environments. As another approach to achieving

Table 1. Types of echo cancellers

Echo suppressor [SB80]	Cancellation of echo in telephony caused by impedance mismatch by simple switching on/off of lines.
Echo canceller [Son67]	Cancellation of echo in telephony with adaptive filter to estimate accurate echo component.
Acoustic echo canceller (AEC) [CSO72, KIS$^+$73]	Echo canceller for acoustic echo caused by reverberation in hands-free speech input.
AEC with two echo path model [AOO77]	Double-talk robust AEC with dual filter structure.
Stereo AEC [MS99]	AEC for stereo loudspeakers.
Wave-domain adaptive filtering [BSK04]	AEC for wave synthesis with multiple loudspeakers.
Integration with adaptive beamformer [HBK03]	Simultaneous adaptation of AEC and adaptive beamformer to eliminate both acoustic echo and interfering noise.

robustness against double-talk, some researchers have proposed AECs with dual filters, but they are equivalent to DTDs based on fluctuation of adaptation.

The problem of the DTD cannot be avoided even by combining an AEC with some noise reduction methods such as an adaptive beamformer (ABF) [Fro72, GJ82], because adaptation of beamformer also requires the detection of the single-talk duration of the interfering noise. Typical beamformers are listed in Table 3. Furthermore, many approaches have been adopted to avoid the divergence of adaptation in a double-talk situation [AOO77, BBGK03, BDH$^+$99]. However, those methods are intended only to prevent divergence and cannot improve the echo cancellation performance in the double-talk duration.

1.2 Contribution

In this dissertation, I propose a new framework of a spoken dialogue interface without DTDs comprising a combination of sound field control and source sepa-

Table 2. Typical filter adaptations for AEC

LMS [WH60]	The most basic on-line adaptation to update filter in each of input samples.
RLS [LMF78]	Adaptation by shifting block of training samples with certain length. More robust and faster convergence than LMS.
Subband LMS [Fur84, Kel84]	LMS with faster convergence by subband analysis of training samples to reduce autocorrelation.

ration.

Response sound elimination using sound field reproduction

DTDs are required in AECs to follow the fluctuation of the acoustical transfer system, and the performance of AECs degrades when the adaptation cannot follow the fluctuation. The cause of the degradation is the weakness of AECs against fluctuations. With a robust elimination mechanism of the response sound, its filter is not required to adapt in real time. The robustness can be obtained by increasing the channels of the transfer system and dispersing the effect of fluctuation. To increase the number of transfer channels, I adopt sound field control with multiple loudspeakers and microphone array. By sound field control, the response sound is cancelled at each point of the microphone elements. In this dissertation, I propose two versions of sound field control for response sound reduction. The first one reproduces the response sound with high quality at the user's ears, and the second version mitigates the strict reproduction. While the first one is superior to the second in the quality of the reproduction, the second one can eliminate the response sound efficiently with fewer loudspeakers.

Source separation to eliminate both response sound and interfering noise

To reinforce the response sound elimination by sound field control and to eliminate interfering noise, I combine sound field control with adaptive signal processing. To realize speech enhancement without DTDs, I adopt unsupervised source

Table 3. Typical beamformers

Name	Adaptation	Features
Delay and sum [FJZE85]	None	The most basic beamformer for microphone array with simple structure to synchronize signals from specific direction.
Linear constrained minimum variance (LCMV) [Fro72]	Batch	The most basic adaptive beamformer.
Generalized sidelobe canceller (GSC) [GJ82]	On-line	Approximation of LCMV for on-line adaptation with least mean squares.
Adaptive microphone array for noise reduction [KO86]	On-line	Modification of GSC to reduce distortion by making virtual speech signal.

separation. To eliminate both the known response sound and unknown interfering noise efficiently, I propose a new structure of partially unsupervised source separation.

1.3 Overview of dissertation

In Chapter 2, the basic idea of the adaptation of AEC and ABF are reviewed. In the discussion of the optimal filter, I point out the problem that the DTD is indispensable in the adaptation.

In Chapter 3, a new mechanism of response sound elimination using sound field reproduction and a microphone array is proposed. The mechanism of obtaining robustness is discussed both theoretically and experimentally.

In Chapter 4, a modified version of the sound field control to realize efficient elimination with fewer loudspeakers is proposed.

In Chapter 5, a new source separation technique is proposed to separate both the known and unknown sources, and it is combined with the elimination based

on sound field reproduction.

Chapter 6 gives a conclusion of the entire dissertation and the plans for future work.

2. Adaptive Signal Processing for Acoustic Echo Canceller and Adaptive Beamformer

2.1 Introduction

In this chapter, I review conventional adaptive signal processing methods used in hands-free robust speech recognition in a spoken dialogue system. In spoken dialogue system, noise can be classified into two categories; known noise as response sound of the spoken dialogue system, and unknown noise as interfering noise in the environment. I describe a basic ideas of AEC to eliminate the response sound and ABF to reduce interfering noise.

The chapter is organized as follows: AEC is reviewed in Chapter 2.2, followed by review of ABF in Chapter 2.3. Finally, the chapter is summarized in Chapter 2.4.

2.2 Acoustic echo canceller

AEC is a problem to reduce feedback of the response sound from the loudspeaker to the microphone. Configuration of AEC is shown in Fig. 2. The observed signal $x(\omega)$ of the microphone can be described as

$$x(\omega) = g(\omega) r_{\mathrm{src}}(\omega) + e(\omega), \tag{1}$$

where ω shows angular frequency, $g(\omega)$ denotes transfer function between the loudspeaker and the microphone, $r_{\mathrm{src}}(\omega)$ shows source of the response sound, and $e(\omega)$ is a signal component related to user's speech or interfering noise. The processed signal $y(\omega)$ is written as

$$y(\omega) = x(\omega) + \hat{g}(\omega) r_{\mathrm{src}}(\omega), \tag{2}$$

and the goal is to conform $y(\omega)$ to $e(\omega)$ by optimizing $\hat{g}(\omega)$ to identify antiphase of $g(\omega)$ as

$$g(\omega) + \hat{g}(\omega) = 0. \tag{3}$$

Here the power of the processed signal $y(\omega)$ can be expanded as

$$\begin{aligned} |y(\omega)|^2 =& |e(\omega)|^2 + |(g(\omega) + \hat{g}(\omega))\, r_{\mathrm{src}}(\omega)|^2 \\ &+ 2\ \mathrm{real}\left[e^*(\omega)(g(\omega) + \hat{g}(\omega) r_{\mathrm{src}}(\omega)\right], \end{aligned} \tag{4}$$

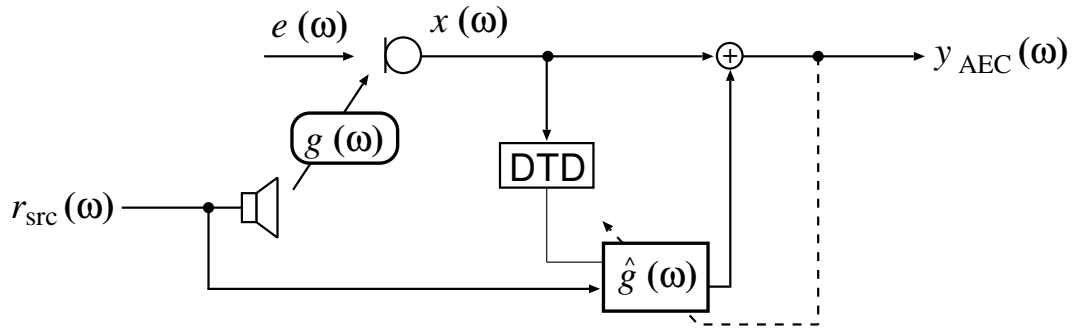

Figure 2. Configuration of adaptation in AEC.

where real[·] denotes real part of conjugate value and $\{\cdot\}^*$ denotes complex conjugate. Since $e(\omega)$ does not include component due to the response signal $r_{\rm src}(\omega)$, uncorrelation between $e(\omega)$ and $r_{\rm src}(\omega)$ can be assumed, i.e.,

$$\frac{E\left[e(\omega)r^*_{\rm src}(\omega)\right]_\omega}{\sqrt{E\left[|e(\omega)|^2\right]_\omega}\sqrt{E\left[|r(\omega)|^2\right]_\omega}} = 0, \tag{5}$$

where $E[\cdot]_i$ denotes expectation with respect to i. Thus the third term of the right side in Eq. (4) can be neglected and optimum condition in Eq. (3) minimizes $|y(\omega)|^2$. Thus $|y(\omega)|^2$ can be used as the criterion, and as shown in [WGM+75], the optimum solution can be obtained as

$$\hat{g}(\omega) = \frac{x(\omega)r^*_{\rm src}(\omega)}{|r_{\rm src}(\omega)|^2}. \tag{6}$$

However, with finite length of signals, assumption of uncorrelation in Eq. (5) is not satisfied because of statistical bias. For real-time usage in speech application, adaptation is required to be obtained with short length of observed signals. To realize this, adaptation is generally conducted in the detected time duration by DTD when $s(\omega) = 0$.

Although I have shown batch-wise adaptation, desirable real-time solution is on-line adaptation rather than batch-wise adaptation. Many types of on-line adaptation, e.g., least mean squares or recursive least squares approximate batch optimum solution Eq. (6). However, problem of statistical bias affects more severely to such on-line adaptation, and mis-detection of on-line DTD results as large error in adaptation.

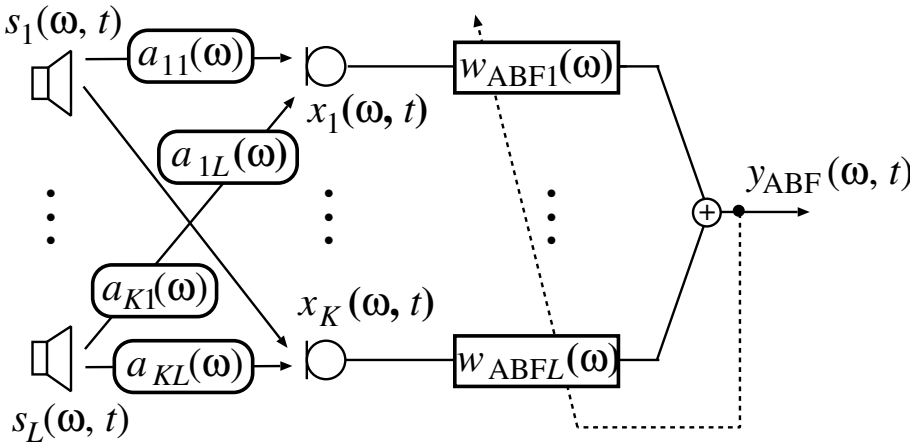

Figure 3. Configuration of adaptation in AEC.

2.3 Adaptive beamformer

ABF is a noise-reduction method by forming directional null to interfering noise. Here I mention linear-constrained minimum variance (LCMV) beamformer [Fro72], which is the most basic adaptation of ABF. Configuration of ABF is shown in Fig. 3. Assume L source signals $\boldsymbol{s}(\omega) = [s_1(\omega), \ldots, s_L(\omega)]^{\mathrm{T}}$ are observed at K microphones as $\boldsymbol{x}(\omega) = [x_1(\omega), \ldots, x_K(\omega)]^{\mathrm{T}}$, and the first source $s_1(\omega)$ is the desired signal while the others are interfering noise. The observed signals $\boldsymbol{x}(\omega)$ can be written as

$$\boldsymbol{x}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{s}(\omega), \tag{7}$$

$$\boldsymbol{A}(\omega) = [a_{kl}(\omega)]_{kl} \text{ for } k = 1, \ldots, K, l = 1, \ldots, L, \tag{8}$$

where $[x]_{ij}$ denotes a matrix who has an entry x in the i-th row and j-th column, and $a_{kl}(\omega)$ denotes a transfer function from the l-th source to the k-th microphone. Noise reduction using the beamformer is achieved by optimizing beamformer coefficients $\boldsymbol{w}_{\mathrm{ABF}}(\omega)$, denoted as

$$\boldsymbol{w}_{\mathrm{ABF}}(\omega) = [w_{\mathrm{ABF1}}(\omega), \ldots, w_{\mathrm{ABF}K}(\omega)], \tag{9}$$

orthogonal to the transfer functions related to the interfering noise, as

$$\boldsymbol{w}_{\mathrm{ABF}}(\omega)\boldsymbol{a}_l(\omega) = 0 \text{ for } l = 2, \ldots, L, \tag{10}$$

where $\boldsymbol{a}_l(\omega) = [a_{1l}, \ldots, a_{Kl}(\omega)]^{\mathrm{T}}$. Then its output $y_{\mathrm{ABF}}(\omega)$ includes only the desired signal $s_1(\omega)$ as

$$y_{\mathrm{ABF}}(\omega) = \boldsymbol{w}_{\mathrm{ABF}}(\omega)\boldsymbol{x}(\omega) = \boldsymbol{w}_{\mathrm{ABF}}(\omega)\boldsymbol{a}_1(\omega)s_1(\omega). \tag{11}$$

To optimize $\boldsymbol{w}_{\mathrm{ABF}}(\omega)$, first short-time Fourier analysis is applied to the observed signals $x(\omega)$. I denote the short-time spectrum of $\boldsymbol{x}(\omega)$ as $\boldsymbol{x}(\omega, t) = [x_1(\omega, t), \ldots, x_K(\omega, t)]^{\mathrm{T}}$, where $[\cdot]^{\mathrm{T}}$ denotes matrix transposition and t shows index of analysis frame. The filter coefficients $\boldsymbol{w}_{\mathrm{ABF}}(\omega)$ are optimized by minimizing the output power when the desired source $s_1(\omega)$ is inactive. Assume a DTD successfully finds set $\mathcal{T}$ of the frame index t when $s_1(\omega, t)$ is inactive. Then the optimal solution can be written as

$$\text{minimize } \langle|\boldsymbol{w}_{\mathrm{ABF}}(\omega)\boldsymbol{x}(\omega, t)|^2\rangle_{t\in\mathcal{T}}, \tag{12}$$

where $\langle\cdot\rangle_t$ denotes time-averaging operation. Here, to prevent directional null to the direction of the desired signal, the gain against the direction is constrained to 1 as

$$\boldsymbol{w}_{\mathrm{ABF}}(\omega)\boldsymbol{q}(\omega) = 1, \tag{13}$$

where a column vector $\boldsymbol{q}(\omega)$ is a rough approximation of $\boldsymbol{a}_1(\omega)$ called steering vector, which should be estimated in advance. Optimal solution of $\boldsymbol{w}_{\mathrm{ABF}}(\omega)$ to minimize Eq. (12) under the constraint Eq. (13) can be obtained as

$$\boldsymbol{w}_{\mathrm{ABF}} = \frac{\boldsymbol{q}^{\mathrm{H}}(\omega)\boldsymbol{R}^{-1}(\omega)}{\boldsymbol{q}^{\mathrm{H}}(\omega)\boldsymbol{R}^{-1}(\omega)\boldsymbol{q}(\omega)}, \tag{14}$$

where $\{\cdot\}^{\mathrm{H}}$ denotes conjugate transposition and $\boldsymbol{R}(\omega)$ is a covariance matrix of the observed noise given by

$$\boldsymbol{R}(\omega) = \langle\boldsymbol{x}(\omega, t)\boldsymbol{x}^{\mathrm{H}}(\omega, t)\rangle_{t\in\mathcal{T}}. \tag{15}$$

LCMV gives optimal solution of batch-wise adaptation, and there are many modifications to on-line adaptation, e.g., generalized sidelobe canceller [GJ82]. As discussed above, DTD is indispensable in adaption of ABF to detect time duration when only the signal component due to the interfering noise is included in the observed signals. DTD is difficult when both desired signal and interfering noise are speech.

In addition, when combined with AEC, DTDs are more difficult to be implemented because of the complication; DTD for AEC is to detect only the response sound while DTD for ABF is to detect only the interfering noise. In [HK02], integration of AEC and ABF is discussed. In addition, their structure (AEC-first or their joint optimization) and their DTDs are discussed in [HNK05] and [HBNK05], respectively. The optimum structure depends on SNR. However, SNR is hardly known in advance, and it is heavily variable. Also, DTDs are very complicated even in the simple environment with few sources of noise in [HBNK05], and the features for the classification of noise have much overlap in each pair of classes.

2.4 Conclusion

In this chapter, I reviewed two adaptive signal processing methods for robust speech recognition in a spoken dialogue system. While AEC is generally used to eliminate response sound feed back to observed signal, an adaptive beamformer is used to eliminate interfering noise. I pointed out that both of them requires DTDs for adaptation of their filters, which are difficult to be implemented.

3. Response Sound Elimination Based on Sound Field Reproduction

3.1 Introduction

In this chapter, I describe a new elimination technique of the response sound using sound field reproduction, which is the most important part of the proposed framework. As discussed in the previous chapter, AEC cannot always adapt its filter coefficients because of the double-talk problem. Since AEC should obtain accurate antiphase of the echo return, the elimination performance degrades extremely when AEC cannot be adapted to the acoustical environment, which is heavily variable in general. However, since the mechanism of the elimination is robust against the fluctuation of the acoustical room transfer system, the response sound can be eliminated without adaptation by designing the filter using the acoustical impulse responses measured when the devices are settled.

In order to achieve robustness, I propose a new interface for a barge-in robust spoken dialogue system that combines multichannel sound field control and beamforming. At first, to prevent the response sound from being observed at the microphone elements, I utilize the sound field reproduction technique via multiple loudspeakers and an inverse filter of the room transfer functions [MK88]. The sound field reproduction is generally used in a *transaural* system [BC96], which presents a three-dimensional sound image to a user at a fixed position. I apply this technique to the response sound elimination by controlling sound field around the microphone to be silent alongside the transaural reproduction at the user's ears, denoted as the multiple-output and multiple-no-input (MOMNI) method. In the next step, the user's speech is enhanced by beamforming. By increasing the numbers of loudspeakers and microphone elements, the control of the MOMNI method becomes robust against the fluctuation of the room transfer functions. With sufficient numbers of loudspeakers and microphones, the MOMNI method enables us to eliminate the response sound with enough robustness to sustain speech recognition accuracy.

Although the MOMNI method requires many loudspeakers and the cost for the hardware is higher than the conventional acoustic echo canceller, the MOMNI

method uses a fixed filter designed in advance and real-time adaptation is unnecessary. As a result, computational cost can be reduced. In addition, the MOMNI method has an advantage that sound virtual reality [TSS01] can be achieved with transaural reproduction.

In Chapter 3.2, I describe the weakness of the conventional acoustic echo canceller against fluctuation of the room impulse responses. In Chapter 3.3, I describe the principle of the MOMNI interface. In Chapter 3.4, an experimental comparison of response sound elimination performances is carried out. In Chapter 3.5, the effectiveness of the MOMNI method is validated in the speech recognition experiment. In Chapter 3.6, I assess the quality of the response sound reproduced by the MOMNI method.

3.2 Response sound elimination error of the acoustic echo canceller when fluctuation of the room transfer function occurs

The room transfer functions are easily changed with the variation of the system's state such as the movement of people. In this section, the response sound elimination error signal $d'(\omega)$ of the AEC is examined in the case where the transfer function is changed. Suppose that the room transfer function $g(\omega)$ in Eq. (1) is changed to $g'(\omega)$ by additive variation $\Delta g(\omega)$ caused by the fluctuation of room transfer function, described as

$$g'(\omega) = g(\omega) + \Delta g(\omega). \tag{16}$$

In this case, the observed response sound $d'(\omega)$ can be written as

$$d'(\omega) = [g(\omega) + \Delta g(\omega)] r_{\mathrm{src}}(\omega). \tag{17}$$

The elimination error signal $\epsilon(\omega)$ of the response sound is written using the estimated filter $\hat{g}(\omega)$ as

$$\epsilon(\omega) = \Delta g(\omega) r_{\mathrm{src}}(\omega), \tag{18}$$

where it is assumed that the filter was exactly estimated so as to satisfy the condition in Eq. (3) and

$$g(\omega) r_{\mathrm{src}}(\omega) + \hat{g}(\omega) r_{\mathrm{src}}(\omega) = 0. \tag{19}$$

Since the acoustic echo canceller has no mechanism for improving the robustness of the elimination (unless it contains a suitable post-processing for that case), the fluctuation of the transfer function effects directly to its error. Therefore, if the fluctuation occurs when the adaptation stops because of barge-in, its elimination performance degrades. To prevent the degradation, adaptation of AEC must follow the fluctuation.

3.3 Multiple-output and multiple-no-input method

In this section, I propose a new response sound elimination technique using sound field reproduction, which is robust against the fluctuation of the room transfer function. The MOMNI method mainly consists of two steps. First, sound field control with multiple loudspeakers realizes silent zones at the microphone elements while the dialogue system gives the response sound to the user. Next, by delay-and-sum-type beamforming using a microphone array, the residual component of the response sound caused by the fluctuation of the transfer function is suppressed and the user's utterance is emphasized. The response sound signal is outputted from the multiple loudspeakers and cancelled at multiple control points. With this mechanism, the response sound is prevented from being inputted to the speech recognition system. Thus this technique is so-called Multiple-Output/Multiple-No-Input (MOMNI) method. I describe the relation between the robustness of the control and the number of transfer channels. Then it is proved that the MOMNI method can improve its robustness against the fluctuation of the transfer functions by increasing the numbers of loudspeakers and microphone elements. With sufficient numbers of loudspeakers and microphones, the MOMNI method can eliminate the response sound with enough robustness using fixed filter coefficients. Needless to say, this processing requires no double-talk detection.

3.3.1 Sound field reproduction

Here, I describe the sound field control used to eliminate the acoustic echo of the response sound from the system. The configuration of the MOMNI system is shown in Fig. 4. Let M be the number of secondary sound sources $\mathrm{S}_1, \ldots, \mathrm{S}_M$ and

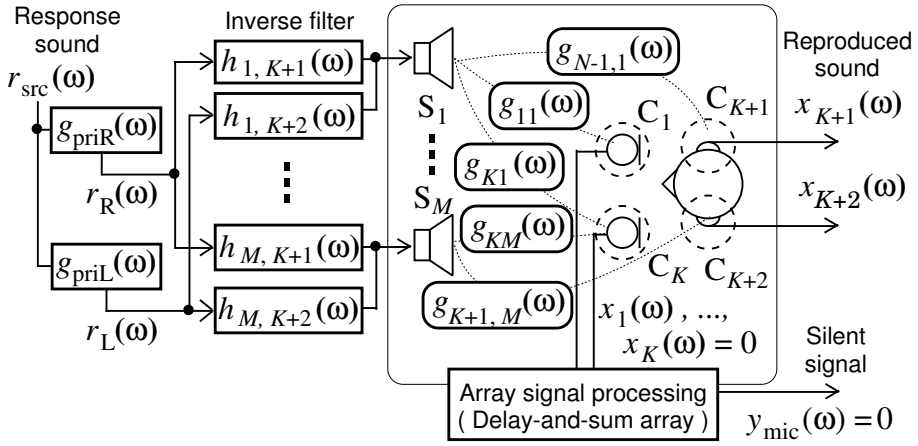

Figure 4. Configuration of the proposed system.

N be the number of control points $\mathrm{C}_1, \ldots, \mathrm{C}_N$. The control points $\mathrm{C}_1, \ldots, \mathrm{C}_K$ ($K = N - 2$) are arranged to the elements of a microphone array for acquisition of a user's speech, and C_{K+1} and C_{K+2} are set at both ears of the user. The signals to be reproduced at the control points $\mathrm{C}_1, \ldots, \mathrm{C}_{K+2}$ are described by

$$\boldsymbol{r}(\omega) = [r_1(\omega), \ldots, r_K(\omega), r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]^{\mathrm{T}}, \tag{20}$$

where T denotes matrix transposition. Similarly, the signals observed at these control points are represented by

$$\boldsymbol{d}(\omega) = [d_1(\omega), \ldots, d_{K+2}(\omega)]^{\mathrm{T}}. \tag{21}$$

Using, e.g., chirp signal [SA95], all of the transfer functions from secondary sound sources S_m should be measured in advance to control points C_n, denoted by $g_{nm}(\omega)$, where $n = 1, \ldots, N$, and $m = 1, \ldots, M$. Here, to design an inverse filter of the transfer functions with nonminimum phases, the condition

$$M > N \tag{22}$$

must hold [MK88]. To use fixed filter coefficients for the inverse filter, the positions of the loudspeakers and the microphones should not be changed after the measurement. In addition, the position for the user to listen to the response sound is specified by, for example, setting a chair at the position. Here in the phase of

the measurement, to obtain the transfer function of the user's ears, since it is a burden for the user to sit on the position wearing microphones at his/her ears, the user can be substituted by a head and torso simulator (HATS) with microphones at the ears. Let $\boldsymbol{G}(\omega) = [g_{nm}(\omega)]_{nm}$ be an $N \times M$ matrix and $\boldsymbol{H}(\omega) = [h_{mn}(\omega)]$ be an $M \times N$ inverse filter of $\boldsymbol{G}(\omega)$, where $[x]_{ij}$ denotes matrix who has an entry x in the i-th row and j-th column, and $g_{nm}(\omega)$ denotes a transfer function from the m-th loudspeaker to the n-th control point. The inverse filter $\boldsymbol{H}(\omega)$ is then designed so that

$$\boldsymbol{G}(\omega)\boldsymbol{H}(\omega) = \boldsymbol{I}_N(\omega), \tag{23}$$

where $\boldsymbol{I}_N(\omega)$ denotes an $N \times N$ identity matrix. Using the transfer function matrix $\boldsymbol{G}(\omega)$ and the inverse filter matrix $\boldsymbol{H}(\omega)$, the relation between the observed signals $\boldsymbol{d}(\omega)$ and the reproduced signals $\boldsymbol{r}(\omega)$ is written as

$$\boldsymbol{d}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{H}(\omega)\boldsymbol{r}(\omega). \tag{24}$$

In Eq. (24), the response sound of a dialogue system is reproduced at both of the user's ears (i.e., $[d_{K+1}(\omega), d_{K+2}(\omega)] = [r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]$), and reproduce silent signals with zero amplitudes at the microphone elements (i.e, $[d_1(\omega), \ldots, y_K(\omega)] = [0, \ldots, 0]$) as

$$\boldsymbol{r}(\omega) = [\underbrace{0, \ldots, 0}_{K}, r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]^{\mathrm{T}}. \tag{25}$$

By this sound reproduction, a sound field can be actualized in which the response sound is presented to the user while the response sound cancels at the microphone elements.

To remove the redundant filtering process of the zero signals, the matrix $\boldsymbol{H}(\omega)$ is truncated into $\boldsymbol{H}_2(\omega) = [h_{m,k+K}]_{mk}$ for $m = 1, \ldots, M$, $k = 1, 2$ which is an $M \times 2$ truncated filter matrix of $\boldsymbol{H}(\omega)$. By inputting the response sound to this filter matrix, the following equation holds

$$\begin{aligned} \boldsymbol{d}(\omega) &= \boldsymbol{G}(\omega)\boldsymbol{H}_2(\omega)[r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]^{\mathrm{T}} \\ &= [\underbrace{0, \ldots, 0}_{K}, r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]^{\mathrm{T}}. \end{aligned} \tag{26}$$

Therefore, the condition equivalent to Eq. (25) can be realized with an $M \times 2$ filter matrix $\boldsymbol{H}_2(\omega)$.

Since the MOMNI method uses an inverse filter of the room transfer function, the response sound can be shown to the user in the form of a *transaural* system, say, a three-dimensional sound field localization [BC96]. In transaural system, a clear sound image of a primary sound source can be shown to the user by reproducing a binaural signal [Bla97], say, a convolution of a signal and transfer functions from the sound source to a person's ears. To provide a practical application of this property, I generate the response sound signals $r_{\mathrm{R}}(\omega)$ and $r_{\mathrm{L}}(\omega)$ by multiplying a monaural source of the response sound signal $r_{\mathrm{src}}(\omega)$ and the room transfer functions $\boldsymbol{g}_{\mathrm{pri}}(\omega)$ between a primary sound source and both of the user's ears, as

$$\boldsymbol{g}_{\mathrm{pri}}(\omega) = [g_{\mathrm{priR}}(\omega), g_{\mathrm{priL}}(\omega)]^{\mathrm{T}} \tag{27}$$

and

$$[r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]^{\mathrm{T}} = \boldsymbol{g}_{\mathrm{pri}}(\omega) r_{\mathrm{src}}(\omega). \tag{28}$$

Such reproduction can be written as

$$\boldsymbol{d}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega) r_{\mathrm{src}}(\omega) = [0, \ldots, 0, r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]. \tag{29}$$

In the transaural reproduction described above, the sound image is degraded when the user is not at the prepared position because the perceived response sound is not an accurate binaural sound. However, the sound quality away from the prepared position is sufficient for the presentation of the response sound for the spoken dialogue system. I will justify this argument in the experiment in Chapter 3.6.

3.3.2 Delay-and-sum beamformer

In this section, I will focus attention on array signal processing. In this study, a delay-and-sum beamforming [FJZE85] is adopted to emphasize the user's utterance. The filter of the k-th element in the delay-and-sum beamformer is denoted by $w_k(\omega)$ for $k = 1, \ldots, K$. Then $w_k(\omega)$ can be expressed as

$$w_k(\omega) = \frac{1}{K} \cdot e^{-j\omega\tau_k}, \tag{30}$$

where τ_k stands for the arrival time difference of the user's utterance between a suitable standard point and the k-th element position. I set τ_k to form a directivity

to the look direction of the user. Suppose that the signal added through the beamforming filters is a signal for speech recognition. Then the response sound $d_{\mathrm{DS}}(\omega)$ contained in the processed signal is expressed as

$$y_{\mathrm{DS}}(\omega) = \sum_{k=1}^{K} x_k(\omega). \tag{31}$$

When this delay-and-sum-type beamformer is used, the system's response sound which arrives 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. As a result, only the user's speech can be emphasized in the $y_{\mathrm{DS}}(\omega)$. Thus this signal is given to the speech decoder to recognize the user's speech.

3.3.3 Inverse system design for sound field reproduction

In a multipoint control system which controls multiple control points with many loudspeakers, large amounts of calculation and memory are needed to design an inverse filter in the time domain. Therefore, I design the inverse filter matrix $\boldsymbol{H}(\omega)$ by using the least-norm solution (LNS) in the frequency domain [TSS01]. The method has advantages that the amount of calculation is small in the frequency domain, and the designed system is stable because the output from each sound source is suppressed to the minimum. As an inverse filter, I adopt Moore-Penrose generalized inverse matrix as the inverse matrix which gives the least-norm solution. I obtain a singular value decomposition of $\boldsymbol{G}(\omega)$ as

$$\boldsymbol{G}(\omega) = \boldsymbol{U}(\omega)[\boldsymbol{\Gamma}_N(\omega), \boldsymbol{O}_{N,M-N}(\omega)]\, \boldsymbol{V}^{\mathrm{H}}(\omega), \tag{32}$$

$$\boldsymbol{\Gamma}_N(\omega) \equiv \mathrm{diag}[\mu_1(\omega), \mu_2(\omega), \ldots, \mu_N(\omega)], \tag{33}$$

where $\boldsymbol{U}(\omega)$ and $\boldsymbol{V}(\omega)$ are $N\times N$ and $M\times M$ unitary matrices, respectively, $\mu_n(\omega)$ for $n = 1, 2, \ldots, N$ are the singular values of $\boldsymbol{G}(\omega)$, and are arranged so that $\mu_n(\omega)\geq\mu_{n+1}(\omega)$ in matrix $\boldsymbol{\Gamma}_N(\omega)$, $\boldsymbol{O}_{N,M-N}(\omega)$ denotes an $N\times(M-N)$ null matrix, and $\{\cdot\}^{\mathrm{H}}(\omega)$ represents a conjugate transposition.

Then the Moore-Penrose generalized inverse matrix $\boldsymbol{G}^{+}(\omega)$ of $\boldsymbol{G}(\omega)$ is given by

$$\boldsymbol{G}^{+}(\omega) = \boldsymbol{V}(\omega) \begin{bmatrix} \boldsymbol{\Lambda}_N(\omega) \\ \boldsymbol{O}_{M-N,N}(\omega) \end{bmatrix} \boldsymbol{U}^{\mathrm{H}}(\omega), \tag{34}$$

$$\boldsymbol{\Lambda}_N(\omega) \equiv \mathrm{diag}\left[\frac{1}{\mu_1(\omega)}, \frac{1}{\mu_2(\omega)}, \ldots, \frac{1}{\mu_N(\omega)}\right]. \tag{35}$$

Then I utilize the Moore-Penrose generalized inverse matrix for the inverse filter as $\boldsymbol{H}(\omega) = \boldsymbol{G}^{+}(\omega)$.

3.3.4 Response sound elimination error for fluctuation of room transfer functions

In an acoustic echo canceller, because the transfer function should be reestimated when it is changed, there is a problem that the response sound elimination accuracy degrades during the estimation process. In contrast, it is proven that the MOMNI technique is robust against the fluctuation of room transfer functions, even when the fixed filter coefficients are used. Here, suppose that an inverse filter matrix computed before the fluctuation is used to control the sound field.

Supposing that the variation $\Delta g_{nm}(\omega)$ caused by the fluctuation of transfer functions is added to a transfer function $g_{nm}(\omega)$, the transfer function matrix after the fluctuation will become $\boldsymbol{G}(\omega) + \Delta\boldsymbol{G}(\omega)$, where $\Delta\boldsymbol{G}(\omega)$ is an $N \times M$ matrix composed of $\Delta g_{nm}(\omega)$. Then, by using an inverse filter matrix $\boldsymbol{H}(\omega)$ designed before the fluctuation of transfer functions, the response sound signals $\boldsymbol{d'}(\omega)$ at each of the control points are expressed as

$$\begin{aligned} \boldsymbol{d'}(\omega) &= [\boldsymbol{G}(\omega) + \Delta\boldsymbol{G}(\omega)]\boldsymbol{H}(\omega)\boldsymbol{r}(\omega) \\ &= [\boldsymbol{I}_N(\omega) + \Delta\boldsymbol{G}(\omega)\boldsymbol{H}(\omega)]\boldsymbol{r}(\omega), \end{aligned} \tag{36}$$

and the errors caused by the fluctuation are represented as $\Delta\boldsymbol{G}(\omega)\boldsymbol{H}(\omega)\boldsymbol{r}(\omega)$. In this case, the error $\Delta y_{\mathrm{DS}}(\omega)$ in the processed signal $\epsilon(\omega)$ of the DS beamformer in Eq. (31) is written as

$$\epsilon(\omega) = \sum_{k=1}^{K} w_k(\omega)\Big\{\sum_{m=1}^{M} \Delta g_{(k+2)m}(\omega) \cdot [h_{m1}(\omega) r_{\mathrm{R}}(\omega) + h_{m2}(\omega) r_{\mathrm{L}}(\omega)]\Big\}. \tag{37}$$

because the MOMNI method controls $y_{\mathrm{mic}}(\omega)$ such that it is 0 before the fluctuation of transfer functions, $\Delta y_{\mathrm{DS}}(\omega)$ after the fluctuation is the response sound elimination error signal $\epsilon(\omega)$. This is expressed as

$$\begin{aligned} \epsilon(\omega) &= y_{\mathrm{DS}}(\omega) + \Delta y_{\mathrm{DS}}(\omega) \\ &= \Delta y_{\mathrm{DS}}(\omega). \end{aligned} \tag{38}$$

Next, let the singular values of $\boldsymbol{G}(\omega)$ be $\mu_j(\omega)$ for $j = 1, 2, \cdots, N$ and let the eigenvalues of $\boldsymbol{G}^{\mathrm{H}}(\omega)\boldsymbol{G}(\omega)$ be $\lambda_j(\omega)$ for $j = 1, 2, \cdots, N$. Then, the norm $\|\boldsymbol{G}(\omega)\|$ is given by

$$\begin{aligned} \|\boldsymbol{G}(\omega)\| &= \sqrt{\max_j(\lambda_j(\omega))} \\ &= \sqrt{\max_j(\{\mu_j(\omega)\}^2)} \\ &= |\mu_1(\omega)|, \end{aligned} \tag{39}$$

where $\max_j(a_j)$ denotes the largest element of a_j for any j. The relation $\lambda_j(\omega) = \{\mu_j(\omega)\}^2$ is used here.

Alternatively, since the singular values of $\boldsymbol{G}^{+}(\omega)$ are given by $1/\mu_j(\omega)$, the norm $\|\boldsymbol{G}^{+}(\omega)\|$ is expressed as

$$\begin{aligned} \|\boldsymbol{G}^{+}(\omega)\| &= \sqrt{\max_j\left(\frac{1}{\lambda_j(\omega)}\right)} \\ &= \sqrt{\max_j\left(\frac{1}{\{\mu_j(\omega)\}^2}\right)} \\ &= \frac{1}{|\mu_N(\omega)|}. \end{aligned} \tag{40}$$

Since the secondary sound source is arranged with almost equal distance for each control point, if the number of secondary sound sources, M, increases, the norm of $\boldsymbol{G}(\omega)$ is directly proportional to M, i.e., $\|\boldsymbol{G}(\omega)\| \propto M$. Moreover, although the condition number of $\boldsymbol{G}(\omega)$, which is expressed by the ratio between the maximum and minimum singular values, i.e.,

$$\mathrm{cond}(\boldsymbol{G}) = \mu_1/\mu_N, \tag{41}$$

is known that the condition number will be close to unity when the number of secondary sound sources arranged is much larger than that of control points (this is experimentally proven in Chapter 3.4.3 and shown in Fig. 12). Therefore, the following relation can be derived from Eqs. (39) and (40):

$$\begin{aligned} \|\boldsymbol{H}(\omega)\| &= \|\boldsymbol{G}^{+}(\omega)\| = \frac{1}{|\mu_N(\omega)|} \\ &\simeq \frac{1}{|\mu_1(\omega)|} = \frac{1}{\|\boldsymbol{G}(\omega)\|} \propto \frac{1}{M}. \end{aligned} \tag{42}$$

Substituting Eq. (30) into Eq. (37) obtains

$$\begin{aligned}\Delta\hat{y}_{\mathrm{DS}}(\omega) &= \|\boldsymbol{H}(\omega)\|\frac{1}{K}\Big\{\sum_{k=1}^{K}\sum_{m=1}^{M}\Delta g_{km}(\omega) \\ &\quad \cdot\Big[\bar{h}_{m(K+1)}(\omega)r_{\mathrm{R}}(\omega)+\bar{h}_{m(K+2)}(\omega)r_{\mathrm{L}}(\omega)\Big] \\ &\quad \cdot e^{-j\omega\tau_k}\Big\}, \end{aligned} \tag{43}$$

where $\bar{h}_{mn}(\omega)=h_{mn}(\omega)/\|\boldsymbol{H}(\omega)\|$. Assume that $\Delta g_{nm}(\omega)$ for $n=1,2,\cdots,N$ and $m=1,2,\cdots,M$ are mutually independent and follow the same Gaussian distribution with zero-mean and variance σ^2. Furthermore, since $\bar{h}_{mn}(\omega)$ is a function normalized by $\|\boldsymbol{H}(\omega)\|$ and independent on M, the deviation of $\{\cdot\}$ in Eq. (43) can be represented by $\eta\sqrt{MK}\sigma$, where η is a suitable constant. Therefore, the elimination error $\epsilon(\omega)$ of response sound is obtained from Eq. (42) as

$$\begin{aligned}\epsilon(\omega)=\Delta y_{\mathrm{DS}}(\omega) &\propto \frac{1}{M}\cdot\frac{1}{K}\cdot\sqrt{MK} \\ &= \frac{1}{\sqrt{MK}}. \end{aligned} \tag{44}$$

In other words, Eq. (44) shows that the elimination error of the response sound for the fluctuation of the transfer functions is inversely proportional to $\sqrt{MK}$. Thus, if the number of transfer channels from loudspeakers to microphones increases, the response sound elimination of the MOMNI method improves its robustness against the fluctuation of the transfer functions.

In the real environment, it is remarked to be difficult to prove whether or not the variations $\Delta g_{nm}(\omega)$ caused by the fluctuation of the room transfer functions are mutually independent for every channel from a loudspeaker to a microphone. However, in the next section, the simulations using impulse responses measured in the real environment show that the error estimation in Eq. (44) is valid.

3.3.5 Computational complexity

Here computational complexity of the MOMNI method is discussed. In the following, N_{MOMNI} denotes filter length in taps of the MOMNI method's inverse filter, and N_{AEC} denotes that of the AEC. Typical values of N_{MOMNI} and N_{AEC}

Table 4. Computational cost of the AEC and the MOMNI method.

Method	Order	Typical values
Single-channel AEC	$3B \log_2 2N$	264 ($B = 8, N = 1024$)
Stereo AEC	$6B \log_2 2N$	528 ($B = 8, N = 1024$)
MOMNI method	$(M + 2) \log_2 2N + 2M$	268 ($M = 16, N = 8192$)

are 1024 and 8192, respectively. We summerized the computational cost for the AEC and the MOMNI method in Table 4.

If AEC is adapted using simple normalized LMS, the number of multiplications required is N_{AEC} per sample. When the AEC uses the typical filter length $N_{\mathrm{AEC}} = 1024$, the computational quantity is very high. If adaptation is conducted in the frequency domain with block-partitioned samples, the significant factors are the FFT of the observed and the reference signals, and IFFT of the processed signal; the number of the multiplications required is $3 \log_2 2N_{\mathrm{AEC}}$ per sample if the blocks have no overlap. When the multiple microphones and loudspeakers are used, the computation increases in proportion to the numbers. In addition, to improve convergence speed, the blocks should be overlapped. Then the number of the multiplications are $B(2K + M) \log_2 2N_{\mathrm{AEC}}$ per sample where B is the number of the block overlaps, K the number of the microphones, M the number of the loudspeakers. Thus the typical stereo AEC with $K = 2$, $M = 2$, $B = 8$ requires 528 multiplications. In this estimation of the computation, I ignored the computation required for the control of the adaptation. For further detailed discussion, please see [Her04] where joint adaptation of AEC and ABF is implemented in real time. Note that many DTD algorithm requires similar computation to that of adaptation.

In real-time implementation of the MOMNI method, only the computation of filtering matters because the filter is designed off line. Since the inverse filter of the MOMNI method is the fixed filter, simple overlap-add filtering with rectangle window can be used. FFT and IFFT are required for input signals of two channels for the user's ears and the M-channel loudspeaker outputs. Thus the number of the multiplications required for the FFT is $(2 + M) \log_2 2N_{\mathrm{MOMNI}}$. In addition, $2M$ multiplications are required for the filtering in frequency domain. Note that

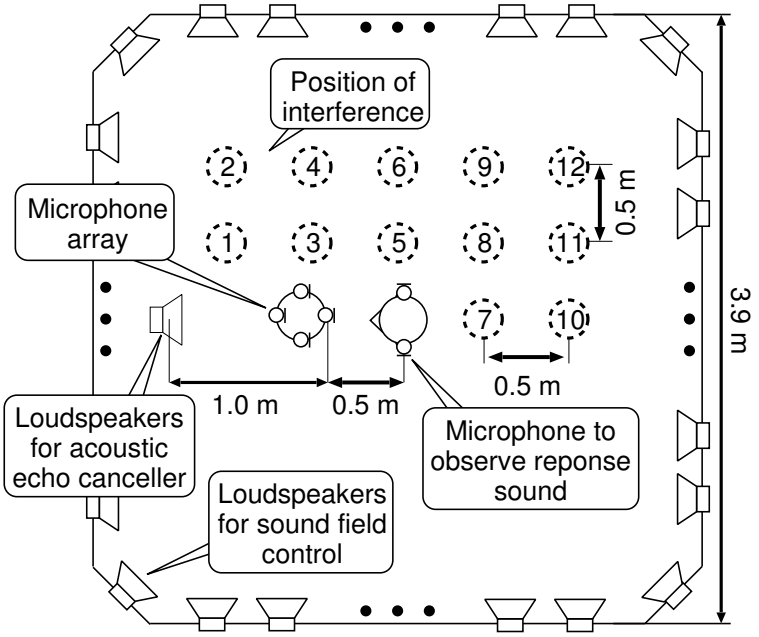

Figure 5. Layout of acoustic experiment room.

number of the microphone-array elements does not affect the complexity of the filtering. In total, $(2 + M) \log_2 2N_{\mathrm{MOMNI}} + 2M$ multiplications are required for the filtering in the MOMNI method; 268 multiplications in the typical setting with $M = 16$. Therefore, the computational complexity of the MOMNI method is comparable or lower level to the AEC.

3.4 Experimental comparison of response sound elimination performance

To assess the robustness of the MOMNI method against the fluctuation of the room transfer functions, the response sound elimination performance of the MOMNI method is evaluated by simulations. Its performance is compared with that of conventional acoustic echo canceller.

3.4.1 Experimental conditions

The simulations are carried out by using impulse responses measured in a real acoustic environment. Figure 5 shows the arrangement of the apparatuses. To imitate the user at the center of the room, I set a HATS. To cause fluctuations of the room transfer functions intentionally, I placed a life-size mannequin as an interference near a user, under the assumption that a person approaches to the user. I measured in a total of 13 patterns of the room impulse responses: 12 patterns are for the state in which the interference is allocated, and the remaining pattern is for the state in which no interference exists. The transfer functions before fluctuation are used to design filters both for the acoustic echo canceller and the MOMNI method, and I evaluated the performance under static transfer functions after fluctuations. To prevent the effect of the change of condition to observe the user's utterance, I did not change the user's position in these fluctuations. A loudspeaker set in front of the user is used both as an acoustic echo canceller and as a primary sound source of the MOMNI method. The reverberation time is about 160 ms. The room impulse responses are sampled at a frequency of 48 kHz and the magnitudes are quantized to 16 bits. I used a circular array with 12 elements, and equally spaced elements were selected for use.

Conventional Acoustic Echo Canceller Our interest is focused on the robustness against the fluctuation of room transfer functions. Therefore, the experiment is carried out under the assumption that the filter coefficients of the acoustic echo canceller are once estimated precisely, and then the fluctuation occurs when the estimation stops because of barge-in. To imitate this situation, I used the transfer function before fluctuation as the estimated transfer function of the acoustic echo canceller, and fixed its filter coefficients. The microphone element closest to the user is chosen as a microphone for acquisition of the user's speech.

MOMNI method The inverse filter in the MOMNI method is calculated only using the impulse responses in the case where there is no fluctuation. The design conditions of the inverse filters are as follows: the number of secondary sound

sources $M = 4$ to 36, the number of control points $N = 3$ to 8, the filter length 16384, and the passband range 150 to 4000 Hz.

3.4.2 Evaluation score

The response sound elimination performance is evaluated using Echo Return Loss Enhancement (ERLE) as

$$\mathrm{ERLE[dB]} = 10 \log_{10} \frac{\sum_\omega \{d_{\mathrm{micref}}(\omega)\}^2}{\sum_\omega \{\epsilon(\omega)\}^2}, \tag{45}$$

where $d_{\mathrm{micref}}(\omega)$ is the response sound reproduced at a standard microphone, and $\epsilon(\omega)$ is the response sound elimination error signal derived from Eqs. (18) or (38).

3.4.3 Experimental results and discussion

Figures 6–8 show that frequency characteristics of the response sound elimination error signal in the conventional acoustic echo canceller and the MOMNI method after the room transfer function has changed. In these evaluations, I used a female utterance selected from the ASJ database [HIKT93] as a response sound. From these figures, it turns out that both of the methods can suppress the response sound independent of frequency in the passband.

The ERLE for each position of the interference in the case of the typical number of loudspeakers and 2 elements is shown in Fig. 9, and that for each position of interference in the case of 24 loudspeakers and the typical number of microphones in Fig. 10. In these evaluations, to remove the effect of the bias of frequency characteristics, I used a white noise as a response sound. It can be seen that increasing both the number of microphone elements and the number of loudspeakers improves the performance of the MOMNI method, and can make the control robust against the fluctuation of room transfer functions. Regardless of the position of the interference, the performance of the MOMNI method is superior to that of the conventional echo canceller. Hereafter, I show only the averaged ERLE of 12 types of fluctuations.

In Fig. 11, ERLE is shown as a function of the number of transfer channels ($= MK$) from the loudspeakers to the microphone elements. The theoretical

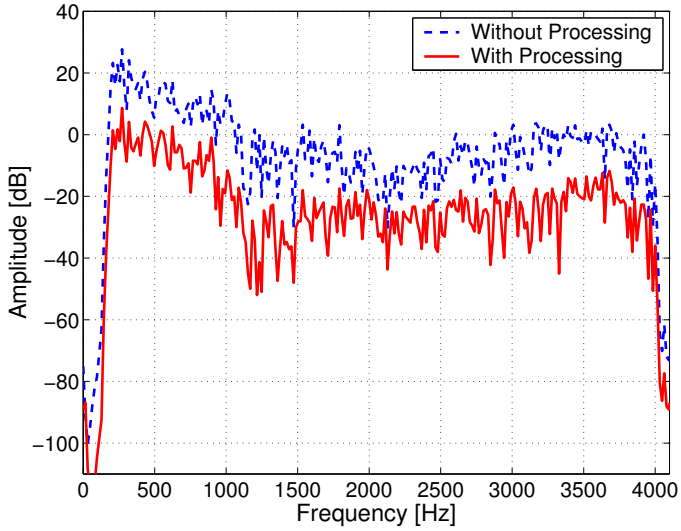

Figure 6. Example of frequency characteristics of observed signal obtained by acoustic echo canceller. The signal is observed at the microphone near the user. The position of interference is no.1 in Fig. 5.

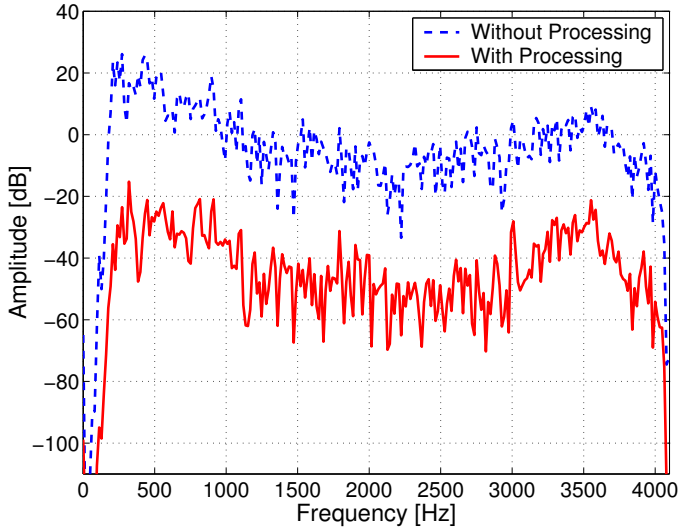

Figure 7. Example of frequency characteristics of observed signal obtained by the MOMNI method with 36 loudspeakers and 1 microphone element. The signal is observed at the microphone near the user. The position of interference is no.1 in Fig. 5.

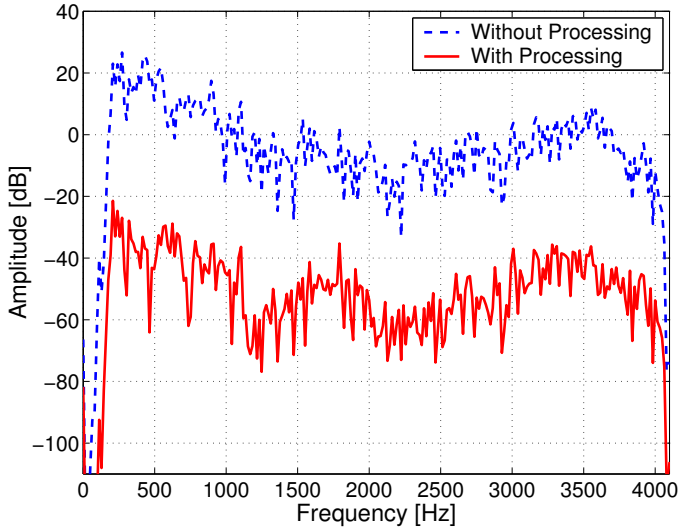

Figure 8. Example of frequency characteristics of observed signal obtained by the MOMNI method with 36 loudspeakers and 6 microphone elements. The signal is observed at the microphone near the user. The position of interference is no.1 in Fig. 5.

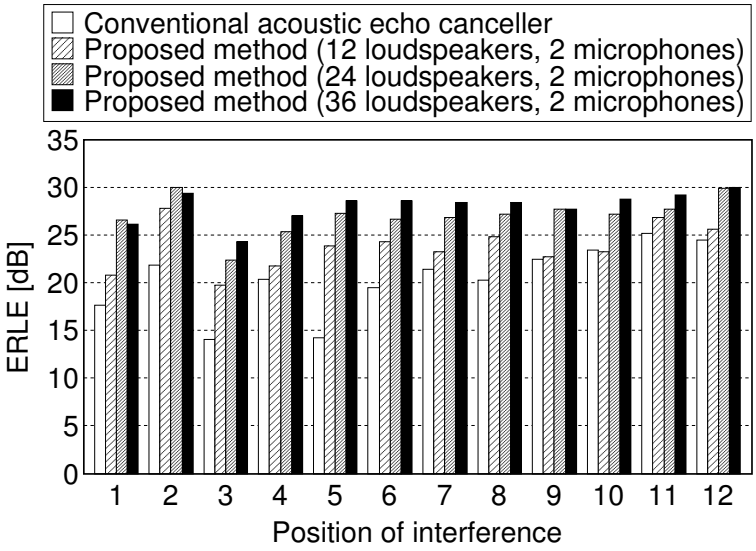

Figure 9. ERLE for each position of interference in 2 microphone elements. The horizontal axis represents the position of interference in Fig. 5.

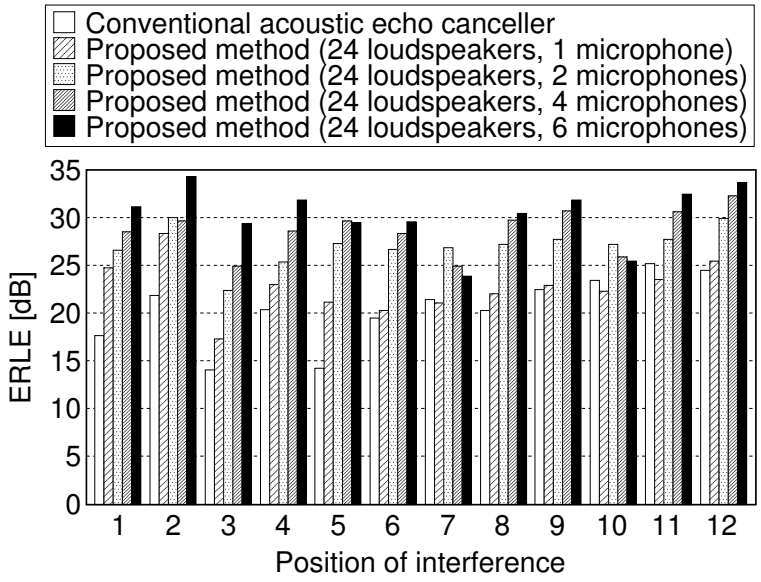

Figure 10. ERLE for each position of interference in 24 loudspeakers. The horizontal axis represents the position of interference in Fig. 5.

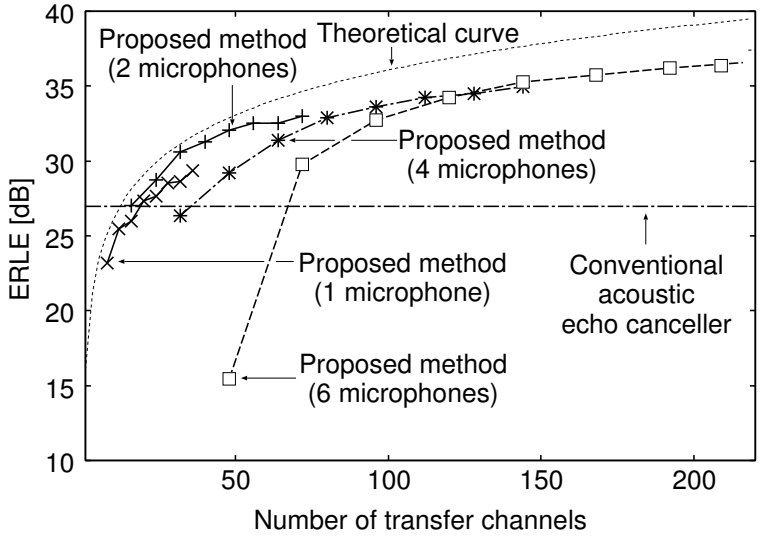

Figure 11. ERLE for different numbers of room transfer channels from loudspeakers to microphone elements.

curve in the figure is drawn by plotting the ERLE derived from Eq. (44), which is given by

$$
\begin{aligned}
\mathrm{ERLE}_{\mathrm{theory}}\ [\mathrm{dB}] &= 10\log_{10}\frac{\sum_{\omega}\{d_{\mathrm{micref}}(\omega)\}^2}{\sum_{\omega}\{\epsilon(\omega)\}^2} \\
&= 10\log_{10}\frac{\sum_{\omega}\{y_{\mathrm{mic}}(\omega)\}^2}{\sum_{\omega}\{\Delta\hat{y}_{\mathrm{mic}}(\omega)\}^2} \\
&\propto \xi + 10\log_{10}\frac{1}{1/(MK)} \\
&\propto \xi + 10\log_{10}(MK), \qquad (46)
\end{aligned}
$$

where ξ is a suitable constant.

From this figure, it can be seen that the response sound elimination performance is improved if the number of transfer channels increases. It also turns out that the deviation between the experimental and theoretical values arises when the number of microphone elements increases. The reasons are as follows:

(A) The stability margin of the inverse filters becomes small when the number of control points is close to that of the secondary sound sources.

(B) When there exist too many transfer channels, the independence of each channel is no longer valid. Consequently, the performance is saturated.

To prove the above claim (A), I show the condition number of transfer functions in Fig. 12. The condition number, expressed as $\mathrm{cond}(\boldsymbol{G}(\omega))$ in Eq. (41), represents the unstableness of the inverse filters. This figure shows that the condition number becomes close to 1 when the number of loudspeakers is much larger than that of the microphone elements (equal to the number of control points minus two), as argued in Chapter 3.3.4. However, when the number of microphone elements increases, the condition number increases. In addition, such a tendency becomes remarkable when the number of the secondary sound sources is small. This causes an appreciable degradation in ERLE.

Comparing the conventional acoustic echo canceller with the MOMNI method in Fig. 11, it can be seen that the MOMNI method is more robust against the fluctuation of transfer functions if the number of transfer channels increases.

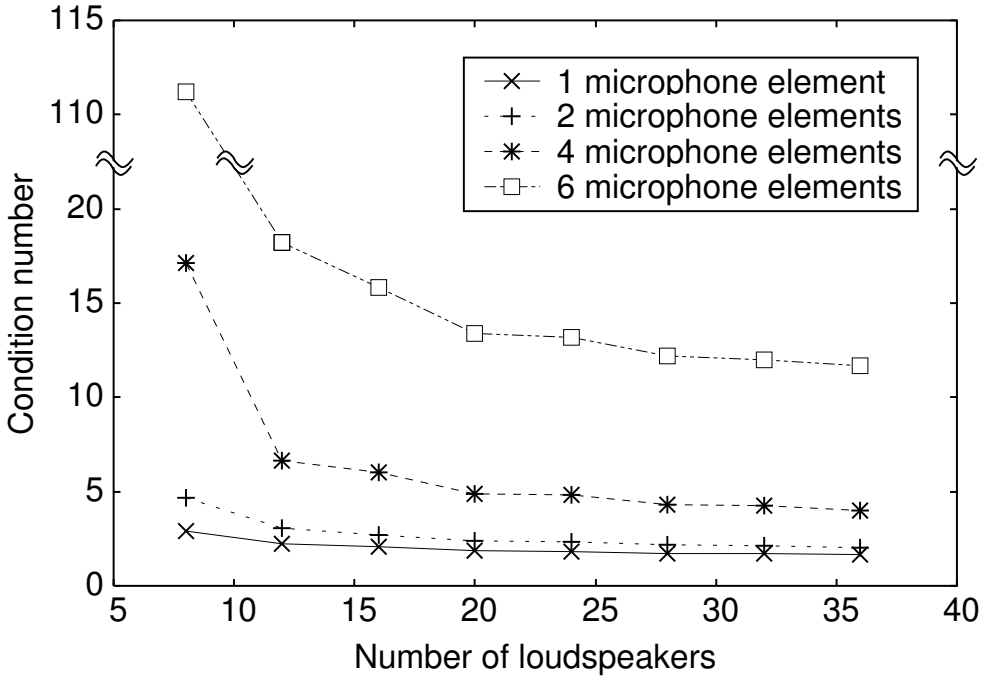

Figure 12. Condition number of average in passband.

3.5 Speech recognition experiment

The experiment involving large-vocabulary speech recognition is carried out to investigate the efficacy of the MOMNI method, compared to that of the conventional acoustic echo canceller.

3.5.1 Experimental conditions

In the recognition experiment, I use the speech signal obtained by imposing the response sound elimination error signal $\epsilon(\omega)$ on the user's input speech. A large-vocabulary recognition engine Julius ver. 3.4.2 [LKK01] is used as a speech decoder. I used two kinds of speaker-independent phonetic tied mixtures [LKTS00] as phoneme models. One is an ordinary clean model. The other is generated by a known-noise imposition technique [YLSS03] (see Appendix). I imposed a known noise of 30 dB on the observed signals to mask the redundant response sound, and to match its phoneme features, and imposed the noise of 25 dB on

the speech in the learning data. A language model is made from newspaper dictation with a vocabulary of 20,000 words [IYT+98]. As the user's speech, 200 sentences obtained from 23 males and 23 females are used through the JNAS database [IYT+99]. As the response sound of the dialogue system, a sentence of a female's speech from the ASJ database is used. Experimental conditions such as interference arrangements to cause changes of the transfer functions are the same as in the previous section.

3.5.2 Evaluation score

In order to evaluate the speech recognition performance, I adopt the word accuracy (WA) as an evaluation score. Word accuracy is defined as follows:

$$\text{WA } [\%] = \frac{W - S - D - I}{W}, \tag{47}$$

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. The resultant recognition score is computed using the average value of data derived from the 200 sentences.

3.5.3 Experimental results and discussions

The speech recognition results obtained by the MOMNI method are shown in Fig. 13 for the clean model, and in Fig. 14 for the known noise imposition. The results of the recognition experiment show that the word accuracy is -8.0% and -13.2% without any processing, and 47.1% and 64.6% when using the conventional acoustic echo canceller, for the clean model and known-noise imposition, respectively. By masking the redundant component of the response sound, all the results are improved compared with the results with the clean model. All the performances of the MOMNI method in the figure are superior to those of the conventional acoustic echo canceller. Note that neither system is adapted, i.e. optimal weights for system before acoustic change is used. The results show that when the transfer functions are changed, the degradation of speech recognition accuracy can be prevented by increasing the number of transfer channels. From these results, the effectiveness of the MOMNI response sound elimination technique is ascertained.

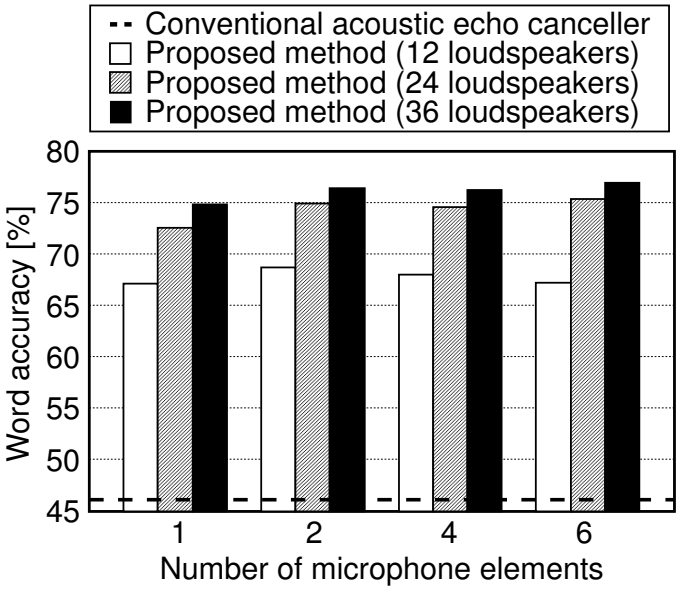

Figure 13. Word accuracy with clean model.

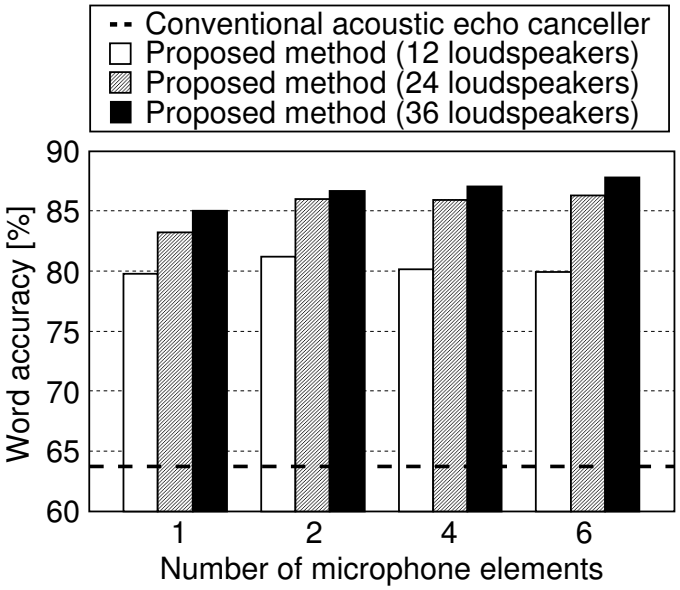

Figure 14. Word accuracy when known-noise imposition technique is applied.

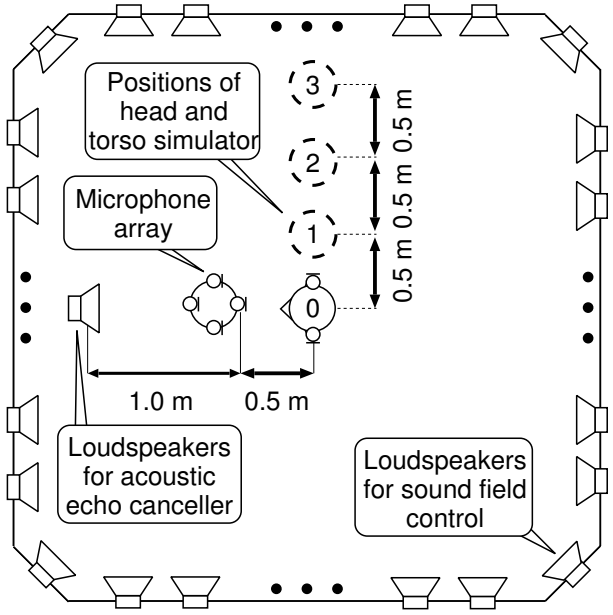

Figure 15. Layout of the experimental room in the sound quality assessment.

3.6 Sound quality assessment at various user positions

The sound quality of the MOMNI method is guaranteed and clear sound image is presented only when the user's ears are at the control points where the response sound is reproduced. However, even when the user moves away from the controlled area, the quality of the response sound is sufficient for the spoken dialogue system. To prove this argument, I assess the quality of the response sound which is perceived by the user at various positions. The quality is assessed from two aspects; objective and subjective evaluations.

3.6.1 Objective evaluation

The objective evaluation is carried out via a simulation using impulse responses measured in a real acoustic environment. Figure 15 shows the arrangement of the apparatuses. The room is the same one used in the experiments of Chapters 3.4 and 3.5. I measured four patterns of impulse responses changing the positions of the HATS from position 0 to position 3. The control points of the the MOMNI method are two microphone elements in the microphone array and the ears of the

HATS at the position 0. The primary sound source of the response sound is the loudspeaker of the acoustic echo canceller.

As an evaluation score, I introduce cepstral distance (CD [RJ93]), which is often used in various speech processing. CD is given by

$$\text{CD [dB]} = \frac{1}{F}\sum_{t=1}^{F}\frac{20}{\log 10}\sqrt{\sum_{l=1}^{20}2(C_{\text{obs}}(l,t)-C_{\text{ref}}(l,t))^2}, \tag{48}$$

where F denotes the number of speech frames, $C_{\text{obs}}(l,t)$ is the l-th FFT-based cepstrum of the observed signal at the t-th frame, and $C_{\text{ref}}(l,t)$ is a reference cepstrum for evaluating the distance. The number of liftering points is 20. A smaller CD value indicates better sound quality. The reference cepstrum $C_{\text{ref}}(l,t)$ is obtained from the source signal of the response sound. I average the CDs at both ears. Note that to express CD in dB the term $20/\log 10$ is multiplied to the Euclidean distances between the cepstrum coefficients which are obtained from natural logarithm of the waveforms. In addition, because of symmetry of cepstrum coefficients, liftered cepstrum is obtained from twice of the cepstrum coefficients from $l = 1$ to $l = 20$.

Figures 16 and 17 show the CDs of the MOMNI method compared with those of the acoustic echo canceller. Since the MOMNI method reproduces the output sound of the acoustic echo canceller at the position 0, its CD is similar to that of the acoustic echo canceller. When the HATS is not at the position 0, the CDs increase. However, its difference is only within 1 dB. Thus, the sound-quality degradation of the MOMNI method is not significant.

3.6.2 Subjective evaluation

To ascertain that the distortion caused by the MOMNI method is not discomfort, I conduct a subjective evaluation of the sound quality reproduced by the MOMNI method in a real environment. I changed the positions of the subjects and let them answer mean opinion score (MOS). The opinion score for evaluation was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad).

The room used in this experiment is the same one where the impulse responses are measured in the other experiments. I directed the positions of the subjects by setting chairs at the position 0, the position 1 and the position 2 in the

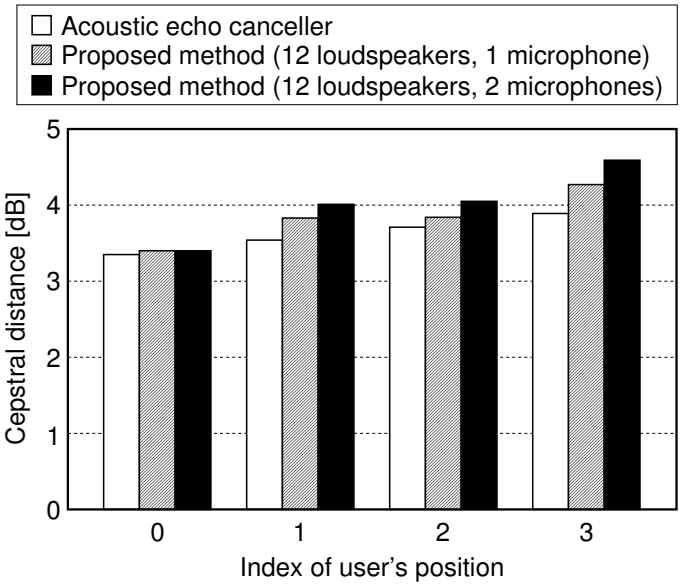

Figure 16. Cepstral distance in various positions when 12 loudspeakers are used for the MOMNI method.

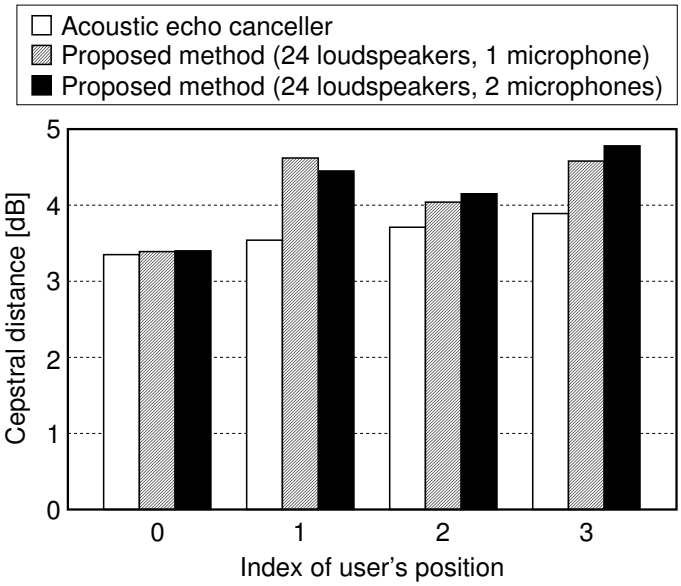

Figure 17. Cepstral distance in various positions when 24 loudspeakers are used for the MOMNI method.

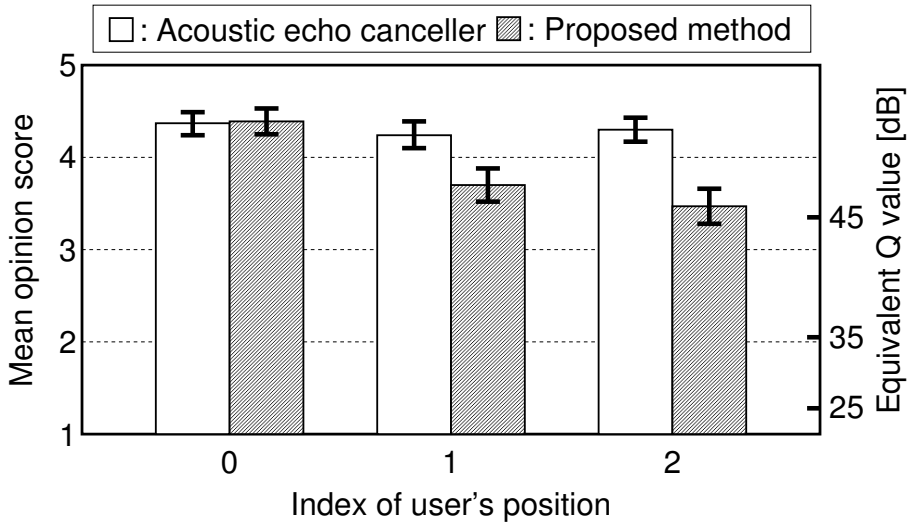

Figure 18. Mean opinion score for the positions of the subjects. The blocks show the means and the error bars show the 95% confidence intervals.

Fig. 15. The filter of the MOMNI method was designed using measured impulse responses where the HATS is set at the position 0. The primary sound source of the response sound is the loudspeaker of the acoustic echo canceller. The number of the secondary sound sources are 24 and the microphone elements of the silent reproduction is two.

I compared the MOSs of the MOMNI method and the acoustic echo canceller. In addition, to give the MOSs objective meaning, I evaluated opinion equivalent Q value [DHP93]. To obtain opinion equivalent Q value, I made three kinds of response sound imposed white noises whose segmental SNRs are 25 dB, 35 dB and 45 dB. Then these noise-added response sound is outputted from the acoustic echo canceller. Therefore, the forms of the reproductions are five, i.e., the MOMNI method, the acoustic echo canceller and the three noise-added response sound. For each of these forms, I prepared 15 sentences of the speech uttered by four males and three females. Then for each of the three positions, I evaluated the MOSs in random orders.

Figure 18 shows the MOSs for each of the subjects' positions. The scores of the acoustic echo canceller rated at more than four four in any of the positions. For the MOMNI method, the score at the position 0 is similar to that of the acoustic

echo canceller. Even at the position 0, the binaural response sound is degraded by the difference of the shapes of the head and the sitting heights between the subjects and the HATS. However, it can be seen that the degradation does not influence the MOSs. Although the MOSs decrease as the subjects move away from the position 0, the degradation of the score is within one. In addition, even in the worst score at the position 3, the opinion equivalent Q value is over 45 dB. From these findings, it is ascertained that the MOMNI method can present the response sound with sufficient quality even when the user is out of the prepared position.

3.7 Conclusion

A barge-in robust spoken dialogue interface combining sound field reproduction and beamforming called MOMNI method is proposed. It is shown that the response sound elimination performance for the fluctuation of room transfer functions depends on the number of transfer channels. By using an adequate number of loudspeakers and microphone elements, the performance of the MOMNI method is better than that of the conventional acoustic echo canceller. In the experiment where the MOMNI method is compared with acoustic echo canceller in the condition that the filter coefficients are fixed, the efficacy of the MOMNI method is ascertained. Although the MOMNI method requires multichannel filtering and multiple loudspeakers, the MOMNI method can maintain the high speech recognition performance in barge-in situation without adaptation.

4. Response Sound Elimination Based on Null-Space Based Sound Field Control

4.1 Introduction

In this chapter, I propose small-scale variation of the sound field control to eliminate the response sound at the microphones. The MOMNI method can make control of silence by increasing the number of loudspeakers, as discussed in Chapter 3.3.4. However, a large number of loudspeakers are needed to achieve sufficient robustness for speech recognition. Furthermore, the MOMNI method reproduces the response sound 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, I 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, called nullspace-based sound field control (NBSFC). 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 NBSFC is more robust against fluctuation of the room transfer function than the conventional methods even when there are few loudspeakers. However, although the NBSFC with many microphone elements improves its speech recognition performance, the quality of response sound is speculated to slightly degrade. I discuss the trade-off between speech recognition performance and sound quality in our sound quality assessment experiment.

The outline of this chapter is as follows: The proposed filter design method is described in Chapter 4.2. The performance of the NBSFC method is discussed in Chapter 4.3. Finally, the conclusion of this study is provided in Chapter 4.4.

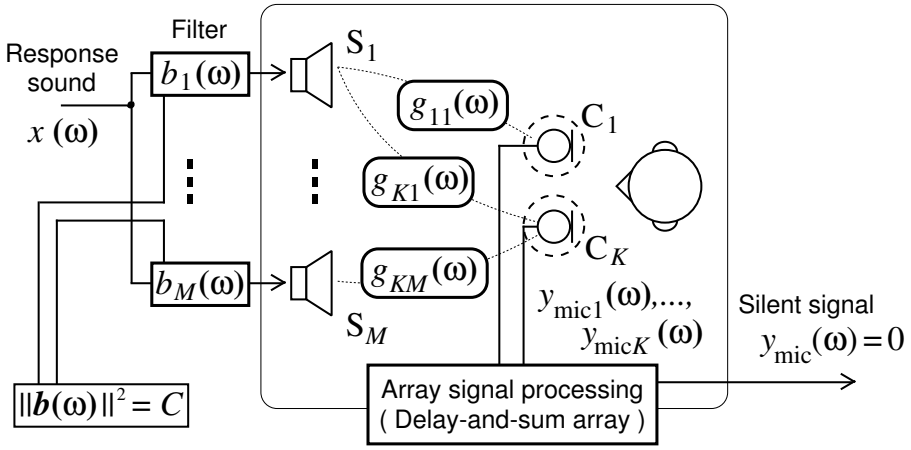

Figure 19. Configuration of NBSFC.

4.2 Response sound cancellation based on nullspace-based sound field control

In this section, I 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.

4.2.1 Sound field control for cancelling out response sound

In Fig. 19, $\mathrm{S}_m (m = 1, \ldots, M)$ 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

$$M > K. \tag{49}$$

The observed signals at the control points are designated as

$$\boldsymbol{d}(\omega) = [d_1(\omega), \ldots, d_K(\omega)]^{\mathrm{T}}, \tag{50}$$

where $d_k(\omega)$ $(k = 1, \ldots, K)$ are the signals observed at the microphones C_k. The response sound is monaural and denoted by a scalar $r_{\mathrm{src}}(\omega)$. The response sound

is outputted from the loudspeakers after being processed by filters. The filter coefficients are represented by

$$\boldsymbol{b}(\omega) = [b_1(\omega), \ldots, b_M(\omega)]^{\mathrm{T}}, \tag{51}$$

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

$$\boldsymbol{G}(\omega) = \begin{bmatrix} g_{11}(\omega) & \ldots & g_{1M}(\omega) \\ \vdots & \ddots & \vdots \\ g_{K1}(\omega) & \ldots & g_{KM}(\omega) \end{bmatrix}, \tag{52}$$

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

$$\boldsymbol{d}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{b}(\omega)r_{\mathrm{src}}(\omega). \tag{53}$$

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

$$\boldsymbol{G}(\omega)\boldsymbol{b}(\omega) = \boldsymbol{0} \tag{54}$$

$$\text{subject to} \quad ||\boldsymbol{b}(\omega)|| = C \tag{55}$$

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.

4.2.2 Extracting vectors that span nullspace

Equation (54) 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

$$\boldsymbol{G}(\omega) = \boldsymbol{U}(\omega) \left[\boldsymbol{\Gamma}(\omega) \;\middle|\; \boldsymbol{O}_{K,M-K} \right] \boldsymbol{V}^{\mathrm{H}}(\omega), \tag{56}$$

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

$$\boldsymbol{\Gamma}(\omega) = \begin{bmatrix} \mu_1 & & & & & \\ & \ddots & & & & \\ & & \mu_{R_\omega} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}. \tag{57}$$

R_ω is the rank of $\boldsymbol{G}(\omega)$, and its diagonal elements $\{\mu_1(\omega), \ldots, \mu_{R_\omega(\omega)}, 0, \ldots, 0\}$ 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 $\{\mu_1(\omega), \ldots, \mu_{R_\omega(\omega)}\}$, 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 needed. 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. (54) as shown by

$$\begin{aligned} \boldsymbol{G}(\omega)\boldsymbol{v}_r(\omega) &= \boldsymbol{U}(\omega)\left[\boldsymbol{\Gamma}(\omega), \boldsymbol{O}_{K,M-K}\right] \\ &\quad \cdot \boldsymbol{V}^{\mathrm{H}}(\omega)\boldsymbol{v}_r(\omega) \\ &= \boldsymbol{U}(\omega)\left[\boldsymbol{\Gamma}(\omega), \boldsymbol{O}_{K,M-K}\right] \\ &\quad \cdot [\underbrace{0, \ldots, 0}_{r-1}, 1, \underbrace{0, \ldots, 0}_{M-r}]^{\mathrm{T}} \\ &= \boldsymbol{U}(\omega)\boldsymbol{0} \\ &= \boldsymbol{0}. \end{aligned} \tag{58}$$

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

$$\boldsymbol{b}'(\omega) = \sum_{r=R_\omega+1}^{M} \alpha_r(\omega)\boldsymbol{v}_r(\omega), \tag{59}$$

where α_r $(r = R_\omega + 1, \ldots, M)$ are arbitrary complex values. This vector $\boldsymbol{b}'(\omega)$ satisfies Eq. (54) as shown by

$$\begin{aligned} \boldsymbol{G}(\omega)\boldsymbol{b}'(\omega) &= \sum_{r=R_\omega+1}^{M} \alpha_r(\omega)\,\boldsymbol{G}(\omega)\boldsymbol{v}_r(\omega) \\ &= \sum_{r=R_\omega+1}^{M} \alpha_r(\omega)\mathbf{0} \\ &= \mathbf{0}. \end{aligned} \tag{60}$$

In another denotation,

$$\boldsymbol{b}'(\omega) = \boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega), \tag{61}$$

where

$$\boldsymbol{W}(\omega) = \left[\boldsymbol{v}_{R_\omega+1}(\omega), \ldots, \boldsymbol{v}_M(\omega)\right], \tag{62}$$

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

$$\boldsymbol{\alpha}(\omega) = \left[\alpha_{R_\omega+1}, \ldots, \alpha_M\right]^{\mathrm{T}}. \tag{63}$$

Since $\boldsymbol{b}'(\omega)$ does not satisfy the norm condition of Eq. (55), the resultant filter $\boldsymbol{b}(\omega)$ is obtained by normalizing $\boldsymbol{b}'(\omega)$ with its norm, as

$$\begin{aligned} \boldsymbol{b}(\omega) &= C\frac{\boldsymbol{b}'(\omega)}{\sqrt{\|\boldsymbol{b}'(\omega)\|^2}} \\ &= C\frac{\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega)}{\sqrt{\boldsymbol{\alpha}^{\mathrm{H}}(\omega)\,\boldsymbol{W}^{\mathrm{H}}(\omega)\,\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega)}}. \end{aligned} \tag{64}$$

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

$$\boldsymbol{v}_i^{\mathrm{H}}(\omega)\boldsymbol{v}_j(\omega) = \begin{cases} 1 & \text{for} \quad i = j; \\ 0 & \text{for} \quad i \neq j. \end{cases} \tag{65}$$

Accordingly,

$$\boldsymbol{W}^{\mathrm{H}}(\omega)\,\boldsymbol{W}(\omega) = \boldsymbol{I}_{M-R_\omega}, \tag{66}$$

where $\boldsymbol{I}_{M-R_\omega}$ denotes $(M - R_\omega) \times (M - R_\omega)$ identity matrix. The substitution of Eq. (66) in Eq. (64) leads to

$$\boldsymbol{b}(\omega) = C\frac{\boldsymbol{W}(\omega)\boldsymbol{\alpha}(\omega)}{\sqrt{\boldsymbol{\alpha}^{\mathrm{H}}(\omega)\boldsymbol{\alpha}(\omega)}}. \tag{67}$$

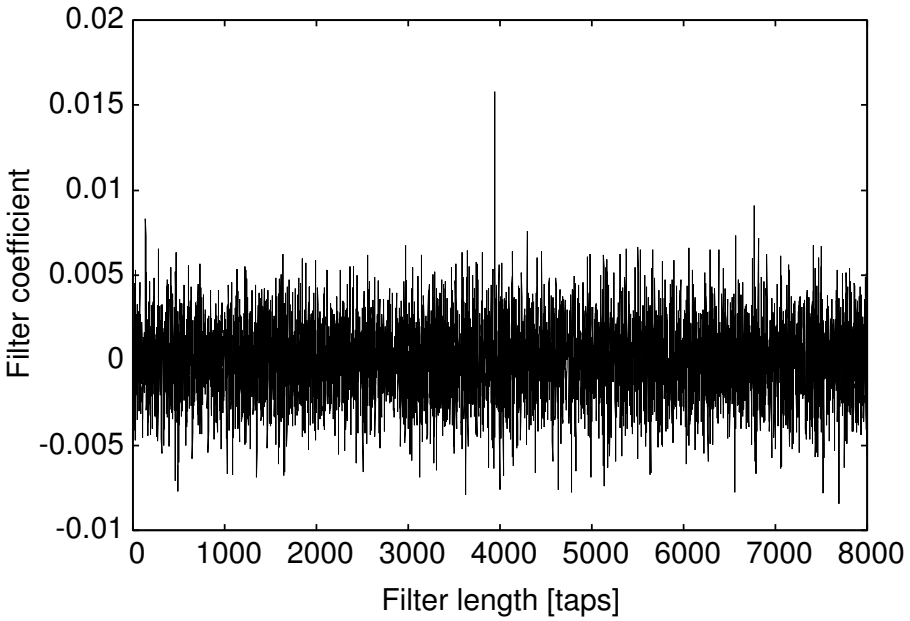

Figure 20. Example of waveform of filter coefficients designed by random summation of the nullspace vectors. The filter is designed with eight loudspeakers and two microphones, and corresponds to one loudspeaker. The filter is designed with FFT points of 16384, and cut by a rectangle window of 8000 points. The sampling frequency is 16 kHz and the bandwidth is 150–4000 Hz.

4.2.3 Filter coefficients closest to impulses

In the previous section, I showed that the conditions of Eqs. (54) and (55) are satisfied by $\boldsymbol{b}(\omega)$ in Eq. (67), 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. 20 shows an example of the filter designed by random summation of the nullspace vectors, where a large undesired pre/post-echo can be seen. In the following, I propose an algorithm for designing a filter with a small distortion by utilizing the solution closest to impulses.

I 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

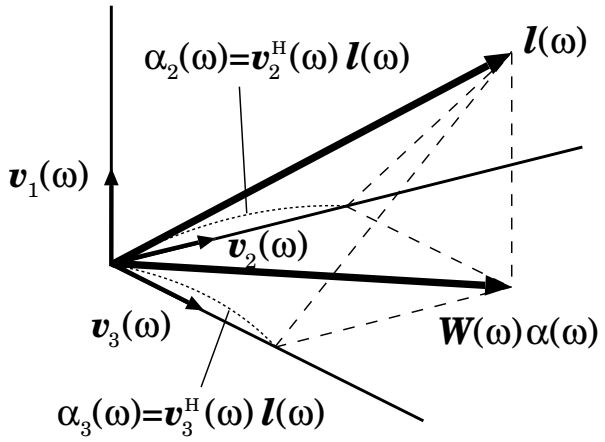

Figure 21. Example of the projection of $\boldsymbol{l}(\omega)$ to the nullspace of $\boldsymbol{G}(\omega)$ when the row space of $\boldsymbol{G}(\omega)$ is spanned by $\boldsymbol{v}_1(\omega)$, and $\boldsymbol{W}(\omega) = [\boldsymbol{v}_2(\omega)\ \boldsymbol{v}_3(\omega)]$.

τ.

$$\boldsymbol{l}(\omega) = e^{j\omega\tau}[\underbrace{1, \ldots, 1}_{M}]^{\mathrm{T}} \tag{68}$$

Then I 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. The optimal expanded coefficient vector $\boldsymbol{\alpha}(\omega)$ can be obtained by solving the following least squares problem:

$$\min_{\boldsymbol{\alpha}(\omega)} \| \boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{\alpha}(\omega) - \boldsymbol{l}(\omega)\|^2. \tag{69}$$

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. 21. Such expanded coefficients $\alpha_r(\omega)$ $(r = R_\omega + 1, \ldots, M)$ that satisfy Eq. (69) can be given by the inner product of $\boldsymbol{v}_r(\omega)$ and $\boldsymbol{l}(\omega)$ as

$$\alpha_r(\omega) = \frac{\boldsymbol{v}_r^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega)}{\boldsymbol{v}_r^{\mathrm{H}}(\omega)\boldsymbol{v}_r(\omega)} = \boldsymbol{v}_r^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega). \tag{70}$$

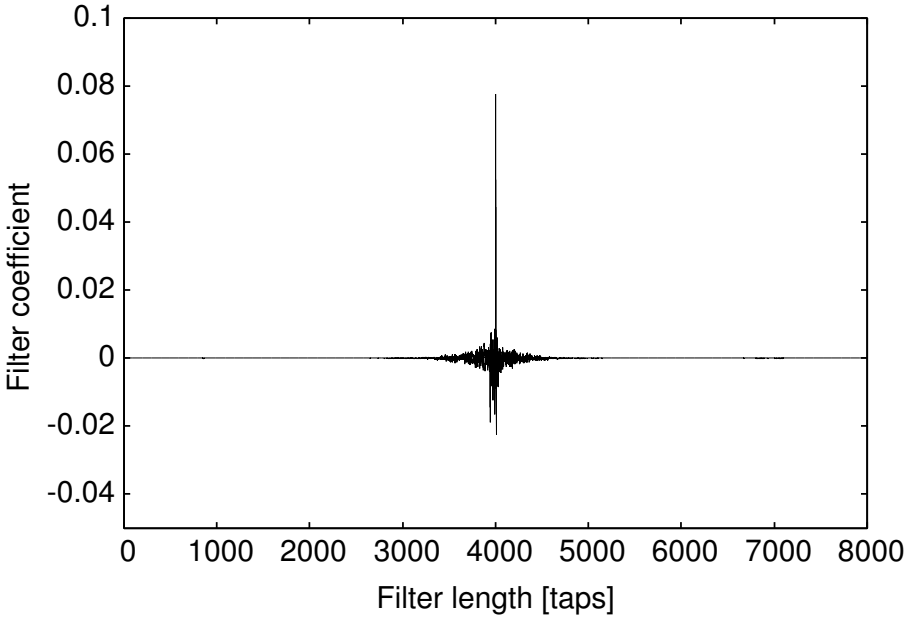

Figure 22. Example of waveform of filter coefficients designed by NBSFC. The filter is designed with eight loudspeakers and two microphones, and corresponds to one loudspeaker. The filter is designed with FFT points of 16384, and cut by a rectangle window of 8000 points. The sampling frequency is 16 kHz and the bandwidth is 150–4000 Hz.

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

$$\boldsymbol{\alpha}(\omega) = \begin{bmatrix} \boldsymbol{v}^{\mathrm{H}}_{R_\omega+1}(\omega) \\ \vdots \\ \boldsymbol{v}^{\mathrm{H}}_{M}(\omega) \end{bmatrix} \boldsymbol{l}(\omega) = \boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega). \tag{71}$$

Then the resultant filter coefficients $\boldsymbol{b}(\omega)$ is obtained by substituting Eq. (71) in Eq. (67) as

$$\boldsymbol{b}(\omega) = C\frac{\boldsymbol{W}(\omega)\,\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega)}{\sqrt{\boldsymbol{l}^{\mathrm{H}}(\omega)\,\boldsymbol{W}(\omega)\,\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{l}(\omega)}}. \tag{72}$$

Figure 22 shows an example of a filter designed by the NBSFC. It can be seen that its distortion is considerably lower than that in Fig. 20.

4.2.4 Response sound elimination error when changing room transfer functions

In this section I describe the theoretical estimation of error caused by fluctuation in the NBSFC, as shown for the MOMNI method in Chapter 3.3.4. 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, 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

$$\Delta \boldsymbol{g}_k(\omega) = \sum_{m=1}^{M} \phi_{km}(\omega) \boldsymbol{b}_m^{\mathrm{H}}(\omega), \tag{73}$$

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

$$\begin{aligned}
\epsilon(\omega) &= \frac{1}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} A_k(\omega) \phi_{km}(\omega) \boldsymbol{b}_m^{\mathrm{H}}(\omega) \boldsymbol{b}(\omega) \\
&= \frac{C}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} \phi_{km}(\omega) \boldsymbol{b}_m^{\mathrm{H}}(\omega) \boldsymbol{b}_1(\omega) e^{-j\omega\tau_k} \\
&= \frac{C}{K} \sum_{k=1}^{K} \phi_{k1}(\omega) e^{-j\omega\tau_k} \\
&\propto \frac{1}{\sqrt{K}}.
\end{aligned} \tag{74}$$

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 NBSFC 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 NBSFC, as can be seen in Eqs. (22) and (49), because of the sound field reproduction at the user's ears. The second reason is that the condition in Eq. (44) 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

Table 5. Computational cost of the AEC and the MOMNI method.

Method	Order	Typical values
Single-channel AEC	$3B \log_2 2N$	528 ($B = 8, N = 1024$)
NBSFC	$M + 1 \log_2 2N + M$	134 ($M = 8, N = 8192$)

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 NBSFC performs better than the MOMNI method when the number of loudspeakers is small.

4.2.5 Computational complexity

Since the input signal is single channel, the computation is simpler than the MOMNI method. We list the computational cost of the AEC and the NBSFC in Table 5. Similarly to the discussion in Chapter 3.3.5, single-channel FFT of the input signal, M-channel filtering in the frequency domain and M-channel IFFT of the loudspeaker outputs require $(M + 1) \log_2 2N_{\mathrm{NB}} + M$ multiplications per sample where N_{NB} is the filter length in taps. Thus, typical setting with $M = 8$, $N_{\mathrm{NB}} = 8192$ requires 134 multiplications per sample. The single-channel AEC with filter length of N_{AEC} taps and B block overlaps requires $log_2 2N_{\mathrm{AEC}}$ multiplications per sample; 264 multiplications per sample with typical values $B = 8$ and $N_{\mathrm{AEC}} = 1024$. Thus the computational complexity of the NBSFC is lower than that of the AEC.

4.3 Experiments and results

In this section, I present two experiments in which the conventional methods (an acoustic echo canceller and the MOMNI method) and the NBSFC are compared. To validate the robustness of the NBSFC against the fluctuation of the room transfer functions, a response sound elimination experiment is conducted where changes in the transfer functions are simulated. Then the sound quality of the response sound is evaluated. Subsequently, the performance of each method on

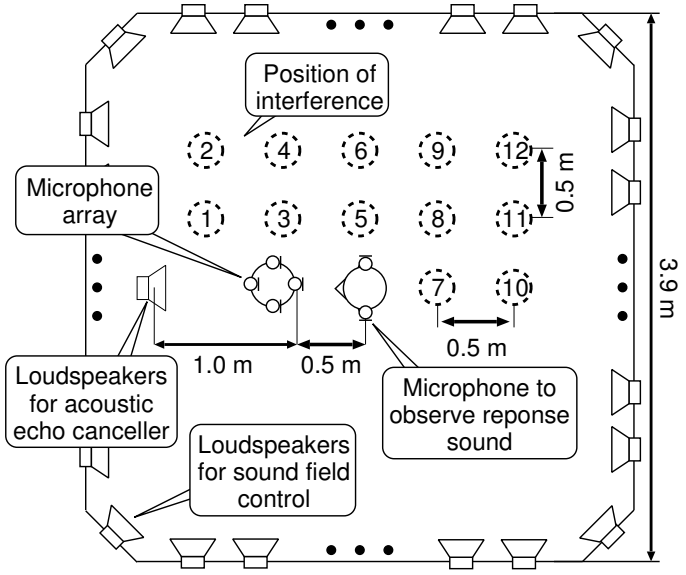

Figure 23. Layout of acoustic experiment room.

the basis of a speech recognition experiment is evaluated to verify the applicability of the NBSFC. Finally, the sound quality of these methods is assessed.

4.3.1 Experimental conditions

In the experiments, it is premised that the fluctuation of transfer functions is caused by changes in positions of an interference, 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. I 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. I used impulse responses without the mannequin as those before fluctuation, and I evaluated the average performance in 12 types of fluctuation. Figure 23 shows the arrangement of the apparatuses. As shown in Fig. 23, I place a dummy head, which has an average human head and an upper body, at the user's position. I designed the filters used in the MOMNI method and NBSFC with room transfer functions before fluctuation. I gave the acoustic echo canceller the room transfer function before fluctuation as its filter coefficients,

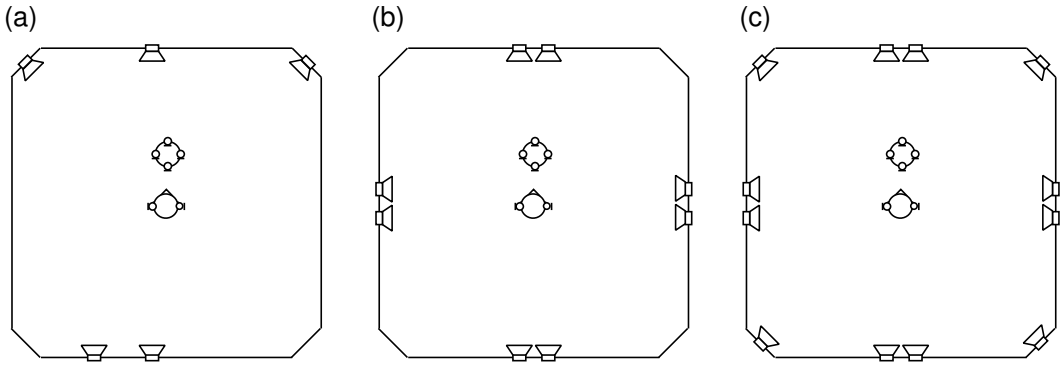

Figure 24. Exact locations of loudspeakers when (a) five, (b) eight and (c) twelve loudspeakers are used.

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. I evaluated the performances with the average of 12 kinds of impulse responses with the mannequin.

The impulse responses used in this experiment are 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 the MOMNI method and the NBSFC 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, I label the transfer system "M-K system"), are designed. The exact locations of the loudspeakers are shown in Fig. 24. The passband range is 150–4000 Hz. I use a circular microphone array with 12 elements and select the elements that are spaced equally.

Table 6. 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 7. BRRs [dB] of the NBSFC 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

4.3.2 Response sound elimination experiment

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

$$\mathrm{BRR} = 10\log_{10}\frac{\sum_{\omega}|y_{\mathrm{ear}}(\omega)|^2}{\sum_{\omega}|y_{\mathrm{out}}(\omega)|^2} \quad [\mathrm{dB}], \tag{75}$$

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 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, I use a female utterance selected from the ASJ database [HIKT93]. Although the sampling frequency of the response sound is 16 kHz, I use the signal in which the frequency components beyond 4 kHz are eliminated.

I show the BRRs of the MOMNI method and the NBSFC before fluctuation in Tables 6 and 7. 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).

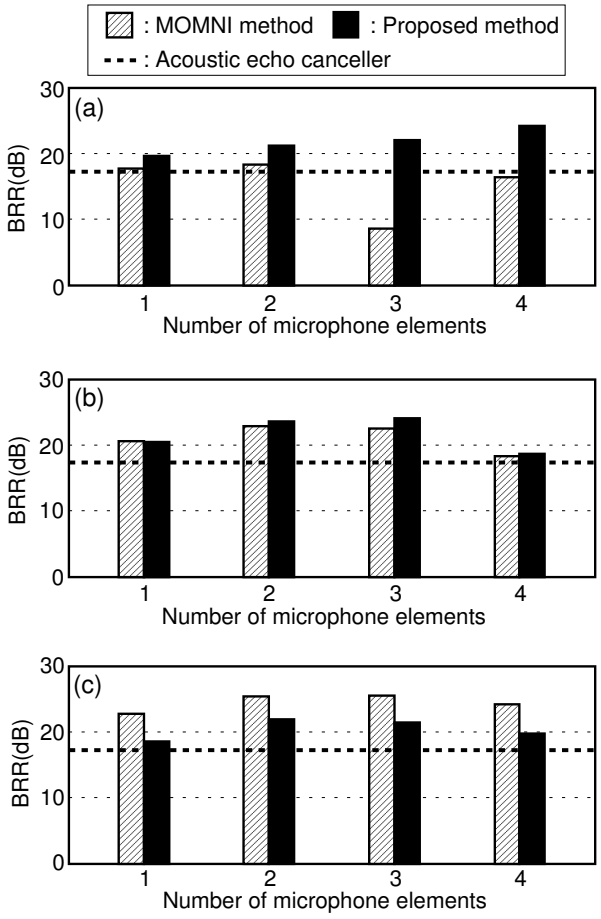

Figure 25. Comparison of BRR for M loudspeakers: (a) $M = 5$, (b) $M = 8$ and (c) $M = 12$.

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 6 and 7, they degrade after fluctuation. Figure 25 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 NBSFC shows a higher performance than the acoustic echo canceller in all combinations. For the 5 loudspeakers, the proposed NBSFC shows a higher performance than the MOMNI method. In particular, the NBSFC 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 NBSFC can not be seen. This is because the performance of the NBSFC is independent of the number of loudspeakers as discussed in Chapter 4.2.4. Thus, the NBSFC is highly beneficial for application to a few loudspeakers.

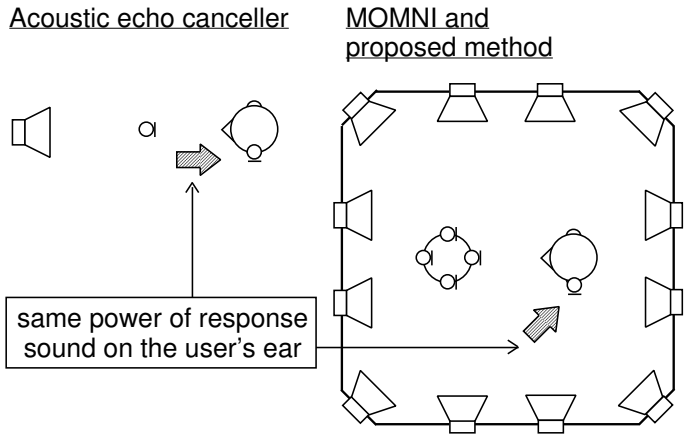

Figure 26. Configuration of the speech recognition.

4.3.3 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, I adopt word accuracy (WA) shown in Eq. (47), where I average each WA obtained from 200 utterances.

I show the configuration of the condition of the speech recognition in Fig. 26. 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

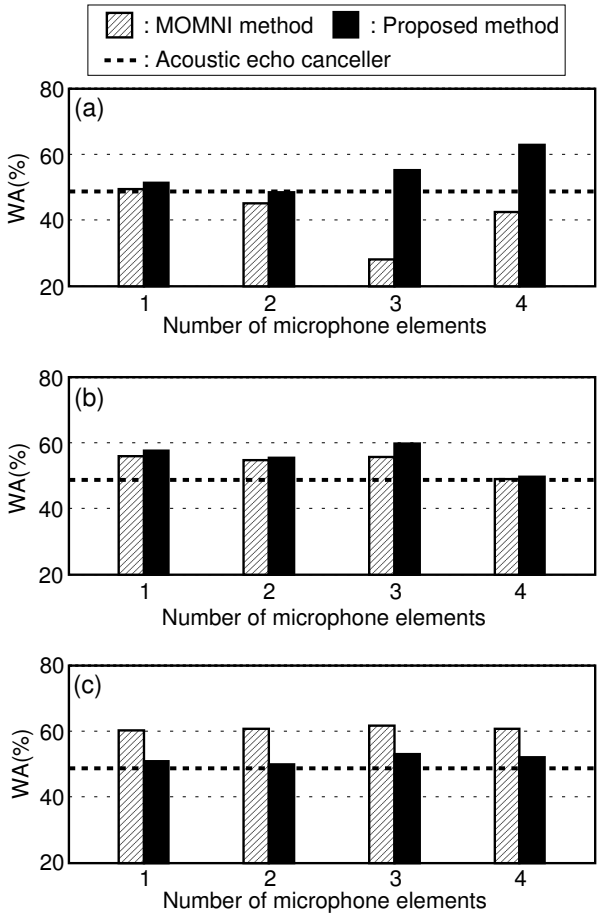

Figure 27. Comparison of WA with clean model for M loudspeakers: (a) $M = 5$, (b) $M = 8$ and (c) $M = 12$.

speech at the microphone is set to 0 dB. In the MOMNI method and NBSFC, I 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. I use two speaker-independent 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 27 and 28 show the WA for all the combinations. Figure 27 shows the speech recognition performance with a clean model, and Fig. 28 shows that with known-noise imposition. All

the scores are similar to those in Fig. 25, 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 the speech recognition performance considerably. The NBSFC for the 5-4 system with known-noise imposition has the highest score of 76.1%. Using the NBSFC together with known-noise superposition, an improvement of 27.3% over the performance of the conventional acoustic echo canceller with clean speech is achieved.

4.3.4 Sound quality assessment

Although the NBSFC 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. (55)). In this section, I assess the quality of the response sound produced by the NBSFC. I evaluate the sound quality against the user's movement. I show the index of the user's position in Fig. 29. I apply CD shown in Eq. (48) as an evaluation score. I average the CDs at both ears. The less CD shows the better sound quality.

Figure 30 shows the CDs between the observed signals at the user's position and original response sound signal (dry source). The NBSFC 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 Chapter 4.2.3.

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.

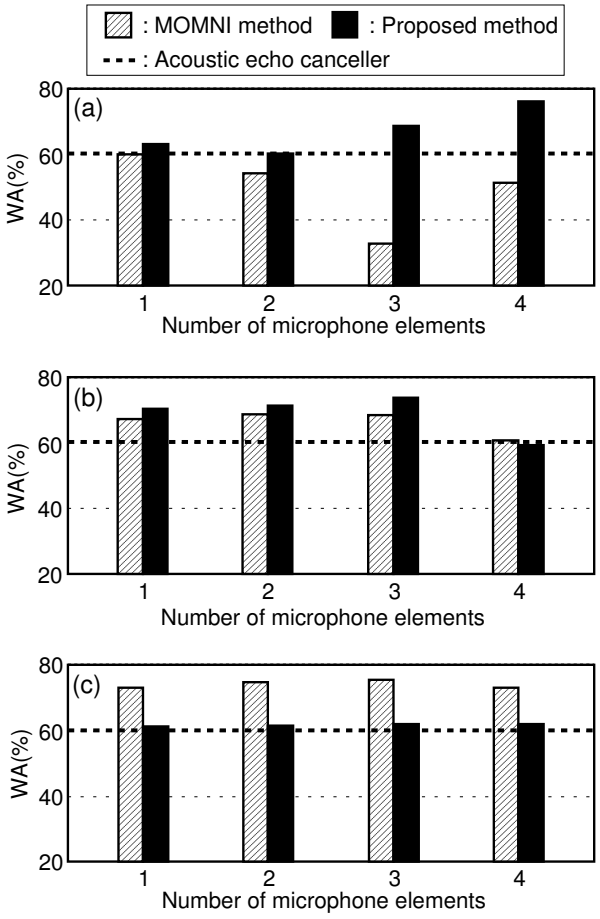

Figure 28. Comparison of WA with model imposed 25 dB known-noise in the case of M loudspeakers: (a) $M = 5$, (b) $M = 8$ and (c) $M = 12$.

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

In the NBSFC, 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 NBSFC has no specific control points where sound quality is high, the distortion of the output signal from filter

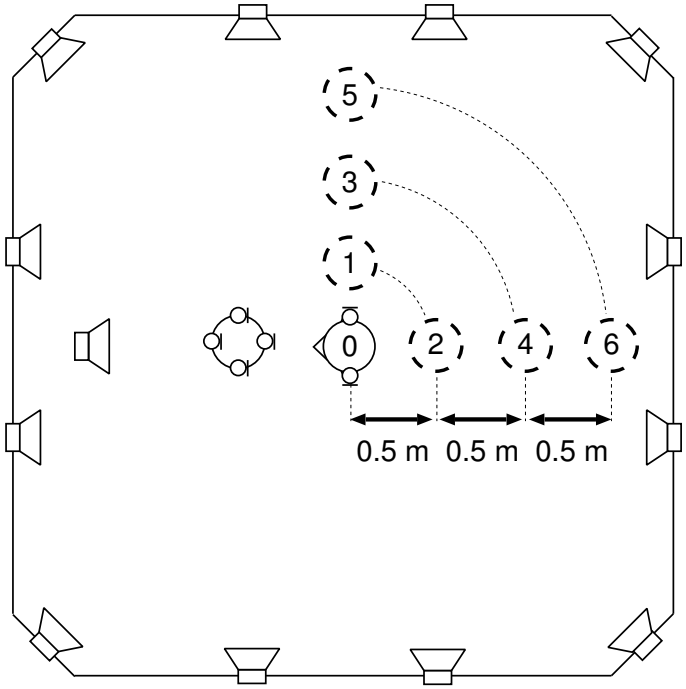

Figure 29. Configuration of user's movement. The symbol "0" indexes the position where the MOMNI method presents the response sound. Symbols 1, 3 and 5 index the positions 0.5 m, 1.0 m and 1.5 m right side from position 0, respectively. Similarly, symbols 2, 4 and 6 index the positions 0.5 m, 1.0 m and 1.5 m behind position 0, respectively.

coefficients close to impulses is not very high throughout the room. The sound quality of the NBSFC 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 NBSFC 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 NBSFC has a trade-off between sound quality and speech recognition performance. The summary of the results is listed as follows.

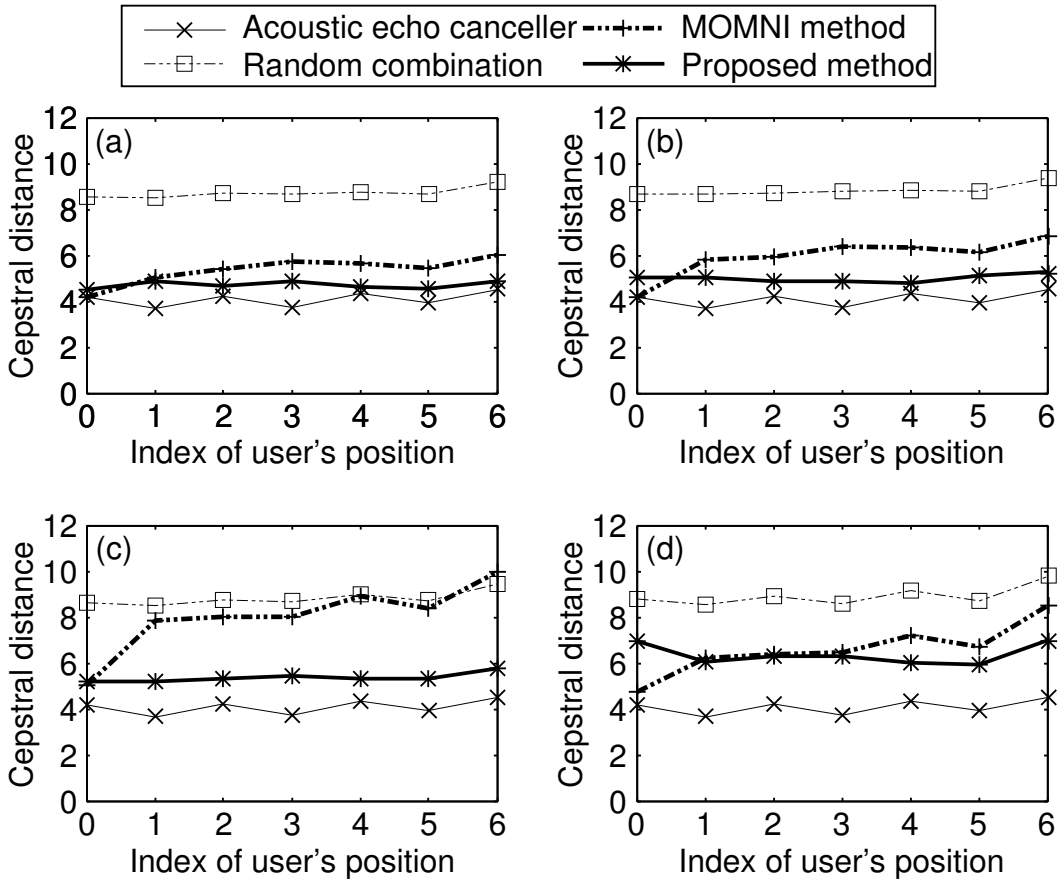

Figure 30. Comparison of cepstral distance from original response sound signal. The experiments are performed with 5 loudspeakers and (a) 1, (b) 2, (c) 3, and (d) 4 microphone elements except for the acoustic echo canceller.

With one or two microphone elements: On the average, the difference in the sound quality of the NBSFC 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 NBSFC 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, it can be said that the NBSFC 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.

4.4 Conclusion

I proposed a new small-scale barge-in robust interface using sound field control called NBSFC to realize response sound cancellation. Since the NBSFC 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 NBSFC together with an appropriate number of microphone elements does not degrade the quality of the response sound. From these findings, the availability of the NBSFC for a spoken dialogue interface is ascertained.

5. Introducing Semiblind Source Separation to MOMNI Method

5.1 Introduction

In this chapter, to adapt a beamformer to the environmental noise and fluctuation of transfer system, source separation is introduced to the MOMNI method. To enhance speech by reducing both known and unknown noise, a new source separation technique is proposed.

Since the MOMNI method uses the simplest DS-type beamformer with a fixed filter, the inherent problem is that the residual response sound caused by fluctuation of transfer function cannot be eliminated perfectly. In addition, the DS beamformer is relatively weak against the reduction of interfering noise, especially in low-frequency regions with limited number of microphones. Indeed in the Chapter 3, we could not address any acoustical scenarios where the interfering noise exists as well as the response sound and user's voice. One intuitive solution is to use an ABF instead of the DS beamformer to eliminate interfering noise sufficiently. However, similarly to the combination of AEC and ABF, many filter adaptation techniques also require DTD, but DTD in heavily noisy environments is still difficult and impractical. Therefore, our next step is to proceed into a new combination of sound field control and *unsupervised adaptation method*, and we mainly deal with an applicability of the combination to a barge-in- and noise-robust spoken dialogue system in this work.

As an unsupervised adaptation of the beamformer, blind source separation (BSS) based on independent component analysis (ICA) [Com94, Lee98] has been studied. Many types of ICA-based BSS methods are presented for an acoustical source separation problem, e.g., time-domain method (TD-ICA) [NSS03, TNSS04, BAK05] and frequency-domain method (FD-ICA) [Sma98, PS00, SKN$^+$06], and some researchers reported an approach to apply TD-ICA to the AEC problem [POL03, ES07]. However, there are no concrete reports on the effective combination of response-sound elimination and noise reduction by ICA, as far as we know. In this chapter, to realize the barge-in- and noise-robust spoken dialogue interface, we extend BSS to a new semiblind source separation (SBSS)

[MTM+06, MTS+07] to separate both the known and unknown sources. By inputting the known response sound signal directly to ICA, its output signals are separated from the echo return of the response sound. Its filter coefficients can be adapted even in the barge-in and noisy situations, and can eliminate the response sound efficiently. In addition, the SBSS can be learned together with BSS to eliminate interfering noise [MTS+07]. Combined with SBSS, the MOMNI method achieves more robustness against the fluctuation of the transfer function and elimination capability of interfering noise.

This chapter is organized as follows. In Chapter 3.3, we will review the principle of the MOMNI method. In Chapter 5.2, we will describe the strategy and the algorithm of the proposed combination of the MOMNI method and SBSS. In Chapters. 5.3–5.4 we will present experimental results and compare the proposed method with the conventional AEC and that combined with ABF using an ideal DTD. The conclusion will be summarized in Chapter 5.5.

5.2 Introducing ICA to the MOMNI method

5.2.1 Motivation

As discussed in Chapter 3.3.4, the MOMNI method can eliminate the response sound with high robustness using many loudspeakers. However, there are two remaining requirements:

(R1) As shown in [MHS+07], robustness of the MOMNI method is improved according to the number of loudspeakers. To reduce the expense of the loudspeakers, adaptation of the elimination is required.

(R2) For a hands-free system, elimination of interfering noise is an important issue, and adaptive signal processing to reduce interfering noise is required.

To satisfy (R1), adaptation is effective in either sound field control or signal processing applied to the observed signals. However, adaptation only in sound field control is invalid for (R2). To satisfy both of them, we try to apply adaptive signal processing to the observed signals of the microphones.

As adaptive signal processings, AEC and ABF are often used for (R1) and (R2), respectively. However, both of them requires DTD and inappropriate for

our purpose, i.e., the double-talk robust system. As an unsupervised filter adaptation without DTD, BSS based on ICA is a strong candidate. Since it is known that frequency-domain ICA (FD-ICA) has advantages over time-domain ICA both for computational simplicity and separation performance [NSS03], we try to adopt FD-ICA in the MOMNI framework.

Here we formulate the problem. The purpose is to separate the sources using under the condition where only the response sound is known without using DTD. Suppose L source signals $\boldsymbol{s}(\omega)$ are observed as mixed signals $\boldsymbol{x}(\omega)$ at K microphones. Among the L sources, the first $s_1(\omega)$ is the desired source, i.e., user's utterance. The L-th source $s_L(\omega)$ is the response sound $r_{\mathrm{src}}(\omega)$ played back by the MOMNI method, i.e.,

$$s_L(\omega) = r_{\mathrm{src}}(\omega). \tag{76}$$

The remaining $L-2$ sources, $s_l(\omega)$ for $l = 1, \ldots, L-1$, are interfering noise. Note that only the L-th source signal $s_L(\omega)$ is known to the system and the others are unknown. In this section the goal is to extract the user's utterance without noise from the mixture by separating the observed signals into each of the sources. Then the mixture of the sources in the observed signals can be described by Eq. (7) similarly to the case of ABF.

$$\boldsymbol{x}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{s}(\omega). \tag{77}$$

While $a_{kl}(\omega)$ for $k = 1, \ldots, K$, $l = 1, \ldots, L-1$ are the room transfer function from the l-th source to the k-th microphone, $a_{kL}(\omega)$ for $k = 1, \ldots, K$ is the complicated transfer function influenced by the binaural room transfer functions $\boldsymbol{g}_{\mathrm{pri}}(\omega)$ in Eq. (28), the filter $\boldsymbol{H}_2(\omega)$ of the MOMNI method, and the M transfer functions $\boldsymbol{g}'_k(\omega)$ from the M loudspeakers to the k-th microphone. The transfer functions $\boldsymbol{g}'_k(\omega)$ are the altered version of the transfer functions $\boldsymbol{g}_k(\omega) = [g_{k1}(\omega), \ldots, g_{kM}(\omega)]$ which are measured in advance to design $\boldsymbol{H}_2(\omega)$. By the effect of the fluctuation in the acoustical condition of the room, $\boldsymbol{g}_k(\omega)$ is altered to $\boldsymbol{g}'_k(\omega)$ by a small additional component $\Delta\boldsymbol{g}_k(\omega)$, described as

$$\boldsymbol{g}'_k(\omega) = \boldsymbol{g}_k(\omega) + \Delta\boldsymbol{g}_k(\omega). \tag{78}$$

For the elimination of the response sound at the microphones, the filter $\boldsymbol{H}_2(\omega)$ of the MOMNI method is designed to satisfy $\boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega) = 0$. Thus the residual

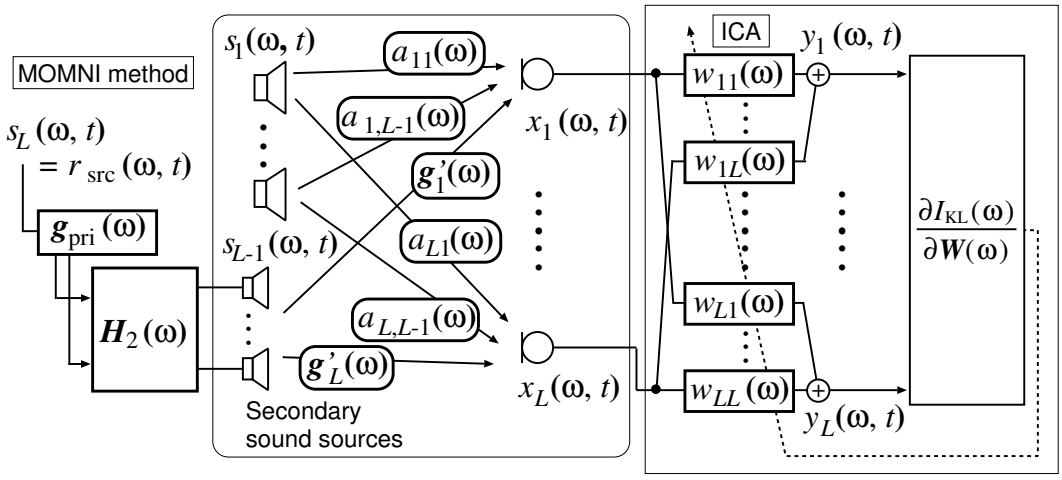

Figure 31. Configuration of BSS based on FD-ICA integrated with MOMNI method.

observation $d'_k(\omega)$ of the response sound at the k-th microphone, caused by the fluctuation, can be written as

$$\begin{aligned} d'_k(\omega) &= \boldsymbol{g}'_k(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega)r_{\mathrm{src}}(\omega) \\ &= (\boldsymbol{g}(\omega) + \Delta\boldsymbol{g}(\omega))\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega)r_{\mathrm{src}}(\omega) \\ &= \Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega)r_{\mathrm{src}}(\omega). \end{aligned} \tag{79}$$

Thus the transfer function $a_{kL}(\omega)$ for $k = 1, \ldots, K$ can be written as

$$a_{kL}(\omega) = \Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega). \tag{80}$$

5.2.2 BSS based on FD-ICA

In this section we review the general principle of BSS based on FD-ICA. Configuration of BSS is shown in Fig. 31. BSS is a problem to estimate unknown source signals only from the observed signals without using DTD and any other a-priori information about sources like steering vector used in ABF. In this framework, even the known source signal $s_L(\omega)$ is regarded as unknown signal. The purpose of BSS is to obtain an $L \times K$ separation filter matrix $\boldsymbol{W}(\omega)$ which makes its output signals,

$$\boldsymbol{y}(\omega) = [y_1(\omega), \ldots, y_L(\omega)]^{\mathrm{T}} = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega), \tag{81}$$

be the estimation of the separated sources.

First, similarly to ABF, short-time analysis of the observed signals is conducted by frame-by-frame discrete Fourier transform. Time-frequency expression of the observed signals $\boldsymbol{x}(\omega)$ are denoted as $\boldsymbol{x}(\omega, t)$. Next, we obtain the separation filter $\boldsymbol{W}(\omega)$ whose time-series output $\boldsymbol{y}(\omega, t)$, written as

$$\boldsymbol{y}(\omega, t) = [y_1(\omega, t), \ldots, y_L(\omega, t)]^{\mathrm{T}} = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega, t), \tag{82}$$

is separated to each of source components. Assuming statistical independence among the sources $\boldsymbol{s}(\omega, t) = [s_1(\omega, t), \ldots, s_L(\omega, t)]$, the necessary and sufficient condition for the separation is statistical independence among $\boldsymbol{y}(\omega, t)$. For the case of $K = L$, such $\boldsymbol{W}(\omega)$ to output independent signals is optimized by, for example, the following iterative updating operation [ACY96]:

$$\boldsymbol{W}(\omega) \rightarrow \boldsymbol{W}(\omega) - \eta\{\boldsymbol{I} - \langle \boldsymbol{\Phi}(\boldsymbol{y}(\omega, t))\boldsymbol{y}^{\mathrm{H}}(\omega, t)\rangle_t\}\boldsymbol{W}(\omega), \tag{83}$$

where $\rightarrow$ denotes updating operation in a single iteration. In our research, we use tangent hyperbolic function based on polar coordinate [SMAM03] as;

$$\boldsymbol{\Phi}(\boldsymbol{y}(\omega, t)) = \begin{bmatrix} \tanh(|y_1(\omega, t)|)\exp\left(j\arg\left(y_1\left(\omega, t\right)\right)\right) \\ \vdots \\ \tanh(|y_K(\omega, t)|)\exp\left(j\arg\left(y_L\left(\omega, t\right)\right)\right) \end{bmatrix}. \tag{84}$$

The optimum solution of $\boldsymbol{W}(\omega)$ satisfies the following condition similar to Eq. (10);

$$\boldsymbol{W}(\omega)\boldsymbol{A}(\omega) = \boldsymbol{\Pi}(\omega)\boldsymbol{C}(\omega), \tag{85}$$

where $\boldsymbol{\Pi}(\omega) = [\delta_{\Pi_\omega(l)k}]_{lk}$ is an $L \times L$ permutation matrix with an arbitrary permutation $\Pi_\omega : \{1, \ldots, L\} \rightarrow \{1, \ldots, L\}$, δ_{lk} is Kronecker delta, and $\boldsymbol{C}(\omega)$ is a $L \times L$ diagonal matrix whose diagonal components are arbitrary complex value.

The separation filter $\boldsymbol{W}(\omega)$ to satisfy the condition Eq. (85) requires some modifications to be used in noise reduction. First, as can be seen as arbitrary diagonal matrix $\boldsymbol{C}(\omega)$ in Eq. (85), the condition of independence has ambiguity in scaling of the output signals, both for amplitudes and phases. To compensate for this, we apply inverse filter of the separation filter $\boldsymbol{W}(\omega)$ [IM99] to estimate the source signals at the microphone points using inverse of the separation filter. Second, as can be seen as $\boldsymbol{\Pi}(\omega)$ in Eq. (85), independence also has ambiguity in

the ordering of the signals to be outputted, refered to as 'permutation problem'. To reconstruct the estimated sources, the ordering must be aligned. To solve the permutation, several approaches have been proposed, e.g., use of directivity pattern of the separation filter [SKT+03] and use of envelopes' correlations among narrow-band signals [IM99], and their combination [SMAM04].

Combination of FD-ICA and the MOMNI method has several problems. The first one is that its output signals are distorted. The mechanism of convolutive separation by ICA is multiple beamformers which extract independent sources separately [AMM+03]. In general, to construct a beamformer with high performance, required filter length is longer than those of the room transfer functions. Since the inverse filter $\boldsymbol{H}_2(\omega)$ used in the MOMNI method has much longer impulse responses than those of the room transfer functions. By necessity the transfer function $a_{kL}(\omega) = \Delta \boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega)$ has long impulse response. Nevertheless, we must use short filter coefficients in a real environment because blind estimation of long filter coefficients requires long input data which is difficult to obtain. The use of the short filter coefficients distorts the output signals as a result of a circular convolution effect. The second problem is difficulty in solving permutation in this case. Since the transfer functions corresponding to the response sound, i.e., $\Delta \boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega)$, have no specific directivity, the permutation solution based on directivity is insufficient. For these reasons, we cannot expect that this integration of BSS and the MOMNI method will perform as well as the ordinary BSS.

5.2.3 Semiblind source separation

In the previous section, we have discussed combination of the MOMNI method and BSS where the response sound is dealt as an unknown signal, and shown its insufficiency. In this section, we propose a new semiblind source separation (SBSS) which separates sources from mixture of known and unknown sources efficiently utilizing information of known source. We give information of known source by inputting the known source directly into to ICA.

Configuration of SBSS is shown in Fig. 32. To separate L sources, ICA requires L mixed reference signals, which are L observed signals in ordinary BSS. However in this case, we can use $s_L(\omega)$ directly as one of the reference input sig-

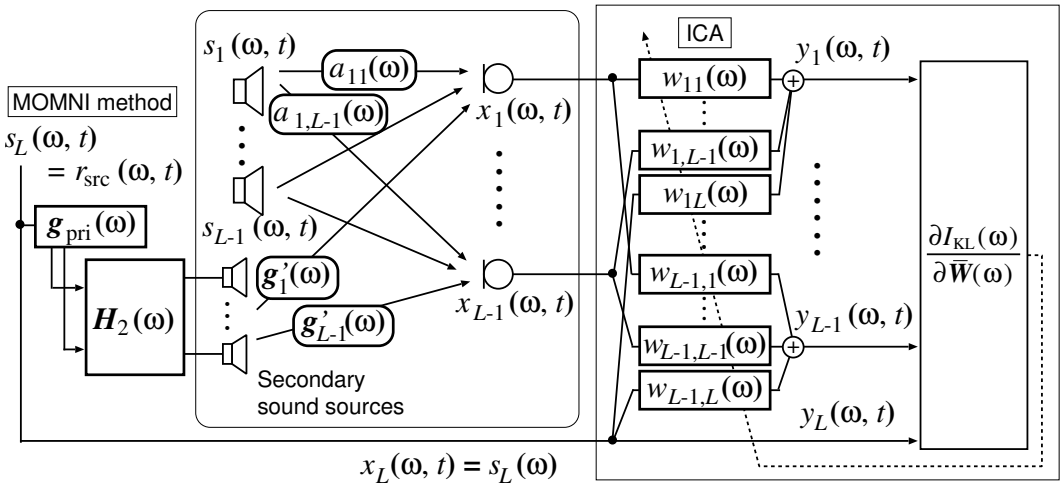

Figure 32. Configuration of semiblind source separation.

nals combined with $L-1$ observed signals. Such input signals can be described by the substitution

$$a_{Lk}(\omega) = \begin{cases} 0 & \text{for} \quad k = 1, \ldots, L-1, \\ 1 & \text{for} \quad k = L, \end{cases}$$

and

$$x_L(\omega) = s_L(\omega), \tag{86}$$

in Eq. (77). Since the L-th input signal $s_L(\omega)$ is already separated, it should be outputted without any modification, i.e.,

$$y_L(\omega) = x_L(\omega) = s_L(\omega). \tag{87}$$

Thus, the L-th row $\boldsymbol{w}_L(\omega)$ of the separation filter $\boldsymbol{W}(\omega)$ should be fixed as

$$w_{Ll}(\omega) = \begin{cases} 0 & \text{for} \quad l = 1, \ldots, L-1, \\ 1 & \text{if} \quad l = L. \end{cases} \tag{88}$$

Since $\boldsymbol{w}_L(\omega)$ is fixed, the components $\bar{\boldsymbol{W}}(\omega)$ in $\boldsymbol{W}(\omega)$ to be updated is the $(L-1) \times L$ truncated submatrix

$$\begin{aligned} \bar{\boldsymbol{W}}(\omega) &= [w_{lk}(\omega)]_{lk} \quad \text{for} \quad l = 1, \ldots, L-1, \ k = 1, \ldots, L \\ &= [\boldsymbol{I}_{L-1}, \boldsymbol{0}_{L-1}] \ \boldsymbol{W}(\omega), \end{aligned} \tag{89}$$

where $\mathbf{0}_i$ denotes i-dimensional column zero vector. As derived in Appendix A, independence among $\boldsymbol{y}(\omega, t)$ can be improved by the following updating formula;

$$\bar{\boldsymbol{W}}(\omega) \leftarrow \bar{\boldsymbol{W}}(\omega) - \eta \left\{ \bar{\boldsymbol{W}}(\omega) - \langle \boldsymbol{\Phi}(\bar{\boldsymbol{y}}(\omega, t)) \boldsymbol{y}^{\mathrm{H}}(\omega, t) \rangle_t \boldsymbol{W}(\omega) \right\}, \tag{90}$$

where

$$\bar{\boldsymbol{y}}(\omega) = [y_1(\omega), \ldots, y_{L-1}(\omega)]^{\mathrm{T}}. \tag{91}$$

The fix of $\boldsymbol{w}_L(\omega)$ has many advantages over conventional BSS. First, with the constraint that the component due to $s_L(\omega)$ is fixed to outputted from $y_L(\omega)$, we need not solve the permutation for $y_L(\omega)$ but for only the remaining $L-1$ outputs. Second, giving part of the answer $y_L(\omega) = x_L(\omega)$ makes the problem easier and helps the avoidance of local minima in the nonlinear optimization. In addition, SBSS has advantage in the length of the separation filter. Though BSS is a problem to obtain a beamformer, SBSS eliminates the component due to $s_L(\omega)$ in $y_l(\omega)$ for $l = 1, \ldots, L-1$ by obtaining opposite phase of mixture just like AEC. Thus required filter length becomes shorter.

Since ICA in SBSS exploits higher-order statistics, more information from training samples is extracted compared with correlation analysis Eq. (6). Thus, with limited length of training samples, SBSS outperforms AEC. The proposed SBSS is equivalent to the complex extension of the instantaneous nonblind source separation proposed in [JMM01]. Extension to convolutive separation problem is already proposed in [POL03] and [ES07] in the form of TD-ICA. However, it is known that FD-ICA is superior to TD-ICA to deal with long impulse response, which is the case in speech application. In addition, as discussed above, we have investigated that the extension to FD-ICA has an advantage to overcome permutation problem, which is one of the most significant weakness of FD-ICA.

The advantage of SBSS discussed above is important especially when combined with the MOMNI method. As discussed in Chapter 5.2.2, the long impulse response of $a_{kL}(\omega) = \Delta \boldsymbol{g}_k(\omega) \boldsymbol{H}_2(\omega) \boldsymbol{g}_{\mathrm{pri}}(\omega)$ requires BSS to have extremely long filter coefficients. Since the required filter length in SBSS is shorter than BSS, the separation works sufficiently. In addition, difficulty in solution of permutation caused by no specific directivity of the MOMNI method can be solved by the fix of the output of the known source. Thus, by combining the MOMNI method and SBSS, the sound elimination before the microphone by the MOMNI method and

adaptation by ICA coordinate successfully, and robust speech recognition in the high quality sound presentation can be realized.

5.2.4 Computational complexity and sound quality

I have discussed that the computational complexity of the MOMNI method is lower than that of the AEC at the sacrifice of the cost of multiple loudspeakers. However, the computational cost in the adaptation of ICA is much larger than those of the AEC and the ABF, and requires (the number of the frames) $\times$ (the number of the microphones) $\times$ (the number of the output signals) for an iteration in a frequency bin. In addition, update of the filter has long latency, and to implement real-time BSS, an older filter than the current update should be used in the separation. Since a certain length of the reference signals are required for an efficient performance, the filter should be updated in batch-wise manner once every certain time, e.g., three seconds.

The MOMNI method combined with ICA plays an important role to cover the late learning and large computation of ICA. When the convergence of ICA is insufficient, the MOMNI method designed in advance eliminates the response sound with a certain performance. Thus the MOMNI method covers the above-mentioned late learning of ICA and degraded performance of ICA by the reduced iteration number for a real-time implementation. In addition, semi-blind structure reduces computation of BSS because of the reduced number of the output signals; output of the known response sound is not required.

As for the quality of the response sound, the output response sound is the same as that of the MOMNI method because the signal processing after the observation does not affect the loudspeaker output. As examined in Chapter 3.6, the quality outside the control point slightly degrades but it is not problematic for the spoken dialogue system.

5.3 Experiments without interfering noise

5.3.1 Experimental conditions and competitive methods

In this section, to validate robustness of the proposed combination of MOMNI method and SBSS to eliminate the response sound, we conduct an experiment un-

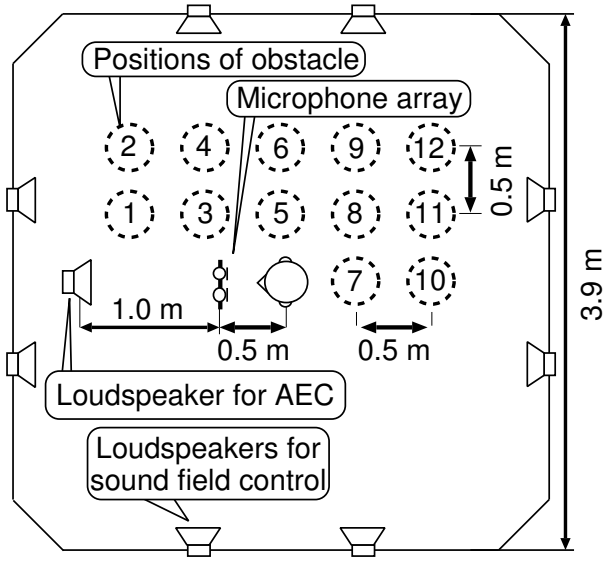

Figure 33. Layout of the acoustic environment room when there is no other noise than the response sound.

der existence of the response sound and the user's utterance, without interfering noise. Here we describe the competitive methods. As the conventional methods, we evaluate performances of two kinds of AECs. The first AEC (**AEC1**) is the AEC adapted by the adaptation Eq. (6) without DTD, to investigate importance of DTD. The second AEC (**AEC2**) is a kind of performance limit of the AEC using ideal DTD. By giving the true single talk by setting threshold to the power of the source which is unknown in fact, we imitated an ideal behavior of DTD. The detection and adaptation is conducted in time-frequency domain. Details of its adaptation is described in Appendix B.1. In addition, conventional the MOMNI method with DS beamformer (**MOMNI+DS**) is evaluated. To investigate superiority of SBSS to BSS and effectiveness of the integration with the MOMNI method, the performances of BSS and SBSS are evaluated in the playback of the response sound both with the single loudspeaker (**BSS** and **SBSS**) and with the MOMNI reproduction (**MOMNI+BSS** and **MOMNI+SBSS**).

The impulse responses used in this experiment are measured in an acoustic experimental room, where the reverberation time is approximately 160 ms. Figure 33 shows the arrangement of the apparatuses. Sampling frequency and resolution

are 16 kHz and 16 bit, respectively. Eight loudspeakers used in the sound field control of the MOMNI method are positioned along the outer circumference of the room. The primary sound source of sound field reproduction is a loudspeaker set in the center of the room. This loudspeaker is also used to play back the response sound in AEC1, AEC2, BSS and SBSS methods. As shown in Fig. 33, we place a dummy head, which is a replica of an average human head and torso, at the user's position. For the AECs, the power of the user's speech and the response sound are arranged to be the same. For BSS and SBSS, the power balances are similarly determined. However, for the methods with sound field control of the MOMNI method, i.e., MOMNI+DS, MOMNI+BSS and the proposed combination, the proportion of power cannot be determined in the same manner. Since the response sound is intended to be presented to the user, its volume of necessity depends on the power to reach the user's ears. Thus we arrange the power of the response sounds to be the same as the AECs at the user's ears, and the user's speech to be the same as the AECs. For the AECs and the ICAs, we use filter with 2048 taps. The length of the training data to adapt their filters is 5 seconds. The passband range of the response sound is set to 150–4000 Hz. For the ICAs used in BSS and MOMNI+BSS, the interelement spacing is 6 cm. For the DS used in MOMNI+DS, the interelement spacing is 30 cm. BSS, MOMNI+DS and MOMNI+BSS use two microphone elements while the other methods use single microphone. All BSS-based methods use the permutation solver proposed in [SMAM03].

When the room transfer functions do not alter from the state where the inverse filter was designed, the performance of the MOMNI method is almost perfect. However, since the transfer functions fluctuate at all times, the performances should be evaluated in the state after fluctuations. To this end, we located an obstacle and we measured various impulse responses by changing its position. Assuming that another person than the user is moving about in the room, we used a life-size mannequin as the obstacle. We measured 13 kinds of impulse responses as follows: one is for the state where the obstacle does not exist, and the other 12 are for the states where the obstacle is located at various positions near the dialogue system. The inverse filter in the MOMNI method was designed with the impulse responses before fluctuation, and we evaluated the average of

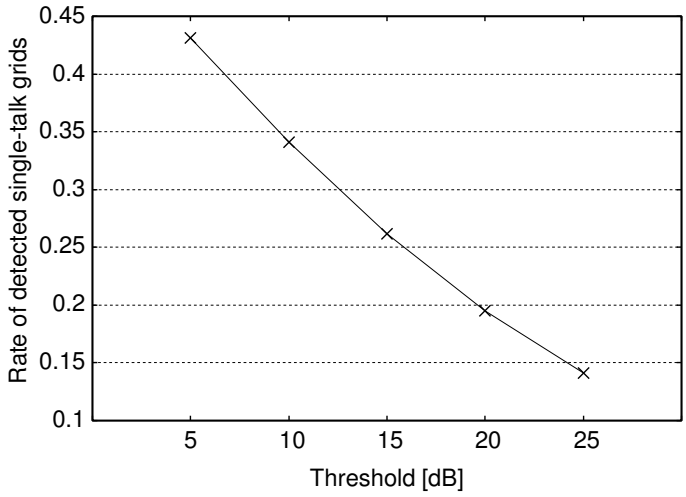

Figure 34. The relationship between the threshold and the rate of time-frequency grids to be judged as single-talk. The horizontal axis shows the rate of the number of whole the grids and the number of the grids where the power ratio of the response sound to the user's speech exceeds. The power balance between the two signals is even at the microphone. The scores are averaged over 200 sentences of the user's speech and 12 fluctuations.

the performances in the latter 12 states after fluctuations. We used a sentence of a male utterance as the response sound. As the user's utterance, we used 200 sentences spoken by 13 male and 13 female selected from JNAS database [IYT+99]. The performances are also averaged by these 200 utterances.

Here we describe behavior of the ideal DTD used in AEC2. There is trade-off between the quality and the quantity of the training data selected by the DTDs used in AEC2. To obtain training data with higher SNR for the adaptation, threshold of the detection must be set higher. However, the high threshold rejects many data and causes shortage of training data. We show the relationship between the threshold and the rate of single-talk time-frequency grids detected by the DTD in Fig. 34. In this experiment, the highest performance is achieved with the threshold of 15 dB among 5, 10, 15, 20 and 25 dB. Therefore in the following discussion, we show the results with 15 dB of the threshold for AEC2.

5.3.2 Performance evaluation of response sound reduction

We evaluate signal-to-noise ratios (SNRs) of the observed signals (observed SNRs) and the processed signals (processed SNRs) for each of the compared methods. The results are shown in Fig. 35. These SNRs are merely the power ratios of the user's speech and the other signals:

$$\mathrm{SNR} = 10 \log_{10} \frac{\sum_\omega |y_{\mathrm{user}}(\omega)|^2}{\sum_\omega |y(\omega) - y_{\mathrm{user}}(\omega)|^2} \ [\mathrm{dB}], \tag{92}$$

where $y(\omega)$ is the observed or processed signal, and $y_{\mathrm{user}}(\omega)$ is the component due to the user's speech in $y(\omega)$. Consequently, it should be noted that these scores are not influenced by distortions of the spectra. To evaluate the performances of the sound field control and the signal processing after observation, we evaluated the SNRs of both the observed signals and the processed signals.

Because AEC1 is adapted to short reference signal without DTD, its performance is quite low. Although AEC2 is adapted to the room transfer functions, its performance is limited within almost the same level of MOMNI+DS without adaptation because of the shortage of single-talk components. The performance of BSS is the worst. It is well known that the separation mechanism of BSS is based on beamforming [AMM+03]. Because the user, the microphone array and the loudspeaker all fall on the same straight line in this experiment, it is difficult for beamformers to separate the response sound and the user's speech well. Since the separation mechanism of SBSS is based on AEC but on the beamformer, such a positioning is of little significance and the performance is as good as AECs despite no use of DTD. However, the improvement of the processed SNR from the observed SNR, i.e., effective improvement by the DS beamformer is not very high. In MOMNI+BSS, the improvement of the SNR is lower than the DS beamformer. This is because the permutation alignment is difficult, as discussed in Chapter 5.2.2. In contrast, in the proposed combination of the SBSS and the MOMNI method, the SBSS sufficiently eliminates the residual response sound of the MOMNI method. Consequently the highest performance is achieved by the proposed combination.

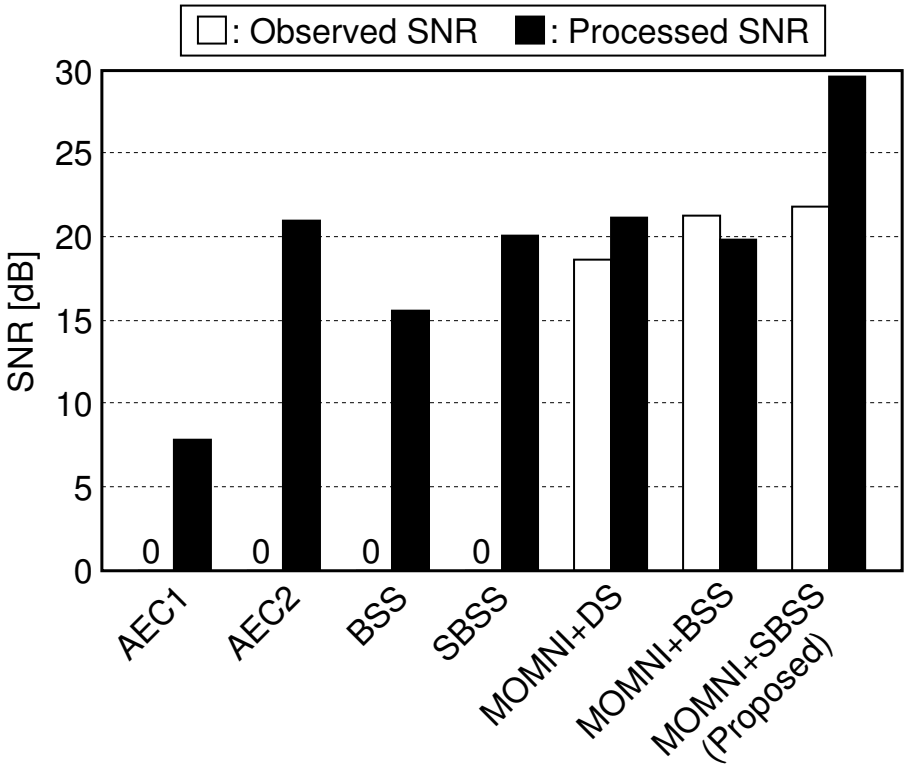

Figure 35. Comparison of observed SNRs without interfering noise. The SNRs are evaluated both for the observed signals and the processed signals. The score is averaged over 200 sentences of the user's speech and 12 fluctuations.

5.3.3 Assessment of distortion

In this section, we assess the quality of the enhanced speech signal. We apply CD shown in Eq. (48) as an evaluation score. The reference is the component of user's speech in the observed signal without mixture of the other components, and distortion is evaluated for the processed signal of this signal without the mixture. Therefore, in contrast to the SNR, this score indicates the distortion of signal processing but is not affected by elimination error.

Since the AEC1, AEC2, SBSS and the proposed combination do not modify the component of the user's utterance, the processed signals are not distorted and their CDs are zero. Although the DS beamformer in MOMNI+DS modifies the

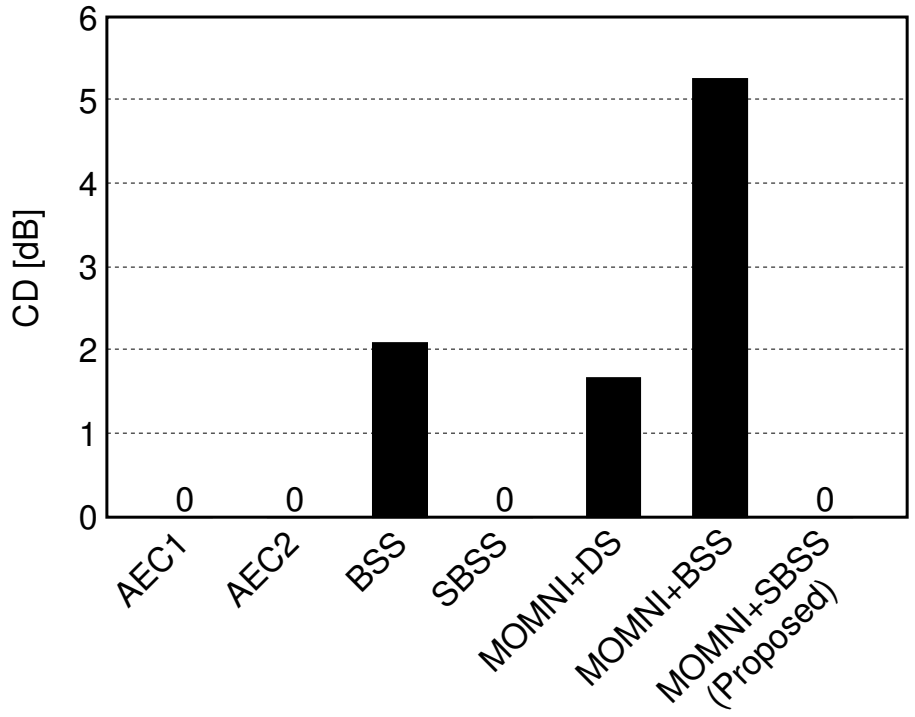

Figure 36. Comparison of CDs without interfering noise. The signal component of user's speech in the observed signal without mixture of the other component is used as reference signal. The CDs evaluate how much the reference signal is distorted by signal processing. The scores are averaged over 200 sentences of the user's speech and 12 fluctuations.

user's speech, its effect is not problematic for speech enhancement. Thus the CD of the conventional MOMNI method with the DS beamformer is low. However, we can see that for BSS and MOMNI+BSS, their outputs are distorted. It can be said for both of them that the permutation distorts their outputs. In addition, as for MOMNI+BSS, the circular convolution effect induces further distortion as discussed in Chapter 5.2.2.

Table 8. Experimental conditions for speech recognition

Test data	JNAS [IYT+99]
Frame length	25 msec (Hamming window)
Frame interval	8 msec
Feature vector	12 MFCCs, 12 ΔMFCCs, Δpower
Language model	Newspaper dictation [IYT+98]
Phoneme model	Phonetic Tied Mixture (PTM) [LKTS00] with known noise imposition [YLSS03]
Decoder	Julius ver. 3.5.1 standard [LKK01]

5.3.4 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 WA as an evaluation score, which is given by Eq. (47). Table 8 lists the experimental conditions for speech recognition. Figure 37 shows the WAs with all the combinations. The speech recognition performances are affected by both elimination performance and distortion. All the WAs are proportional to the processed SNRs except the methods with high distortion, i.e., BSS and MOMNI+BSS. For BSS and MOMNI+BSS, their speech recognition performances are heavily degraded by the distortion. Because of the high elimination performance without distortion, the proposed combination of the SBSS and the MOMNI method shows the highest speech recognition performance and the value is close to the upper limit.

5.4 Experiments in noisy environments

5.4.1 Experimental conditions and competitive methods

To validate the elimination performance of both the response sound and interfering noise, we compare the speech enhancement performances in noisy environments between the proposed combination (**MOMNI+SBSS**) and conventional methods, i.e., two kinds of AEC-integrated ABF, **BSS**, **SBSS**, **MOMNI+DS**, and **MOMNI+BSS**. The ABF is adapted using ideal DTD after echo cancella-

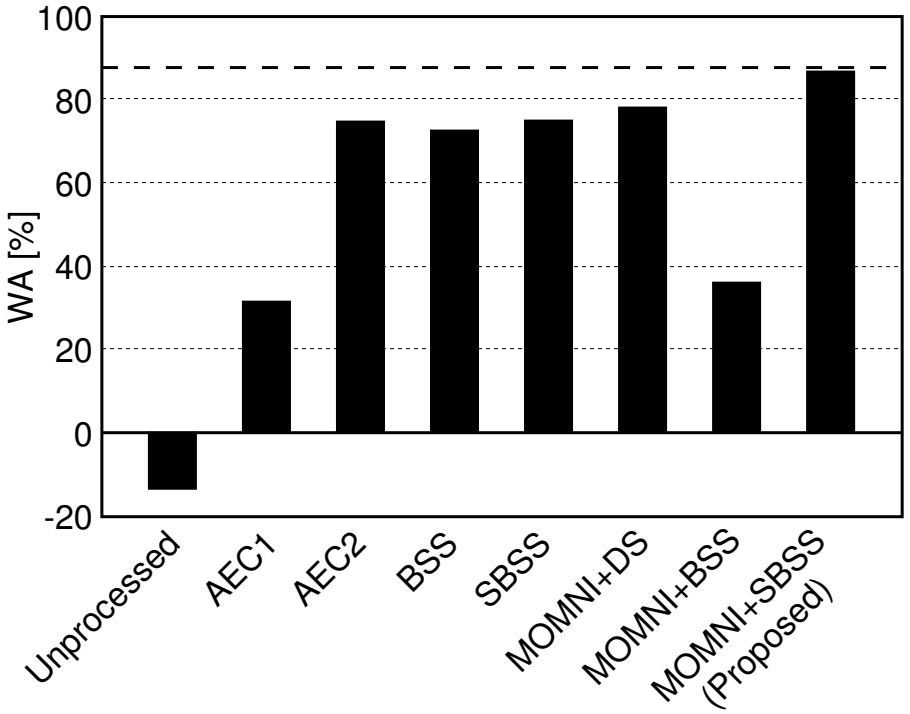

Figure 37. Comparison of WAs of the processed signals without interfering noise. The scores are averaged over 200 sentences of the user's speech and 12 fluctuations. The broken line shows upper limit in the room.

tion with the two kinds of AECs in Chapter 5.3 (**AEC1+ABF** and **AEC2+ABF**). The detail of the adaptation is described in Appendix B.2. Figure 38 shows the arrangement of the apparatuses. Apparatuses are similar to those in Fig. 33 except for the existence of interfering noise at an edge of the room. For AECs and ICAs, we use filter length of 2048 taps. The length of the training data to adapt their filters is 5 seconds. AEC uses single microphone. AEC+ABF, SBSS and the proposed combination use two microphone elements while BSS and MOMNI+BSS use three microphone elements, and their interelement spacings are about 3 cm. MOMNI+DS uses three microphone elements with interelement spacing of about 26 cm. The length of the training data to adapt each of the adaptive filters and ICAs is 5 seconds. All BSS-based methods use the permutation solver proposed in [SMAM03].

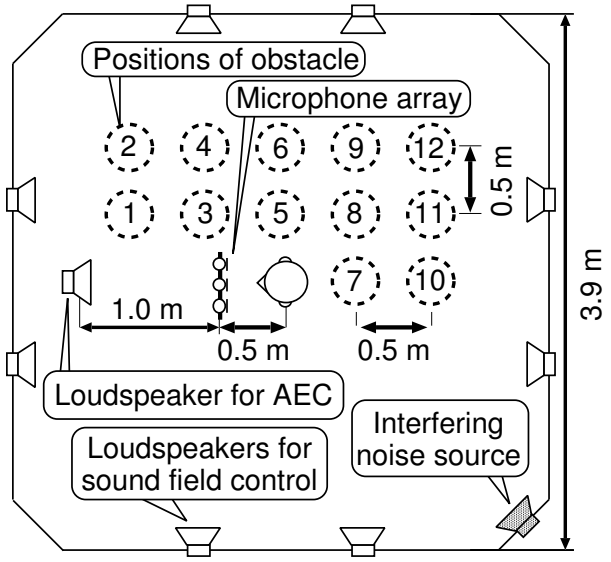

Figure 38. Layout of acoustic environment room when there is interfering noise.

The power balances of the user's speech and the response sound are set in the same manner as discussed in Chapter 5.3. The power of the interfering noise is arranged to be 10 dB lower than that of the user's speech. As interfering noise, we used three signals, i.e., a female utterance, music (a symphony), and stationary noise with -10 dB/octave spectral coloration.

We show the rates of the time-frequency grids detected as single talk by DTD1 for the AEC2 and DTD2 for the ABF in Figs. 39 and 40, respectively. As can be seen in Fig. 39, the order of the sparseness is from female utterance, music to stationary noise. Therefore, for the adaptation of the AEC2, the most single-talk grids of the response sound can be find in the noise of female utterance. However, in contrast, Fig. 40 shows that the single-talk of the female utterance grids for the ABF adaptation are fewest because of the sparseness. In average of these three noises, the best performance was achieved with $T(\omega) = 15$ dB. Therefore, we use this value in the following discussion.

The other conditions are similar to those of the experiments in Chapter 5.3.

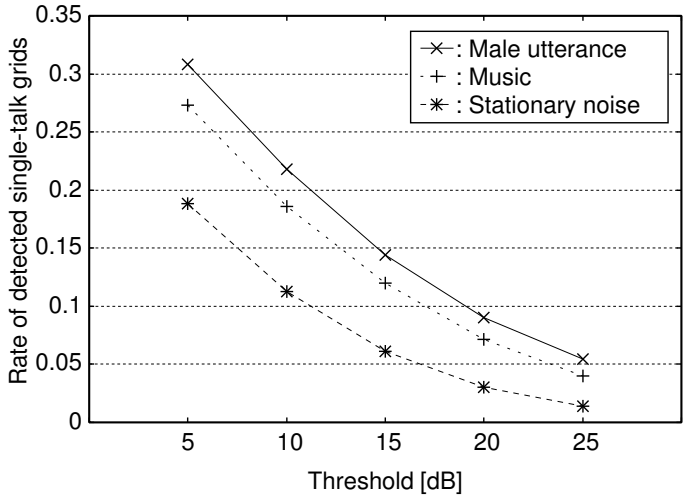

Figure 39. The relationship between the threshold $T(\omega)$ for DTD1 and the rate of time-frequency grids to be judged as single-talk. The horizontal axis shows the rate of the number of whole the grids and the number of the grids where the power ratio of the response sound to the other signals exceeds $T(\omega)$. The powers of the user's speech and the response sound is equal and those of the interfering noises are 10 dB lower. The scores are averaged among 200 sentences of the user's speech and 12 fluctuations.

5.4.2 Performance evaluation of noise reduction

Similarly to the experiment discussed in Chapter 5.3.2, we evaluate the SNRs of the observed and the processed signals to assess the reduction performance of the response sound and the interfering noise. The observed and processed SNRs are shown in Figs. 41 and 42, respectively.

Since the power of the interfering noises are -10 dB, the observed SNRs of the methods with sound field control are around 10 dB. Otherwise, the observed SNRs of the other methods are under 0 dB.

For the AEC and MOMNI+DS, the processed SNRs cannot reach the sufficient level because they have no effective mechanism for eliminating the interfering noise. We can confirm that SBSS can adapt well even in noisy environment. Although the processed SNR of AEC+ABF and SBSS are similarly high, it should be noted that such precise DTDs used in AEC+ABF cannot be implemented in

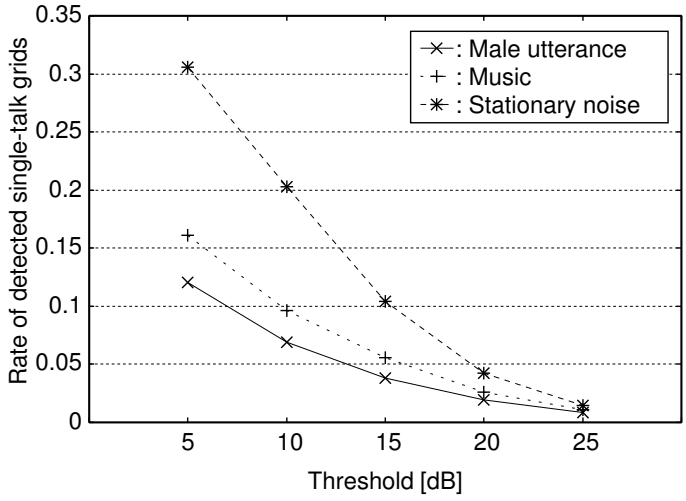

Figure 40. The relationship between the threshold $T(\omega)$ for DTD2 and the rate of the time-frequency grids to be judged as single-talk. The DTD2 detects the single-talk time frequency grids of the interfering noise in the processed signals of the AEC with DTD1. The thresholds of DTD1 and DTD2 are equal. The horizontal axis shows the rate of the number of whole the grids and the number of the grids where the power ratio of the interfering noise to the other signals exceeds $T(\omega)$. The powers of the user's speech and the response sound is equal and those of the interfering noises are 10 dB lower. The scores are averaged among 200 sentences of the user's speech and 12 fluctuations.

practice while SBSS requires no DTD. The performance of MOMNI+BSS is lower than those of AEC+ABF and SBSS. The improvement of SNR is more considerable in the proposed combination of SBSS and the MOMNI method compared with all of the competitive methods, which shows the best processed score.

5.4.3 Assessment of distortion

To evaluate the distortion occurring in each signal processing, we compare the CDs in the same manner as described in Chapter 5.3.3. The results are shown in Fig. 43. While BSS and MOMNI+BSS have high distortion, the distortion of the other methods is not problematic.

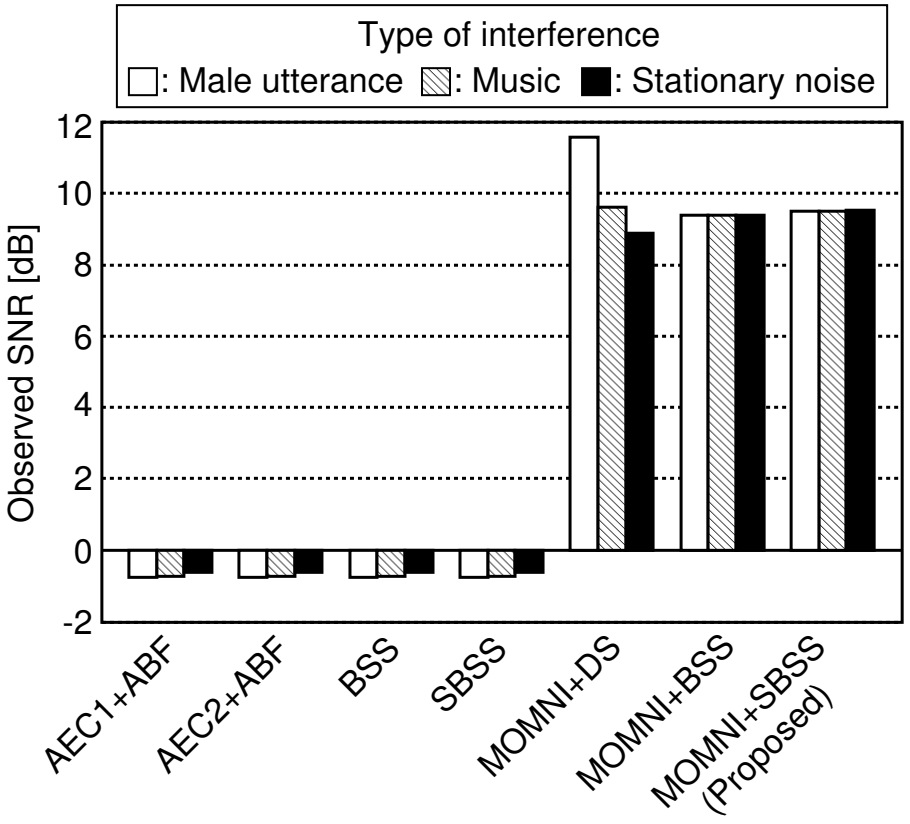

Figure 41. Comparison of observed SNRs for three kinds of interfering noises. The scores are averaged among 200 sentences of the user's speech and 12 fluctuations.

5.4.4 Speech recognition experiment

To assess the efficacy of the proposed combination under the existence of the response sound and interfering noise, we evaluate its speech recognition performance. The experimental conditions are similar to those discussed in Chapter 5.3.4. The results are shown in Fig. 44.

The results are almost proportional to those in Fig. 37. We can say that the performance of SBSS is very high because, despite no use of DTD, the SBSS achieves close performance to AEC2+ABF using ideal DTDs. Although the proposed combination has slightly more distortion than AEC2+ABF, the WA of the proposed combination is the highest because of the remarkably high processed

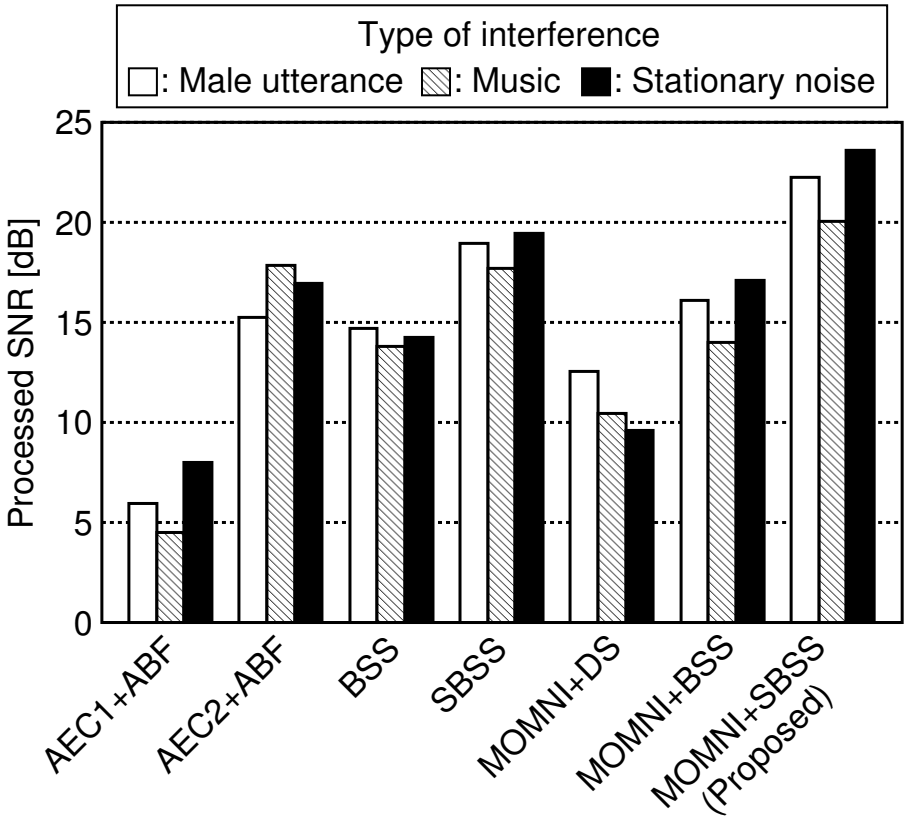

Figure 42. Comparison of processed SNRs for three kinds of interfering noises. The scores are averaged among 200 sentences of the user's speech and 12 fluctuations.

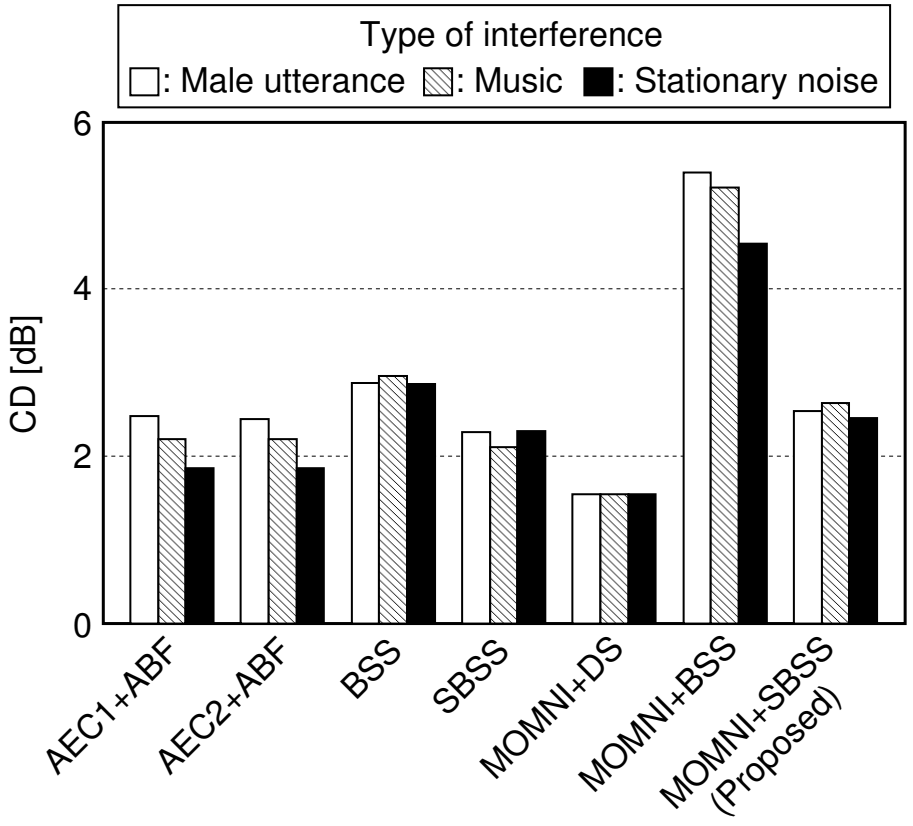

Figure 43. Comparison of CDs for three kinds of interfering noises. The signal component of user's speech in the observed signal without mixture of the other components is used as reference signal. The CDs evaluate how much the reference signal is distorted by signal processing. The scores are averaged among 200 sentences of the user's speech and 12 fluctuations.

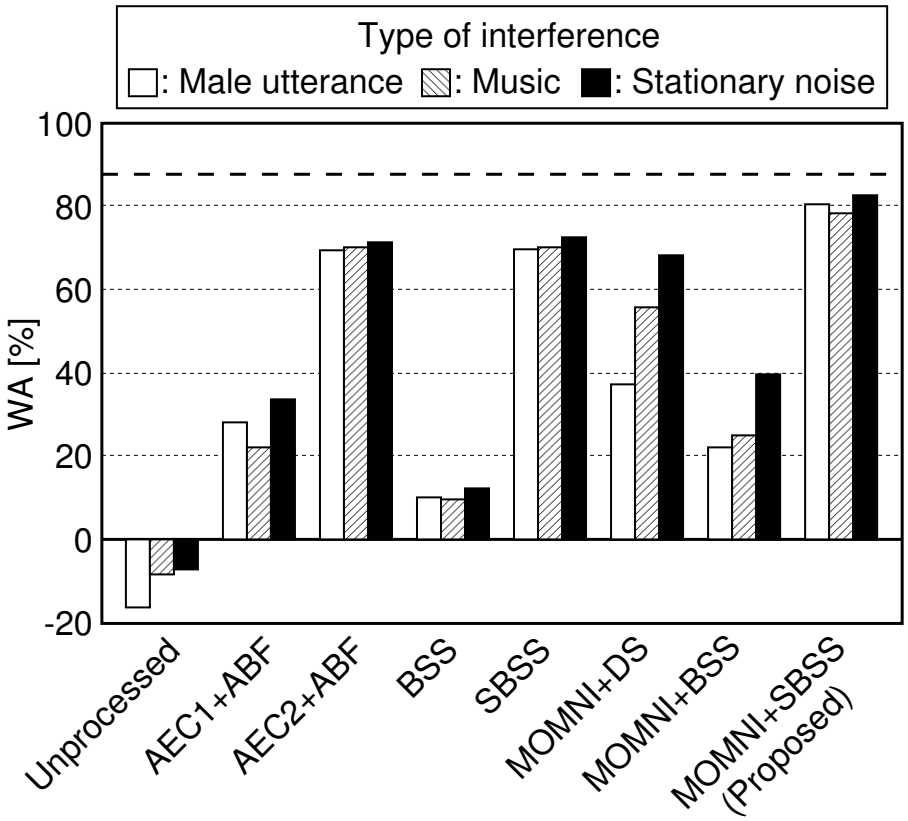

Figure 44. Comparison of WAs of the processed signals for three kinds of interfering noises. The scores are averaged among 200 sentences of the user's speech and 12 fluctuations. The broken line shows upper limit in the room.

SNR. In summary, these results can provide the convincing evidence that our proposed strategy (*combination of sound field control and SBSS*) is more beneficial to the barge-in- and noise-robust spoken dialogue system in comparison to possible conventional speech enhancement methods.

5.5 Conclusion

This chapter mainly addressed the problem on improvement of robustness in a spoken dialogue system. To improve the performance of the conventional MOMNI method using sound field control and a DS beamformer, we proposed to combine

the MOMNI method with SBSS by extending conventional BSS. By inputting known response sound signal directly to ICA, SBSS can separate the user's speech from the response sound and interfering noise without DTD. Because of the simplification of permutation solution and decrease of the required filter length, the possible drawbacks in SBSS are mitigated in comparison to conventional BSS. By replacing a DS beamformer with SBSS, the MOMNI method achieves both more robustness and ability of noise reduction.

In the experiment without interfering noise, the proposed combination showed higher speech recognition score than AEC adapted under an ideal condition. Thus it is revealed that the proposed combination can eliminate the response sound with high accuracy. Also in the experiment with interfering noises, the proposed combination showed higher performance than a combination of AEC and ABF adapted under an ideal condition. The proposed combination can reduce both the response sound and interfering noise regardless of the double-talk. From these findings, the efficacy of the proposed combination is confirmed for the interface of a barge-in- and noise-robust spoken dialogue system.

6. Conclusion

6.1 Thesis summary

In this dissertation, I described a hands-free interface for a spoken dialogue system using sound field reproduction and source separation. The proposed framework is easier to implement in a real-time system than the conventional methods based on adaptive filters, because the proposed framework excludes double-talk detection (DTD).

In Chapter 1, I introduced the background of this research, and briefly described the basic concept of the dissertation.

In Chapter 2, I reviewed optimal solution of the adaptive signal processing for acoustic echo canceller (AEC) and adaptive beamformer (ABF), and pointed out the inherent problem of the DTD.

In Chapter 3, to eliminate the response sound of the spoken dialogue system with a fixed filter, I proposed a new combination of sound field reproduction and a beamformer, named the multiple-output and multiple-no-input (MOMNI) method. By reproducing the response sound at the user's ears and silent signals at the microphones, high-presence response-sound reproduction and robust elimination of the response sound are realized simultaneously. Since the effect of the error caused by the fluctuation of room acoustics is dispersed to multiple channels, the MOMNI method achieves high robustness through increased numbers of loudspeakers and microphones. With sufficient numbers of loudspeakers and microphones, the proposed method achieves satisfactory robustness against fluctuation. In the experiment, the MOMNI method performed over 20% better than the conventional AEC in word accuracy using 24 loudspeakers and four microphones.

In Chapter 3, by removing strict reproduction at the user's ears from the MOMNI method, I proposed a small-scale version of sound field control to eliminate the response sound at the microphones. The MOMNI method requires many loudspeakers to achieve robust elimination, and its performance degrades when the number of loudspeakers is insufficient because of the large condition number of the inverse filter. To address the problem, I proposed a new filter design method by exploiting the nullspace of the transfer system, refered to as

nullsapce-based sound field control (NBSFC). The NBSFC is more robust than elimination based on the MOMNI method when the number of the loudspeakers is small. In the experiment, the NBSFC performed 15% better than the AEC and the MOMNI method with five loudspeakers.

In Chapter 5, I proposed a new semiblind source separation (SBSS) to separate both known and unknown noise efficiently, and combined it with the MOMNI method. To realize adaptation without DTDs, I adopted independent component analysis (ICA). ICA is generally used for unsupervised adaptation of the beamformer. For the efficient separation of noise from its known source, SBSS exploits the known source as a reference signal. In addition, the unknown sources are separated in parallel with the known-noise separation. SBSS is effective for separating the residual response sound in the MOMNI method with a long impulse response and no specific directivity. Thus, SBSS can reinforce the response sound elimination by the MOMNI method and even reduces unknown noise. In the experiment, the performance of the proposed combination was compared with the ideal behavior of the combination of AEC and ABF. As a result, the proposed combination was found to perform 10% better in word accuracy than the performance limit of AEC and ABF.

6.2 Application

Above-discussed new framework combining sound field control and unsupervised noise reduction realizes noise-robust spoken dialogue system. For settled spoken dialogue system which does not require the high-fidelity reproduction such as guidance systems at community centres, the combination of the NBSFC and SBSS satisfies a robust hands-free spoken dialogue interface efficiently.

In car-navigation system, the combination of the MOMNI method and SBSS meets the demands. To avoid accidents, most of the car navigation systems reject the manipulation when the cars are running. A hands-free spoken dialogue system can avoid such risk. Thus car navigation is one of the most important application of the hands-free spoken dialogue system. In addition, the reproduction of the MOMNI method can reproduce music of car stereo with high fidelity and removes its affect on the speech recognition performance. In this situation, the limitation of the reproduction area is not problematic because the user is always sitting on

the seats.

Another important application of the spoken dialogue system is robot. Speech dialogue is desirable for the manipulation of the robots to assist the user in many situations. We can apply the proposed framework to robots by setting multiple loudspeakers on the robot. Although the robot cannot carry many loudspeakers, this situation may require only few loudspeakers because the microphones can be set close to the loudspeakers and the effect of the reverberation is not sufficient. This should be examined in the future research.

In addition, duplex virtual reality of sound realized by the proposed framework also has wide application, for example, remote communication, remote communication, tele-presence, or entertainment such as computer games.

6.3 Future research

Although I evaluated the performances in only an acoustic environmental room in this dissertation, each of the situation in a particular application could have unique problems and the performance for each of the application should be evaluated. In in-car application, the required number of the loudspeakers is expected to be small because of the low reverberation in a small capacity of a car cabin. In addition, the current car interior already has about ten loudspeakers. However, their irregular positioning may degrade the performance.

Also in robot application, the performance should be examined to reveal the expected reduction of the number of the loudspeakers as discussed above. Another problem of the robot application is the effect of the interior noise caused by machinery of the robot. Application of SBSS with additional sensors for observation of the interior noise is also an interesting topic.

Another remaining issue is the robustness of the sound field reproduction against change of the user's position. In sound field reproduction of the MOMNI method, the robustness of the reproduction against the user's movement must be improved. Similarly to the NBSFC in this dissertation, the nullspace of the transfer system does not affect sound at the the ears of a user in a fixed position. Utilizing this feature, the simultaneous realization of two sound reproductions at a fixed position and in the area outside the fixed position can be realized. For the reproduction of binaural recording composed of two components, the

members of the Speech and Acoustics Laboratory and I have already proposed a system based on this idea [MST+06]. Now one of the master's students in the Laboratory is working on the problem with an arbitrary number of sources. In his preliminary experiment, a subjective assessment of the sound localization indicated good impressions even outside the sweet spot. The problem in the source separation is the real-time and online adaptation of ICA. This is a common challenge of almost all methods based on ICA.

Appendix

A. Derivation of Update Formula

Making $y_1(\omega,t),\ldots,y_L(\omega,t)$ mutually independent is equivalent to minimizing Kullback-Leibler divergence $I_{\mathrm{KL}}(\omega)$ between the joint probability distribution $p\left(\boldsymbol{y}(\omega,t)\right)$ and the product of marginal probability distributions $\prod_{l=1}^{L} p\left(y_l(\omega,t)\right)$ [ACY96, Sma98], where $I_{\mathrm{KL}}(\omega)$ is described as

$$\begin{aligned} I_{\mathrm{KL}}(\omega) &= \int p\left(\boldsymbol{y}(\omega,t)\right) \log \frac{p\left(\boldsymbol{y}(\omega,t)\right)}{\prod_{l=1}^{L} p\left(y_l(\omega,t)\right)} d\boldsymbol{y}(\omega,t) \\ &= E\left[\log p\left(\boldsymbol{y}(\omega,t)\right)\right]_t - E\left[\sum_{l=1}^{L} \log p\left(y_l(\omega,t)\right)\right]_t . \end{aligned} \tag{93}$$

The partial differential of $I_{\mathrm{KL}}(\omega)$ by $\bar{\boldsymbol{W}}(\omega)$ can be written as

$$\begin{aligned} \frac{\partial I_{\mathrm{KL}}(\omega)}{\partial \bar{\boldsymbol{W}}(\omega)} =& E\left[\frac{\partial}{\partial \bar{\boldsymbol{W}}(\omega)} \log p\left(\boldsymbol{y}(\omega,t)\right)\right] \\ & - \left[\frac{\partial}{\partial \bar{\boldsymbol{W}}(\omega)} \sum_{l=1}^{L} \log p\left(y_l(\omega,t)\right)\right]_t . \end{aligned} \tag{94}$$

The first term in the right side of Eq. (94) can be written as

$$\begin{aligned} \frac{\partial \log p\left(\boldsymbol{y}(\omega,t)\right)}{\partial \bar{\boldsymbol{W}}(\omega)} =& \frac{\partial}{\partial \bar{\boldsymbol{W}}(\omega)} \log\left(\frac{p\left(\boldsymbol{x}(\omega,t)\right)}{|J(\omega)|}\right) \\ =& \frac{\partial}{\partial \bar{\boldsymbol{W}}(\omega)} \left\{\log\left(p\left(\boldsymbol{x}(\omega,t)\right)\right) - |\det\left[\boldsymbol{W}(\omega)\right]|\right\} \\ =& -\frac{\partial}{\partial \bar{\boldsymbol{W}}(\omega)} |\det\left[\boldsymbol{W}(\omega)\right]| , \end{aligned} \tag{95}$$

where "det" denotes the determinant and $J\left(\omega\right)$ denotes the Jacobian of the variable transform from $\boldsymbol{x}(\omega,t)$ to $\boldsymbol{y}(\omega,t)$. Here, the following relation is well known:

$$\frac{\partial \log p\left(\boldsymbol{y}(\omega,t)\right)}{\partial \boldsymbol{W}(\omega)} = \boldsymbol{W}^{-\mathrm{H}}(\omega), \tag{96}$$

where $\{\cdot\}^{-\mathrm{H}}$ denotes the inversion of the conjugate transposition. Since the differential $\partial/\partial\,\bar{\boldsymbol{W}}(\omega)$ is a truncated submatrix of $\partial/\partial\,\boldsymbol{W}(\omega)$, The following condition

satisfies:

$$\frac{\partial}{\partial \bar{\boldsymbol{W}}(\omega)} = [\boldsymbol{I}_{L-1}, \boldsymbol{0}] \frac{\partial}{\partial \boldsymbol{W}(\omega)}, \tag{97}$$

where $\boldsymbol{0}$ denotes the zero vector. Substituting Eq. (97) in Eq. (96) obtains

$$\frac{\partial \log p\left(\boldsymbol{y}(\omega, t)\right)}{\partial \bar{\boldsymbol{W}}(\omega)} = [\boldsymbol{I}_{L-1}, \boldsymbol{0}]\ \boldsymbol{W}^{-\mathrm{H}}(\omega) \tag{98}$$

Also for the second term in Eq. (94), the following approximation is proposed in [SMAM03];

$$\frac{\partial}{\partial \boldsymbol{W}(\omega)} \sum_{l=1}^{L} \log p\left(y_l(\omega, t)\right) = \boldsymbol{\Phi}(\boldsymbol{y}(\omega, t)) \boldsymbol{x}^{\mathrm{H}}(\omega, t). \tag{99}$$

Substitution of Eq. (97) in Eq. (99) obtains

$$\frac{\partial}{\partial \bar{\boldsymbol{W}}(\omega)} \sum_{l=1}^{L} \log p\left(y_l(\omega, t)\right) = \boldsymbol{\Phi}(\bar{\boldsymbol{y}}(\omega, t)) \boldsymbol{x}^{\mathrm{H}}(\omega, t). \tag{100}$$

Substituting Eqs. (98) and (100), Eq. (94) can be rewritten as

$$\begin{aligned} \frac{\partial I_{\mathrm{KL}}(\omega)}{\partial \bar{\boldsymbol{W}}(\omega)} &= E\left[[\boldsymbol{I}_{L-1}, \boldsymbol{0}]\ \boldsymbol{W}^{-\mathrm{H}}(\omega) - \boldsymbol{\Phi}(\bar{\boldsymbol{y}}(\omega, t)) \boldsymbol{x}^{\mathrm{H}}(\omega, t)\right]_t \\ &= [\boldsymbol{I}_{L-1}, \boldsymbol{0}]\ \boldsymbol{W}^{-\mathrm{H}}(\omega) - E\left[\boldsymbol{\Phi}(\bar{\boldsymbol{y}}(\omega, t)) \boldsymbol{x}^{\mathrm{H}}(\omega, t)\right]_t . \end{aligned} \tag{101}$$

By applying the natural gradient of $\boldsymbol{W}(\omega)$ [ACY96], the update of $\bar{\boldsymbol{W}}(\omega)$ can be written as

$$\begin{aligned} \bar{\boldsymbol{W}}(\omega) \leftarrow\ & \bar{\boldsymbol{W}}(\omega) - \eta \frac{\partial I_{\mathrm{KL}}(\omega)}{\partial \bar{\boldsymbol{W}}(\omega)} \boldsymbol{W}^{\mathrm{H}}(\omega)\, \boldsymbol{W}(\omega) \\ =\ & \bar{\boldsymbol{W}}(\omega) - \eta \big\{ \bar{\boldsymbol{W}}(\omega) - E\big[\boldsymbol{\Phi}\big(\bar{\boldsymbol{y}}(\omega, t)\big) \boldsymbol{y}^{\mathrm{H}}(\omega, t)\big]_t \boldsymbol{W}(\omega) \big\}. \end{aligned} \tag{102}$$

Assuming the ergodicity of the sources, the expectation can be substituted by the time average and the update formula Eq. (90) is obtained.

B. AEC and ABF with Ideal DTD Used in Experiments in Chapter 5

In this section I describe the adaptation of AEC2 and ABF in the experiments. The configuration of the combination of AEC2 and ABF is shown in Fig. 45.

B.1 AEC with ideal DTD

Here I describe the adaptation of AEC with ideal frequency-domain DTD used in the experiments. The filter coefficient $\hat{g}(\omega)$ can be estimated using frequency-domain batch adaptation with the ideal frequency-domain DTD. Since the human speech is sparse in time-frequency domain, there are many time-frequency grids where the response sound is much larger than the user's utterance. Therefore, by finding them with frequency-domain DTD, AEC can be adapted in frequency domain even if there is no full-band single-talk duration of the response sound. Assuming a DTD works without error, I give the AEC2 the durations $t \in \mathcal{T}(\omega)$ when the power ratio of the response sound and the user's speech exceeds the threshold $T(\omega)$ as

$$t \in \mathcal{T}(\omega) \quad \text{if} \quad \frac{|g(\omega) r_{\mathrm{src}}(\omega, t)|^2}{|x_{\mathrm{noresp}}(\omega, t)|^2} > T(\omega), \tag{103}$$

where $x_{\mathrm{noresp}}(\omega, t)$ denotes the observed signal component of the user's utterance. It should be noted that such an accurate DTD without error cannot be implemented in real world; this corresponds to *ideal DTD* for the AEC. Using the result of DTD, the filter is adapted by frame-analysis version of Eq. (6) as

$$\hat{g}(\omega) = -\frac{\left\langle d(\omega, t) r^*_{\mathrm{src}}(\omega, t) \right\rangle_{t \in \mathcal{T}(\omega)}}{\left\langle |r_{\mathrm{src}}(\omega, t)|^2 \right\rangle_{t \in \mathcal{T}(\omega)}}. \tag{104}$$

This corresponds to a batch learning, and is more robust than on-line learning methods. Using this batch learning and the ideal DTD, I evaluate the performance limit of the AEC of DTD framework.

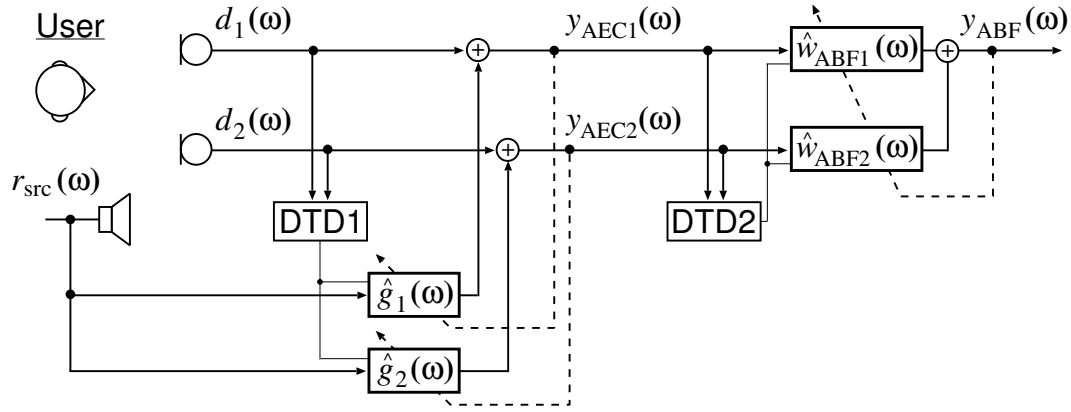

Figure 45. Configuration of the adaptation of the AECs and ABF using ideal frequency-domain DTDs.

B.2 Combination of AEC and ABF using ideal DTDs

Here I describe the combination of the AEC and the ABF using ideal DTDs. As the adaptation algorithm of ABF, I adopt the minimum variance ABF [Fro72]. Since the minimum variance ABF is a batch adaptation algorithm, it consistently outperforms on-line adaptation algorithms like generalized sidelobe canceller [GJ82]. Combined with ideal DTDs, the performance limit of the supervised adaptation based on AEC and ABF can be evaluated. First, I obtain processed signals of AEC $\boldsymbol{y}_{\mathrm{AEC}}(\omega,t) = [y_{\mathrm{AEC}1}(\omega,t),\ldots,y_{\mathrm{AEC}K}(\omega,t)]^{\mathrm{T}}$ for each of the K microphone elements $k = 1,\ldots,K$. Then, to distinguish the single-talk of the interfering noise in $\boldsymbol{y}_{\mathrm{AEC}}(\omega,t)$, an ideal DTD for ABF (DTD2) detects the times $t \in \mathcal{T}_2(\omega)$ when the power ratio of the component of the interfering noise and the other components exceeds a threshold $T(\omega)$, similarly to Eq. (103). A covariance matrix $\boldsymbol{R}_{\mathrm{AEC}}(\omega)$ of AEC outputs $\boldsymbol{y}_{\mathrm{AEC}}(\omega,t)$ in $t \in \mathcal{T}_2(\omega)$ can be obtained as

$$\boldsymbol{R}_{\mathrm{AEC}}(\omega) = \langle \boldsymbol{y}_{\mathrm{AEC}}(\omega,t)\boldsymbol{y}_{\mathrm{AEC}}^{\mathrm{H}}(\omega,t)\rangle_{t\in\mathcal{T}_2(\omega)}. \tag{105}$$

Finally, the optimal $\boldsymbol{w}_{\mathrm{ABF}}(\omega)$ is obtained by applying the adaptation in Eq. (15) to $\boldsymbol{R}_{\mathrm{AEC}}(\omega)$ as

$$\boldsymbol{w}_{\mathrm{ABF}}(\omega) = \frac{\boldsymbol{q}^{\mathrm{H}}(\omega)\boldsymbol{R}_{\mathrm{AEC}}^{-1}(\omega)}{\boldsymbol{q}^{\mathrm{H}}(\omega)\boldsymbol{R}_{\mathrm{AEC}}^{-1}(\omega)\boldsymbol{q}(\omega)}. \tag{106}$$

Acknowledgements

This dissertation is a summary of studies for four years and half carried out at Graduate School of Information Science, Nara Institute of Science and Technology, Japan.

I would like to express my foremost gratitude to Professor Kiyohiro Shikano, my thesis adviser , for his valuable guidance and constant encouragement through my master's course and doctoral course. His insight played a significant role in my study.

I would also like to express my sincere thanks to a member of the thesis committee, Professor Hirokazu Nishitani, for his valuable comments to the master's and doctoral theses.

I would especially like to express genuine gratitude to Associate Professor Hiroshi Saruwatari for his continuous support and valuable advice. Without his guidance, this work could not have been completed. I am always happy to carry out research with him.

I would also like to thank to Assistant Professor Hiromichi Kawanami for his beneficial comments. I could also like to thank Assistant Professor Tomoki Toda for invaluable discussion. I could learn many lessons from his attitude toward study.

I want to thank all members of Speech and Acoustics Laboratory in Nara Institute of Science and Technology for providing fruitful atmosphere with many productive discussions. I would especially like to express my appreciation to Dr. Akinobu Lee, who is currently Associate Professor at Nagoya Institute of Technology, Dr. Yosuke Tatekura, who is currently Assistant Professor at Shizuoka University, Dr. Ryuichi Nishimura, who is currently Assistant Professor at Wakayama University, Dr. Tsuyoki Nishikawa, who is a researcher in Matsushita Electric Industrial Co., Ltd, Dr. Tomoya Takatani, who is a researcher in Toyota Motor Corporation, Dr. Randy Gomez, who is a Post-Doctorate Fellow in Nara Institute of Science and Technology, Mrs. Toshie Nobori, who is a security in Speech and Acoustics Laboratory, Dr. Akira Nakayama, who is a researcher in NTT Cyber Solution Laboratories, Mr. Goshu Nagino, who is a researcher in Asahi Kasei Corporation, Mr. Tobias Cincarek, Yamato Otani and Keigo Nakamura, who are Ph. D candidates of Nara Institute of Science and Technology.

I would like to thank Dr. Jani Even, who is currently a Post-Doctorate Fellow in Nara Institute of Science and Technology, Mr. Tatsunori Asai, who is currently a researcher in Brother Industries, Ltd., Mr. Yusuke Kaibara, who is currently a researcher in Clarion Co., Ltd., Mr. Masayuki Shimada, who is currently a researcher in TOA Corporation, Mr. Yoshimitsu Mori and Mr. Yu Takahashi, who are Ph. D candidates of Nara Institute of Science and Technology, Tadashi Mihashi, who is currently a researcher in Sharp Corporation, who is currently a researcher in Fujitsu Ltd., Yuki Haraguchi, Yuta Yuyama, Kentaro Tachibana, and Keiichi Osako, who are currently in master's course of Nara Institute of Science and Technology, for providing valuable discussion and suggestions about sound field reproduction and independent component analysis.

I would also like to express my gratitude to all of my friends for their support. I would especially like to express my gratitude to Mr. Satoshi Ukai, Mr. Yasuaki Ohashi, Mr. Yasumichi Omoto, Ms. Sachiko Obara, Mr. Mitsuru Samejima, Mr. Koji Takenae, Mrs. Noriko Takenae, Mr. Takeo Higashi, Mr. Daisuke Matsumoto, and Mr. Masashi Yamada.

Finally, I would like to thank my father Noritaka Miyabe, my mother Shihomi Miyabe, and my brother Kazuki Miyabe, for their supports in daily life to carry on this study.

References

[ACY96] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," *Advances in Neural Information Processing Systems,* vol. 8, pp. 757–763, MIT Press, Cambridge MA, 1996.

[AMM+03] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech & Audio Process.*, vol. 11, no. 2, pp. 109–116, 2003.

[AOO77] T. Araseki, K. Ochiai and T.Ogihara, "Echo canceller with two echo path models," *IEEE Trans. Communications,* vol. 25, pp. 589–595, 1977.

[BAK05] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 120–134, 2005.

[BBGK03] H. Buchner, J. Benesty, T. Gaensler, and W. Kellermann, "An outlier-robust extended multidelay filter with application to acoustic echo cancellation," in *Proc. IEEE Int. Workshop on Acoustic Echo and Noise Control,* 2003, pp. 19–22.

[BBK05] H. Buchner, J. Benesty, and W. Kellermann, "Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication," *Signal Process.*, vol. 85, no. 3, pp. 549–570, 2005.

[BC96] J. Bauck and D. H. Cooper, "Generalized transaural stereo and applications", *J.Audio Eng. Soc.,* vol. 44, no. 9, pp. 148–151, Sept. 1996.

[BDH+99] C. Breining, P. Dreitseitel, E. Hänsler, A. Mader, B. Nitsch, H. Pudeer, T. Scheirtler, G. Schmidt, and J.Tilp, "Acoustic echo control — An application of very high order adaptive filters," *IEEE Signal Process. Mag.,* pp. 42–69, July 1999.

[Bla97] J. Blauert, *Spatial Hearing (Revised edition),* MIT Press, Cambridge, MA, 1997.

[BMC99] J. Benesty, D. R. Morgan and J. H. Cho, "A family of doubletalk detectors based on cross-correlation," *Proc. International Workshop on Acoustic Echo and Noise Control,* pp. 108–111, Sept. 1999.

[BSK04] 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.4 pp. 117–120, May 2004.

[CNLS07] T. Cincarek, R. Nisimura, A. Lee, and K. Shikano, "Insights Gained from Development and Long-Term Operation of a Real-Environment Speech-Oriented Guidance System," *Proc. International Conference on Acoustics, Speech, and Signal Processing,* vol. 4, pp. 157–160, Apr. 2007.

[Com94] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.

[CSO72] S. Campanella, H. Suyderhoud, M. Onufry, "Analysis of an adaptive impulse response echo canceller," *Comsat Tech. Rev.,* vol.2, pp.1–36, 1972.

[DHP93] J. R. Deller Jr., J. H. L. Hansen and J. G. Proakis, *Discrete-Time Processing of Speech Signals,* Macmillan, New York, 1993.

[ES07] J. Even, K. Sugimoto, "An ICA approach to semi-blind identification of strictly proper systems based on interactor polynomial matrix," *International Journal of Robust and Nonlinear Control,* vol. 17, no. 8, pp. 752–768, 2007.

[FJZE85] J. L. Flanagan, J. D. Johnston, R. Zahn and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms", *J. Acoust. Soc. Am.,* vol. 78, no. 5, pp. 1508–1518, May, 1985.

[Fro72] O. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE,* vol. 60, pp. 926–935, 1972.

[Fur84] I. Furukawa, "A design of canceller of broad band acoustic echo," *Int. Teleconference Symp.,* pp.1/8–8/8, 1984.

[GB01] T. Gänsler and J. Benesty, "A frequency-domain double-talk detector based on a normalized cross-correlation vector," *Signal Process.,* vol. 81, pp. 1783–1787, Aug. 2001.

[GB06] T. Gänsler and J. Benesty, "The fast normalized cross-correlation double-talk detector," *Signal Process.,* vol. 86, pp. 1124–1139, 2006.

[GJ82] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propagation,* vol. 30, no. 1, pp. 27–34, 1982.

[Han01] E. Hänsler, "Acoustic echo and noise control: Where do we come from — where do we go?," *Proc. International Workshop on Acoustic Echo and Noise Control,* pp. 1–4, Sept. 2001.

[Hay91] S. Haykin, *Adaptive Filter Theory (4th edition),* Prentice Hall, Englewood Cliffs, NJ, 1991.

[HBK03] W. Herbordt, H. Buchner, and W. Kellermann, "An acoustic human-machine front-end for multimedia applications," *EURASIP Journal on Applied Signal Processing,* vol. 2003, no. 1, pp. 1-11, 2003.

[HBNK05] W. Herbordt, H. Buchner, S. Nakamura, W. Kellermann, "Outlier-robust DFT-domain adaptive filtering for bin-wise stepsize controls, and its application to a generalized sidelobe canceller," *Proc. Int. Workshop on Acoustic Echo and Noise Control,* pp.113–116, 2005

[Her04] W. Herbordt, "Combination of robust adaptive beamforming with acoustic echo cancellation for acoustic human/machine interfaces," PhD thesis, University Erlangen-Nuremberg, Germany, January 2004.

[HIKT93] 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, Jan. 1993.

[HK02] W. Herbordt and W. Kellermann, "Frequency-domain integration of acoustic echo cancellation and a generalized sidelobe canceller with improved

robustness," *European Trans. on Telecommunications,* vol. 13, no. 2, pp. 123–132, 2002.

[HNK05] W. Herbordt, S. Nakamura, and W. Kellermann, "Joint optimization of LCMV beamforming and acoustic echo cancellation for automatic speech recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 3, pp.77–80, 2005

[IM99] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation*, 1999, pp. 365–371.

[IYT+98] 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. International Conf. on Spoken Language Processing,* vol. 7, pp. 3261–3264, Dec. 1998.

[IYT+99] 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.

[JLPY00] Y. W. Jung, J. H. Lee, Y. C. Park and D. H. Youn, "A new adaptive algorithm for stereophonic acoustic echo canceller," *Proc. International Conf. on Acoustics, Speech and Signal Processing,* vol. 2 pp. 801–804, Istanbul, Turkey, June 2000.

[JMM01] M. Joho, H. Mathis, and G. Moschytz, "Combined blind/nonblind source separation based on the natural gradient," *IEEE Signal Processing Letters,* vol. 8, no. 8, pp. 236–238, 2001.

[JS01] B. H. Juang and F. K. Soong, "Hands-free telecommunications," *Proc. International Workshop on Hands-Free Speech Communication,* pp. 5–8, Kyoto, Japan, April 2001.

[Kel84] W. Kellerman, "Kompensation akustischer Echos in Frequenzteilbändern," *Aachener Kolloquium 1984,* Aachen, FRG, pp. 322–325, 1984.

[KIS+73] Y. Kato, S. Chiba, T. Ishiguro, Y. Sato, M. Tajima, T. Ohihara, S. Campanella, H. Suyderhoud, and M. Onufry, "A digital adaptive echo canceller," *NEC Res. Dev.,* vol.31, pp.32–41, 1973.

[KO86] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust. Speech Signal Process.,* vol. ASSP-34, no. 6, pp. 1391–1400, 1986.

[KYHS97] H. Kikuchi, M. Yokoyama, K. Hoashi, K. Shirai, "The Role of Non-Verbal Information in Spoken Dialogue between a Man and a Robot," Proc. International Conference on Signal Processing, vol. 2, pp. 539–544, Aug., 1997.

[Lee98] T.-W. Lee, *Independent Component Analysis.* Norwell, MA: Kluwer Academic Publishers, 1998.

[LKK01] A. Lee, T. Kawahara and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," *Proc. European Conf. on Speech Communication and Technology,* pp. 1691–1694, Aalborg, Denmark, Sept. 2001.

[LKTS00] 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, and Signal Processing*, vol.III, pp. 1269–1272, June 2000.

[LMF78] L. Ljung, M. Morf, and D. Falcorner, "Fast calculation of gain matrices for recursive estimation schemes," *Int. J. Control,* vol. 27, pp.1–19, 1978.

[MHS+07] S. Miyabe, Y. Hinamoto, H. Saruwatari, K. Shikano, and Y. Tatekura, "Interface for barge-in free spoken dialogue system based on sound field reproduction and microphone array," *EURASIP J. Advances in Signal Processing,* vol. 2007 , Article ID 57470, 13 pages, 2007.

[MK88] 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, Feb. 1988.

[MS99] 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.

[MST+06] S. Miyabe, M. Shimada, T. Takatani, H. Saruwatari, and K. Shikano, "Multi-Channel Inverse Filtering with Loudspeaker Selection and Enhancement for Robust Sound Field Reproduction," *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC2006),* Sept. 2006 (CD-ROM).

[MTM+06] S. Miyabe, T. Takatani, Y. Mori, H. Saruwatari, K. Shikano and Y. Tatekura, "Double-talk free spoken dialogue interface combining sound field control with semi-blind source separation," *Proc. International Conference on Acoustics, Speech, and Signal Processing,* vol.1 pp. 809–812, 2006.

[MTS+07] S. Miyabe, T. Takatani, H. Saruwatari, K. Shikano, and Y. Tatekura, "Barge-in- and noise-free spoken dialogue interface based on sound field control and semi-blind source separation," in *Proc. European Signal Processing Conference,* vol. 1, pp. 45–48, April 2007.

[NSS03] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," *IEICE Trans. Fundamentals*, vol. E86-A, no. 4, pp. 846–858, 2003.

[POL03] H. M. Park, S. H. Oh, and S. Y. Lee, "A filter bank approach to independent component analysis and its application to adaptive noise cancelling," *Neurocomputing*, vol. 55, no. 3–4, pp. 755-759, 2003.

[PS00] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech & Audio Process.*, vol. 8, pp. 320–327, 2000.

[RJ93] L. Rabiner and B. H. Juang, Fundamentals of speech recognition, Prentice Hall, Englewood Cliffs, NJ, 1993.

[SA95] Y. Suzuki, F. Asano, H.-Y. Kim and Toshio Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Am.,* vol. 97, no. 2, pp. 1119–1123, 1995.

[SB80] M. Sondhi and D. Berkley, "Silencing echoes on the telephone networks," *Proc. IEEE,* no. 64, pp. 1583–1597, 1980.

[SHL+98] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, V. Zue, "Galaxy-II: a reference architecture for conversational system development," *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP98),* pp.931–934, 1998.

[SJH01] A. Sugiyama, Y. Joncour, A. Hirano, "A stereo echo canceller with correct echo-path identification on an input-sliding technique," *IEEE Trans. Signal Process.,* vol. 49, no. 11, pp. 2577–2587, Nov. 2001.

[SKN+06] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Speech Audio Process.*, vol. 14, pp. 666–678, 2006.

[Son67] M. Sondhi, "An adaptive echo canceller," *Bell Syst. Tech. J.,* no.64, vol. 3, pp.497–511, 1980.

[SP00] S. Seneff and J. Polifroni, "Dialogue Management in the Mercury Flight Reservation System," *Proc. NASLPNAACL Satellite Workshop,* pp. 1–6, 2000.

[SKT+03] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing,* vol. 2003, No. 11, pp. 1135–1146, 2003.

[Sma98] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[SMAM03] H. Sawada, R. Mukai, S. Aaraki, and S. Makino, "Polar coordinate based on nonlinear function for frequency domain blind source separation," *IEICE Trans. Fundamentals*, vol. E86-A, no. 3, pp. 590–596, 2003.

[SMAM04] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind

source separation," *IEEE Trans. Speech and Audio Process,*, vol. 12, no. 5, pp. 530–538, 2004.

[TNSS04] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "High-fidelity blind separation of acoustic signals using SIMO-model-based independent component analysis," *IEICE Trans. Fundamentals*, vol. E87-A, no. 8, pp. 2063–2072, 2004.

[TSS01] Y. Tatekura, H. Saruwatari and K. Shikano, "An iterative inverse filter design method for the multichannel sound field reproduction system," *IEICE Trans. Fundamentals,* vol. 84-A, no. 4, pp. 991–998, April 2001.

[TSS02] Y. Tatekura, H. Saruwatari and K. Shikano, "Sound reproduction system including adaptive compensation of temperature fluctuation effect for broadband sound control", *IEICE Trans. Fundamentals,* vol. E85-A, no. 8, pp. 1851–1860, August 2002.

[WEWH04] S. Werner, M. Eichner, M. Wolff, and R. Hoffmann, "Toward spontaneous speech synthesis—utilizing language modelinformation in TTS," *IEEE Transactions on Speech and Audio Processing,* vol.12, no.4, pp.436–446, 2004.

[WGM+75] B. Widrow, J. Glover, J. McCool, J. Kaunitz, C. Williams, R. Hearn, J. Zeidler, E. Dong, and R. Goodlin, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1715, 1975.

[WH60] B. Widrow and M. Hoff, "Adaptive switching circuits," *IRE WESCON Conv. Rec.,* pt.4, pp.64–104, 1960.

[XZF03] O. Xiongbing, C. Zhe, Y. Fuliang, "An echo canceller based on the structure of dual-auxiliary filters," *Proceedings of 2003 International Workshop on Acoustic Echo and Noise Control (IWAENC2003),* pp.35–38, 2003.

[YLSS03] S. Yamade, A. Lee, H. Saruwatari and K. Shikano, "Unsupervised speaker adaptation based on HMM sufficient statistics in various noisy environments," *Proc. European Conf. on Speech Communication and Technology,* vol.2, pp. 1493–1496, Sept. 2003.

[YW91] H. Ye and B. Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Trans. Commun.,* vol. 39, no. 11, pp. 1542–1545, Nov. 1991.

[ZSG+00] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T.J. Hazen, L. Hetherington, "JUPITER: a telephone-based conversational interface for weather information," *IEEE Transactions on Speech and Audio Processing,* vol.8, no.1, pp.100–112, 2000.

List of Publications

Journal papers

1. Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, Yosuke Tatekura, "Interface for Barge-in Free Spoken Dialogue System Using Nullspace based Sound Field Control and Beamforming," *IEICE Transactions on Fundamentals,* vol.E89-A, no.3, 2006.

2. Shigeki Miyabe, Yoichi Hinamoto, Hiroshi Saruwatari, Kiyohiro Shikano, Yosuke Tatekura, "Interface for barge-in free spoken dialogue system based on sound field reproduction and microphone array," *EURASIP Journal on Advances in Signal Processing,* vol.2007, Article ID 57470, pp.1–13, 2007.

International conference

1. Shigeki Miyabe, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "Barge-in- and Noise-Free Spoken Dialogue Interface Based on Sound Field Control and Semi-Blind Source Separation," *Proc. 15th European Signal Processing Conference (EUSIPCO2007),* September 2007 (accepted).

2. Kentaro Tachibana, Hiroshi Saruwatari, Yoshimitsu Mori, Shigeki Miyabe, Kiyohiro Shikano, Akira Tanaka, "Efficient blind source separation combining closed-form second-order ICA and nonclosed-form higher-order ICA," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP2006),* 2007 (accepted).

3. Shigeki Miyabe, Masayuki Shimada, Tomoya Takatani, Hiroshi Saruwatari, and Kiyohiro Shikano, "Multi-Channel Inverse Filtering with Loudspeaker Selection and Enhancement for Robust Sound Field Reproduction," *Proceedings of 10th International Workshop on Acoustic Echo and Noise Control (IWAENC2006),* September 2006 (CD-ROM).

4. Shigeki Miyabe, Masayuki Shimada, Tomoya Takatani, Hiroshi Saruwatari, and Kiyohiro Shikano, "Multi-Channel Inverse Filtering with Loudspeaker

Selection and Enhancement for Robust Sound Field Reproduction," *Proceedings of 10th International Workshop on Acoustic Echo and Noise Control (IWAENC2006),* September 2006 (CD-ROM).

5. Shigeki Miyabe, Tomoya Takatani, Yoshimitsu Mori, Hiroshi Saruwatari, Kiyohiro Shikano, Yosuke Tatekura, "Double-Talk Free Spoken Dialogue Interface Combining Sound Field Control with Semi-Blind Source Separation," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP2006),* vol.1, pp.809–812, May 2006.

6. Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, Yosuke Tatekura, "Barge-in Free Spoken Dialogue Interface Using Nullspace-Based Sound Field Control and Beamforming," *Proceedings of 13th European Signal Processing Conference (EUSIPCO2005),* September 2005 (CD-ROM).

7. Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, Yosuke Tatekura, "Speech Enhancement Using Nullspace-Based Sound Field Control for Barge-in Free Spoken Dialogue Interface," *Proceedings of the 2005 IEEE Workshop on Statistical Signal Processing (SSP '05),* #447, July 2005 (CD-ROM).

8. Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, "Barge-in Free Spoken Dialogue Interface Based on Response Sound Cancellation Using Sound field Control and Microphone Array," *Proceedings of Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA2005),* pp.a-7–8, March 2005.

9. Tatsunori Asai, Shigeki Miyabe, Hiroshi Saruwatari, and Kiyohiro Shikano, "Interface for barge-in free spoken dialogue system using adaptive sound field control," *Proc. 8th International Conference on Spoken Language Processing (ICSLP2004),* FrA1502p-9, vol.4, pp.2665–2668, October 2004.

Invited paper

1. Hiroshi Saruwatari, Masayuki Shimada, Shigeki Miyabe, Tomoya Takatani, and Kiyohiro Shikano, "Robust sound field reproduction using multi-channel

inverse filtering with secondary source selection and enhancement," *Proceedings of 9th Western Pacific Acoustics Conference (WESPAC IX 2006),* June 2006.

2. Yoshimitsu Mori, Keiichi Osako, Shigeki Miyabe, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, "MLSP 2007 Data Analysis Competition: Two-Stage Blind Source Separation Combining SIMO-Model-Based ICA and Binary Masking," *2007 IEEE International workshops on Machine Learning for Signal Processing (MLSP2007),* Aug. 2007.

Technical report

1. Kentaro tachibana, Hiroshi Saruwatari, Yoshimitsu Mori, Shigeki Miyabe, Kiyohiro Shikano, and Akira Tanaka, "Blind source separation using closed-form second-order ICA and nonclosed-form higher-order ICA," *IEICE Technical Report*, EA2006-95, pp.37–42, December 2006 (in Japanese).

2. Yuki Yai, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "Fast Compensation of Temperature Fluctuation Effect for the Multichannel Sound Field Reproduction System," *IEICE Technical Report*, EA2006-26, pp.7–12, July 2006 (In Japanese).

3. Tadashi Mihashi, Tomoya Takatani, Shigeki Miyabe, Yoshimitsu Mori, Hiroshi Saruwatari, and Kiyohiro Shikano, "Multichannel Audio Signal Compressive Coding Method with Independent Component Analysis," *IEICE Technical Report*, EA2006-28, pp.21–26, July 2006 (In Japanese).

4. Masayuki Shimada, Shigeki Miyabe, Tomoya Takatani, Hiroshi Saruwatari, and Kiyohiro Shikano, "Subjective evaluation of multi-channel inverse filter with secondary source selection and enhancement," *IEICE Technical Report,* no.EA2006-4, pp.19–24, April 2006.

5. Shigeki Miyabe, Tomoya Takatani, Yoshimitsu Mori, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "Spoken dialogue interface using sound field control and source separation," *IEICE Technical Report, no.EA2006-5,* pp.25–30, April 2006.

6. Masayuki Shimada, Shigeki Miyabe, Tomoya Takatani, Hiroshi Saruwatari, and Kiyohiro Shikano, "Robust sound field reproduction against user's move based on multi-channel inverse filtering with secondary source selection and enhancement," *IEICE Technical Report,* EA2005-27, pp.13-18, July 2005 (in Japanese).

7. Yusuke Kaibara, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "Minimum Error Relaxation Algorithm of Inverse Filter in Multi-Channel Sound Reproduction System," *IEICE Technical Report,* EA2005-97, pp.7-11, January 2006 (in Japanese).

8. Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, "Speech Recognition Using Barge-in Free Spoken Dialogue InterfaceBased on Multi-Channel Sound Field Control," *IEICE Technical Report,* no.EA2004-117, pp.1–6, January 2005.

9. Shigeki Miyabe, Hiroshi Saruwatari, and Kiyohiro Shikano, "Small-scale barge-in free spoken dialogue interface using response sound cancellation by multichannel sound field control," *IEICE Technical Report,* EA2004-33, pp.19–24, August 2004 (in Japanese).

Meetings

1. Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "Evaluation of spoken dialogue interface using sound field control and semi-blind source separation," *Spring Meeting of Acoust. Soc. Jpn.,* 2-1-17, pp.579–580, March 2007 (in Japanese).

2. Tadashi Mihashi, Tomoya Takatani, Shigeki Miyabe, Yoshimitsu Mori, Hiroshi Saruwatari, and Kiyohiro Shikano, "Evluations of Stereo Audio Signal Compressive Coding with Independent Component Analysis," *Spring Meeting of Acoust. Soc. Jpn.,* 1-1-5, pp.501–502, March 2007 (in Japanese).

3. Yuki Haraguchi, Tadashi Mihashi, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, "Sound-Localization Control of Audio Objects Based on

Blind Souce Separation," *Spring Meeting of Acoust. Soc. Jpn.,* 1-1-6, pp.503–504, March 2007 (in Japanese).

4. Yuta Yuyama, Shigeki Miyabe, Hiroshi Saruwatari, and Kiyohiro Shikano, "Sound Quality Improvement in Robust Sound Field Reproduction Based on Multi-Channel Inverse Filtering with Selected and Enhanced Secondary Sources," *Spring Meeting of Acoust. Soc. Jpn.,* 1-1-16, pp.523–524, March 2007 (in Japanese).

5. Kentaro Tachibana, Hiroshi Saruwatari, Yoshimitsu Mori, Shigeki Miyabe, Kiyohiro Shikano, and Akira Tanaka, "Efficient Blind Source Separation Using Closed-Form Second-Order ICA in Combination with Nonclosed- Form Higher-Order ICA," *Spring Meeting of Acoust. Soc. Jpn.,* 3-P-7, pp.611–612, March 2007 (in Japanese).

6. Yuki Yai, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "Efficient Method for Adaptive Compensation of Temperature Fluctuation Effect in Multichannel Sound Field Reproduction System," *Spring Meeting of Acoust. Soc. Jpn.,* 3-Q-19 , pp.687–688, March 2007 (in Japanese).

7. Shigeki Miyabe, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "A front-end for spoken dialogue system using sound field control and source separation, " *KJCIEE 2006,* G16-15, November 2006 (in Japanese).

8. Yuki Yai, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "Efficient Compensation of Temperature Fluctuation Effect for Multichannel Sound Field Reproduction System," *Fall Meeting of Acoust. Soc. Jpn.,* 2-1-16, pp. 467–468, September 2006 (in Japanese).

9. Tadashi Mihashi, Tomoya Takatani, Shigeki Miyabe, Yoshimitsu Mori, Hiroshi Saruwatari, and Kiyohiro Shikano, "Stereo Audio Signal Compressive Coding with Independent Component Analysis and Projection Back Method,", *Fall Meeting of Acoust. Soc. Jpn.,* 2-1-8, pp. 451–452, September 2006 (in Japanese).

10. Shigeki Miyabe, Yuuki Yai, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "Evaluation of spoken dialogue interface using sound field control and semi-blind source separation," *Fall Meeting of Acoust. Soc. Jpn.,* 3-Q-12, pp.523–524, September 2006 (in Japanese).

11. Masayuki Shimada, Shigeki Miyabe, Tomoya Takatani, Hiroshi Saruwatari, and Kiyohiro Shikano, "Evaluation of Sound Field Reproduction Based on Multi-Channel Inverse Filtering with Selected and Enhanced Secondary Sources," *Spring Meeting of Acoust. Soc. Jpn.,* 3-Q-27, pp.681–682, March 2006 (in Japanese).

12. Yusuke Kaibara, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "Multi-channel sound filed reproduction with minimum error relaxed inverse filters and its evaluation," *Spring Meeting of Acoust. Soc. Jpn.,* 3-Q-28, pp.683–684, March 2006 (in Japanese).

13. Tadashi Mihashi, Tomoya Takatani, Shigeki Miyabe, Yoshimitsu Mori, Hiroshi Saruwatari, and Kiyohiro Shikano, "Multichannel Audio Signal Compressive Coding with Blind Source Separation," *Spring Meeting of Acoust. Soc. Jpn.,* 1-5-10, pp.525–526, March 2006 (in Japanese).

14. Shigeki Miyabe, Tomoya Takatani, Yoshimitsu Mori, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "Barge-in free spoken dialogue interface using sound field control and blind source seperation," *Workshop for Young Researchers,* pp.13, December, 2005 (in Japanese).

15. Masayuki Shimada, Shigeki Miyabe, Tomoya Takatani, Hiroshi Saruwatari, and Kiyohiro Shikano, "Robust Sound Field Reproduction against User ' s Move Based on Multi-Channel Inverse Filtering with Selected and Enhanced Secondary Sources," *Fall Meeting of Acoust. Soc. Jpn.,* 3-Q-12, pp.671-672, September 2005 (in Japanese).

16. Yusuke Kaibara, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, and Yosuke Tatekura, "Multi-channel sound filed reproduction with adaptive inverse filter using temperature compensation and relaxation processing,"*Fall Meeting of Acoust. Soc. Jpn.,* 3-Q-13, pp.673-674, September 2005 (in Japanese).

17. Shigeki Miyabe, Tomoya Takatani, Yoshimitsu Mori, Kiyohiro Shikano, and Yosuke Tatekura, "Barge-in free spoken dialogue interface using sound field control and blind source seperation," *Fall Meeting of Acoust. Soc. Jpn.,* 3-Q-30, September 2005 (in Japanese).

18. Shigeki Miyabe, Tomoya Takatani, Hiroshi Saruwatari, and Kiyohiro Shikano, "Speech Recognition Using Barge-in Free Spoken Dialogue Interface Based on Response Sound Cancellation," *Spring Meeting of Acoust. Soc. Jpn.,* 3-Q-23, pp.551–552, March 2005 (in Japanese).

19. Shigeki Miyabe, Hiroshi Saruwatari, and Kiyohiro Shikano, "Small-Scale Barge-in Free Spoken Dialogue Interface Based on Response Sound Cancellation by Multichannel Sound Field Control," *Fall Meeting of Acoust. Soc. Jpn.,* 3-Q-16, pp.745–746, September 2004 (in Japanese).

Award

1. Winner of best noninear system in *MLSP 2007 Data Analysis Competition on Convolutive Blind Source Separation*.