BARGE-IN- AND NOISE-FREE SPOKEN DIALOGUE INTERFACE BASED ON SOUND FIELD CONTROL AND SEMI-BLIND SOURCE SEPARATION

Shigeki Miyabe[*†], *Tomoya Takatani*[†], *Hiroshi Saruwatari*[†], *Kiyohiro Shikano*[†], *and Yosuke Tatekura*[‡]

†Nara Institute of Science and Technology, 630-0192, Takayama-cho 8916-5, Ikoma-shi, Nara, Japan
Phone: +81-743-72-5287, email: {shige-m, tomoya-t, sawatari, shikano}@is.naist.jp, web: http://spalab.naist.jp/en/
‡Shizuoka University, 432-8561, Johoku 3-5-1, Hamamatsu, Shizuoka, Japan
Phone: +81-53-478-1139 email: tytatek@ipc.shizuoka.ac.jp web: http://spalab.eng.shizuoka.ac.jp/tatekura/index.html

ABSTRACT

For hands-free spoken dialogue system, we propose an interface to eliminate known noise outputted by the dialogue system and unknown noise sufficiently by giving known source of the loudspeaker output. Moreover, elimination of known noise is reinforced by sound field control. The proposed method, at first, eliminates the loudspeaker output at the microphone points using sound field control with fixed filter coefficients. Next, the observed signals are processed by a source separation to eliminate both known and unknown noise. The conventional approach to combine acoustic echo canceller and adaptive beamformer requires double-talk detection (DTD) in noisy environment, which is difficult to implement. Due to no use of DTD, the proposed method works better than the conventional method in real environment. To prove this, we show in the experiment that the performance of the proposed method is superior to the performance limit of the conventional method.

1. INTRODUCTION

To realize an interface for comfortable speech communication between human and machine based on spoken dialogue system, there is two important issues, i.e., hands free and barge-in free [1]. On one hand, hands free is a demand to free a user from spatial constraint to stand near or wear a microphone. On the other hand, barge-in free is the issue to set the user free from temporal forbiddance from inputting his speech command when the user is in outputting mode. These demands are difficult to be satidfied because automatic speech recognition is weak against noise. In this situation, noise is classified into two types, i.e., known and unknown noise. Unknown noise is interfering external noise in acoustic environment, caused by air conditioning or undesired speaker, etc. The most popular solution of unknown noise is beamforming using microphone array. Known noise is undesired observation of loudspeaker playback to present user message of sound from dialogue system. Elimination of such known response sound of dialogue system is called echo cancellation problem.

Most of the algorithms to adapt both adaptive beamformer (ABF) and acoustic echo canceller (AEC) are based on minimization of mean square error [2]. ABF learns filter to extract only desired signal by eliminating undesired signals [3]. For this purpose its filters must be learned only when the desired signal source is inactive. Thus ABF requires double-talk detector (DTD) to find such durations [4]. On the other hand, in principle, AEC does not require DTD because its filter is optimized by exploiting uncorrelation between the known noise and the other sound sources. However, in speech application where the transfer system is heavily variable and has long impulse response, DTD is indispensable to learn the filter rapidly with limited training samples. Although there are several approaches to simultaneous use of AEC and ABF [5], its implementation is hard because of difficulty in DTDs for this purpose. The problem of DTD is highly complicated because DTD for each of these methods has different target; noise for one is target to the other. In addition, noisy environment makes DTD more difficult.

To avoid DTD, one of the authors proposed an alternative method to eliminate response sound using robust sound field control with fixed filter without adaptation, called multiple-output and multiple-no-input (MOMNI) method [6]. This method utilizes sound field reproduction using multi-channel loudspeaker system and inverse filter of room transfer functions. The response sound is presented to user with high quality while its observation at microphones is prevented by construction of silent zone at the microphones. Although robustness of its fixed filters designed a-priori is proven in the literature, this method cannot be applied to reduction of unknown noise. To deal with unknown noise, adaptive signal processing is desired after observation by microphones. In addition, the robustness of the response sound elimination also can be reinforced by adding adaptive process.

As an adaptation of beamformer for unknown noise without DTD, blind source separation (BSS) based on independent component analysis (ICA) is studied [7, 8]. In the previous work we have proposed a semi-blind source separation (SBSS), which can eliminate known noise efficiently, and incorporated SBSS with the MOMNI method [**?**]. SBSS uses known source of response sound as training data together with observed signals at microphones. With these training data, known source is separated from other outputs efficiently. In this paper, we generalize the problem from only known-noise elimination into both known- and unknown-noise elmination. With this generalization, unknown sources can be separated in the same manner as ordinary BSS simultaneously to the elimination of known noise. In an experiment, we compare performances of the proposed method and ideal performance limit of conventional combination of AEC and ABF. As a result, although the performance of the conventional method degrades in practical use, the practical performance of the proposed method is superior to ideal behaviour of the conventional method.

2. MOMNI METHOD

The purpose of the MOMNI method is simultaneous realization of high-presence reproduction and elimination of response sound, which are conflicting issues. Since all information known a-priori is only transfer functions measured in advance, the elimination is not perfect because of fluctuation of room transfer functions. However, it is shown that this elimination is robust against the fluctuation without adaptation. We show configuration of the MOMNI method in Fig. 1.

This sound field reproduction controls sound pressures at $K+2$ control points, i.e., positions of the K microphones, denoted by C_k for $k = 1, \ldots, K$, and left and right ears of the user, denoted by C_{K+1} and C_{K+2}. The stereophonic response sound to be reproduced at the user's right and left ears are denoted by $r_{\mathrm{R}}(\omega)$ and $r_{\mathrm{L}}(\omega)$, where ω shows angular frequency. Our goal is that the sound pressures at the user's ears $d_{K+1}(\omega), d_{K+2}(\omega)$ equal the response sound signals while the sound pressures at the microphones, denoted by $d_k(\omega)$ for equal zero, as

$$\begin{aligned}\boldsymbol{d}(\omega) &= [d_1(\omega), \ldots, d_{K+2}(\omega)]^{\mathrm{T}} \\ &= [0, \ldots, 0, r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]^{\mathrm{T}}, \end{aligned} \quad (1)$$

where $\boldsymbol{d}(\omega)$ shows all the sound pressures at the control points and $\{\cdot\}^{\mathrm{T}}$ shows transposition. With this reproduction, the response sound is prevented from being observed by the microphones while the user listens to the high-quality reproduction of the response sound.

For the reproduction of Eq. (1), the effect of room transfer functions should be cancelled at the control points. such cancellation can be obtained by multi-channel inverse filter of the room transfer functions. To obtain the strict inverse filter of the room transfer functions with non-minimum phases, the number of the loudspeakers M for the reproduction must be larger than the number of the

[*]Research Fellow of the Japan Society for the Promotion of Science.

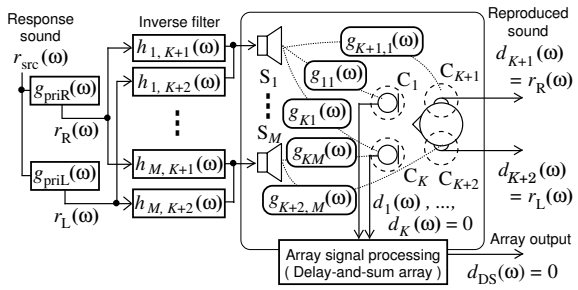

Figure 1: Configuration of the MOMNI method.

control points, i.e., $M > K+2$ [10]. First we measure all the transfer functions $g_{km}(\omega)$ for $k = 1,\ldots,K+2, m = 1,\ldots,M$. We compose an $M \times (K+2)$ transfer function matrix $\boldsymbol{G}(\omega) = [g_{km}(\omega)]_{km}$ where $[x]_{ij}$ denotes a matrix who has an entry x in the i-th row and j-th column. Next, the inverse filter matrix $\boldsymbol{H}(\omega) = [h_{mk}(\omega)]_{mk}$ is obtained by an Moore-Penrose generalized inverse matrix of $\boldsymbol{G}(\omega)$. Then the following condition is obtained;

$$\boldsymbol{G}(\omega)\boldsymbol{H}(\omega) = \boldsymbol{I}_{K+2}(\omega), \tag{2}$$

where I_n denotes an i-dimensional identity matrix. When $K+2$ signals are inputted to the inverse filter, each of them are reproduced at the control points. Using this property, the reproduction of the state in Eq.(1) can be obtained by inputting silent signals with zero amplitudes into the first K channels and the response sound into the remaining two channels as

$$\begin{aligned}\boldsymbol{d}(\omega) &= \boldsymbol{G}(\omega)\boldsymbol{H}(\omega)[0,\ldots,0,r_{\mathrm{R}}(\omega),r_{\mathrm{L}}(\omega)]^{\mathrm{T}} \\ &= [0,\ldots,0,r_{\mathrm{R}}(\omega),r_{\mathrm{L}}(\omega)]^{\mathrm{T}}.\end{aligned} \tag{3}$$

The accurate reproduction of stereophonic signals is effective especially when the input response sound signals are binaural recordings, refered to as transaural reproduction [11]. To generate binaural recordings, we filter a monaural response sound signal $r_{\mathrm{src}}(\omega)$ with head related transfer functions or binaural room impulse responses, denoted by $g_{\mathrm{priR}}(\omega), g_{\mathrm{priL}}(\omega)$, as

$$[d_{\mathrm{R}}(\omega),d_{\mathrm{L}}(\omega)]^{\mathrm{T}} = [g_{\mathrm{priR}}(\omega),g_{\mathrm{priL}}(\omega)]^{\mathrm{T}} r_{\mathrm{src}}(\omega). \tag{4}$$

Since the control points for the reproduction of any signals are only C_{K+1} and C_{K+2}, we truncate $\boldsymbol{H}(\omega)$ on the upper K rows and make an $M \times 2$ matrix $\boldsymbol{H}_2(\omega) = [h_{mk}(\omega)]_{mk}$ for $m = 1,\ldots,M$, $k = K+1, K+2$. Since the two rows of $\boldsymbol{H}_2(\omega)$ are in the null space of the transfer functions related to the microphones, when the repose sound signal $\boldsymbol{d}_{\mathrm{R}}(\omega)$, $\boldsymbol{d}_{\mathrm{L}}(\omega)$ are inputted to $\boldsymbol{H}(\omega)$, the following condition is obtained;

$$\begin{aligned}\boldsymbol{d}(\omega) &= \boldsymbol{G}(\omega)\boldsymbol{H}_2(\omega)[d_{\mathrm{R}}(\omega),d_{\mathrm{L}}(\omega)]^{\mathrm{T}} \\ &= \boldsymbol{G}(\omega)\boldsymbol{H}_2(\omega)[g_{\mathrm{priR}}(\omega),g_{\mathrm{priL}}(\omega)]^{\mathrm{T}} r_{\mathrm{src}}(\omega) \\ &= [0,\ldots,0,r_{\mathrm{R}}(\omega),r_{\mathrm{L}}(\omega)],\end{aligned} \tag{5}$$

which is equivalent to Eq. (1).

Note that alghough high-quality reproduction cannot be obtained when the user does not sit on the arranged position, it is shown in [6] that the degradation is not in problematic level for spoken dialogue system.

3. INTRODUCING ICA TO MOMNI METHOD

3.1 Motivation

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

(R1) As shown in [6], robustness of the MOMNI method is improved according to the number of loudspeakers. To reduce the expense of the loudspearkers, adaptation of the elimination is required.

(R2) For a hands-free system, elimination of interfering noise is an important issue. Thus adaptive signal processing method for the noise reduction 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 by the microphones.

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

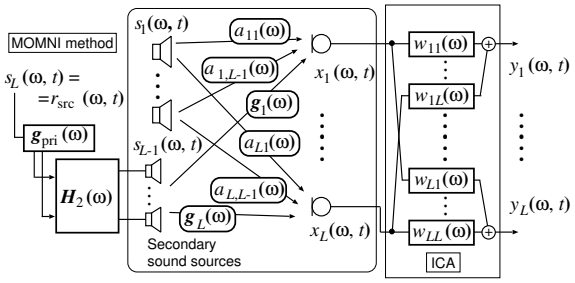

Figure 2: Configuration of BSS based on FD-ICA.

DTD and inappropriate for our purpose. 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 [12], we try to adopt FD-ICA in the MOMNI framework.

3.2 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. 2. BSS is a problem to estimate unknown source signals only from the observed signals, which are linear mixture of sources in unknown system. Suppose that there are L unknown sound sources $\boldsymbol{S}(\omega) = [s_1(\omega),\ldots,s_L(\omega)]^{\mathrm{T}}$. Using K microphones, their observed signals,

$$\boldsymbol{x}(\omega) = [x_1,\ldots,x_K(\omega)]^{\mathrm{T}}, \tag{6}$$

can be written as

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

where $\boldsymbol{A}(\omega) = [a_{kl}(\omega)]_{kl}$ is an unknown $K \times L$ transfer function matrix. To obtain L separated source signals, $K \geq L$ must be satisfied. The purpose of BSS is to obtain an $L \times K$ separation filter matrix $\boldsymbol{W}(\omega)$ which makes its output signals,

$$\boldsymbol{y}(\omega) = [y_1(\omega),\ldots,y_L(\omega)]^{\mathrm{T}} = \boldsymbol{W}\boldsymbol{x}(\omega), \tag{8}$$

be the estimation of the separated sources.

In FD-ICA, first, short-time analysis of the observed signals is conducted by frame-by-frame discrete Fourier transform. By plotting the spectral values in a frequency bin for each microphone input frame by frame, we consider them as a time series. Hereafter, we designate the spectral values as $\boldsymbol{x}(\omega,t) = [x_1(\omega,t),\ldots,x_K(\omega,t)]^{\mathrm{T}}$, where t denotes the time index of the frame. Next, we obtain the separation filter $\boldsymbol{W}(\omega)$ whose time-series output $\boldsymbol{y}(\omega,t)$,

$$\boldsymbol{y}(\omega) = [y_1(\omega,t),\ldots,y_L(\omega,t)]^{\mathrm{T}} = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega,t), \tag{9}$$

are statistically independent. 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 $\boldsymbol{W}(\omega)$ is optimized by, for example, the following iterative updating operation [7]:

$$\boldsymbol{W}^{++}(\omega) = \boldsymbol{W}(\omega) - \eta\{\boldsymbol{I} - \langle\boldsymbol{\Phi}(\boldsymbol{y}(\omega,t))\boldsymbol{y}^{\mathrm{H}}(\omega,t)\rangle_t\}\boldsymbol{W}(\omega), \tag{10}$$

where $\langle\cdot\rangle_t$ denotes the time-averaging operator, $\{\cdot\}^{\mathrm{H}}$ denotes the conjugate transposition, and $\boldsymbol{W}^{++}(\omega)$ is an updated filter matrix. In our research, we use tangent hyperbolic function based on polar coordinate [13] as;

$$\boldsymbol{\Phi}(\boldsymbol{y}(\omega)) = \begin{bmatrix}\tanh(|y_1(\omega)|)\exp(j\arg(y_1(\omega))) \\ \vdots \\ \tanh(|y_K(\omega)|)\exp(j\arg(y_L(\omega)))\end{bmatrix}. \tag{11}$$

The separation filter $\boldsymbol{W}(\omega)$ requires some modifications for usage. First, the condition of independence has ambiguity in scaling of the output signals, both for amplitudes and phases. To compensate for this, we apply projection back [14] to estimate the source signals at the microphone points using inverse of the separation filter. Second, 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 [15] and use of envelopes' correlations among narrow-band signals [14]. In our research we use combination of those approaches [16].

3.3 Straightforward Combination of BSS and MOMNI

The purpose of introducing BSS in the MOMNI framework is to separate user's utterance from the observed user's utterance mixed with unknown noise and the known response sound. Although source of the response sound is known to the system, its observation is mixed with unknown signals and effected by unknown fluctuation of the room transfer functions. The problem is the same as the ordinary BSS except for the availability of the response sound source. Thus, the most straightforward idea to extract the user's utterance is simply to use BSS, treating the response sound as an unknown signal. In this section, we analyze the mechanism of the separation in this strategy, and point out several problems.

Although the residual response sound observed at the microphones is influenced by the multiple transfer channels, it is originally a single source. Thus ICA can separate the response sound as one of the estimated sources. We define an M-dimensional row vector $\boldsymbol{g}_k(\omega)$ $(k = 1,\ldots,K)$ composed of measured room transfer functions $g_{km}(\omega)$ $(m = 1,\ldots,M)$ between the k-th microphone element and all the M loudspeakers before fluctuation. Then we define $\boldsymbol{g}'_k(\omega)$, the unknown room transfer functions fluctuated after the design of the inverse filter, as

$$\boldsymbol{g}'_k(\omega) = \boldsymbol{g}_k(\omega) + \Delta\boldsymbol{g}_k(\omega), \tag{12}$$

where $\Delta\boldsymbol{g}_k(\omega)$ is the differential of $\boldsymbol{g}_k(\omega)$ and $\boldsymbol{g}'_k(\omega)$. If input signals are given by Eq. (4), because of the condition $\boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega) = 0$ the residual response sound $d'_k(\omega)$ observed at the k-th microphone element can be written as

$$\begin{aligned} d'_k(\omega) &= (\boldsymbol{g}_k(\omega) + \Delta\boldsymbol{g}_k(\omega))\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega)r_{\mathrm{src}}(\omega) \\ &= \Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{src}}(\omega)r_{\mathrm{src}}(\omega). \end{aligned} \tag{13}$$

Thus, the residual sound can be written as a multiplication of single source $r_{\mathrm{src}}(\omega)$ and a single scalar transfer function $\Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega)$. Suppose there are $L-1$ mutually independent source $s_l(\omega)$ $(l = 1,\ldots,L-1)$ in the room, excluding the response sound; they should be independent of of the response sound. Then, the observed signal at the k-th microphone element can be written as

$$x_k(\omega) = \sum_{l=1}^{L-1} a_{kl}(\omega)s_l(\omega) + \Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega)r_{\mathrm{src}}(\omega) \tag{14}$$

where there exist L independent signals including the component of the response sound. By the substitutions

$$a_{kL}(\omega) = \Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega) \quad \text{for} \quad k = 1,\ldots,K, \tag{15}$$

and

$$s_L(\omega) = r_{\mathrm{src}}(\omega), \tag{16}$$

in Eq. (7), the mixing system can be described in the same manner as ordinary BSS. Since there are L sources, ICA can separate the signals with L observed signals. Thus, the MOMNI method should make silent zones at $K = L$ microphone elements with the sound field control, and then we input the observed signals of the microphone elements to ICA.

However, this method has several problems. The first one is that its output signals are distorted. The mechanism of separation by ICA is multiple beamformers which extract independent sources separately [17]. In general, to construct beamformer with high performance, required filter length is longer than those of the room transfer functions. In addition, 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 $\Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega)$ 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 method will perform as well as the ordinary BSS.

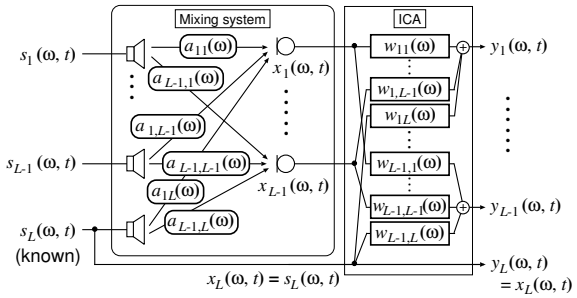

Figure 3: Configuration of semi-blind source separation.

3.4 Semi-Blind 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 semi-blind 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.

Suppose there are L sources $s_l(\omega)$ $(l = 1,\ldots,L)$ and only the L-th source $s_L(\omega)$ is the known source. To separate L sources with ICA, L mixed signals are required. However in this case, we use $s_L(\omega)$ as one of the input signals. These input signals can be expressed by the substitution

$$x_L(\omega) = s_L(\omega) \tag{17}$$

in Eq. (6). In addition, the mixing system of these input signals also can be expressed by the substitution

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

in Eq. (7). Since the L-th input signal $s_L(\omega)$ is already separated, it should be outputted without any modification, i.e.,

$$y_L(\omega) = x_L(\omega) = s_L(\omega). \tag{18}$$

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

$$w_{Ll}(\omega) = \begin{cases} 0 & \text{for} \quad l = 1,\ldots,L-1, \\ 1 & \text{if} \quad l = L. \end{cases} \tag{19}$$

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

$$\begin{aligned} \bar{\boldsymbol{W}}(\omega) &= [w_{lk}(\omega)]_{lk} \quad \text{for} \quad l = 1,\ldots,L-1,\ k = 1,\ldots,L \\ &= [\boldsymbol{I}_{L-1}, \boldsymbol{0}_{L-1}]\,\boldsymbol{W}(\omega), \end{aligned} \tag{20}$$

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

$$\bar{\boldsymbol{W}}^{++} = \bar{\boldsymbol{W}}(\omega) + \eta\left\{\bar{\boldsymbol{W}}(\omega) - \langle\boldsymbol{\Phi}(\bar{\boldsymbol{y}}(\omega))\boldsymbol{y}^{\mathrm{H}}(\omega)\rangle_t\boldsymbol{W}(\omega)\right\} \tag{21}$$

where

$$\bar{\boldsymbol{y}}(\omega) = [y_1(\omega),\ldots,y_{L-1}(\omega)]^{\mathrm{T}}. \tag{22}$$

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 $s_L(\omega)$. 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 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.

3.5 Combination of MOMNI Method and SBSS

Combination of the MOMNI method and SBSS can be realized by just giving the response sound source to ICA as $s_L(\omega) = r_{\mathrm{src}}(\omega)$ in the control of $L-1$ microphones to be silent. In this combination, the advantage of SBSS is significant. As discussed in Sect. 3.3, the long impulse response of $\Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}_2(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega)$ requires BSS to have extremely long filter coefficients. However, as discussed

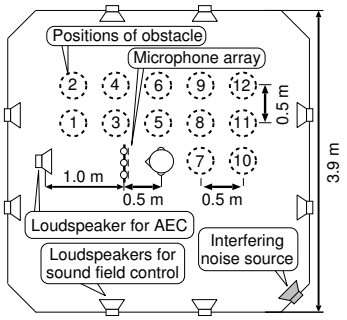

Figure 4: Layout of acoustic environment room.

in the previous section, the filter length can be shorten by SBSS. 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, as also discussed in the previous section.

4. SIMULATION

4.1 Experimental Conditions

To validate the performance of the proposed method to eliminate both response sound and interfering noise, we conducted simulation of speech enhancement using impulse responses measured in real environment. The competitive methods against proposed combination of MOMNI method and SBSS (**proposed**) are combination of ABF and AEC assuming ideal DTDs (**AEC+ABF**), whose details are described in the following section, conventional BSS (**BSS**), SBSS without sound field control of MOMNI method (**SBSS**), conventional MOMNI method with delay-and-sum (DS) beamformer [6] (**MOMNI+DS**), and combination of MOMNI method and BSS but SBSS (**MOMNI+BSS**).

Figure 4 shows the arrangement of the apparatuses in the experimental room. The reverberation time is approximately 160 ms. The sampling frequency and the resolution are 16 kHz and 16 bit, respectively. We used eight loudspeakers for the sound field control and they are positioned along the outer circumstance of the room. The primary source of the 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 AEC+ABF, BSS and SBSS. We place a dummy head, i.e., a replica of an average human head and torso, at the user's position.

When the room transfer functions do not alter from the state where the inverse filter was designed, the performance of the MOMNI method is infinity. 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. Supposing that another person than the user is moving about in the room, we used a life-size mannequin as the obstacle. Note that we did not change the position of the dummy head to fix the distance to the microphones. 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. We evaluated the average of 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 Japanese sentences by 13 male and 13 female speakers. The performances are also averaged by these 200 utterances. The number of microphone elements are three and two for MOMNI+DS and the others, respectively.

The power of the of the user's speech and the response sound are arranged to be the same. The power of the interfering noise is arranged to be 10 dB lower than the user's speech. The source of the interfering noise is set at the edge of the room. As interfering noise, we used three signals, i.e., a female utterance, music (a symphony), and stationary noise with -10 dB/octave spectral coloration.

4.2 Adaptation of AEC and ABF Using Ideal DTDs

In this section we describe adaptation algorithm of AEC+ABF. To validate the performance limit of combination of AEC and ABF, we

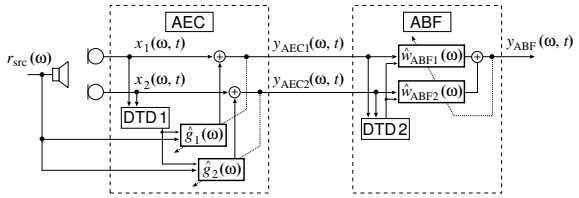

Figure 5: Configuration of the adaptation of the AECs and ABF using ideal frequency-domain DTDs.

estimated the filter coefficients by frequency-domain batch adaptation with ideal DTDs. To evaluate ideal behaviours of DTDs, we manually gave the true durations when the power ratio of the signal to eliminate to the other signals exceeds a threshold. Since human speech is known to be sparse in time-frequency domain, there are many time-frequency grids where the signal due to human speech is much smaller or larger than the other signals. Thus, if we can find such grids where a source to be adapted is dominant, the filters can be adapted in frequency domain even if there is no full-band single-talk duration. The adaptations are not on-line but batch-wise. With this batch adaptation, the adaptation of AEC+ABF can be set on the equal footing with the batch learning of the ICAs. Also, since the system in this experiment is static after fluctuation once occurred, batch adaptation estimates the performance limit of many on-line adaptation algorithms because batch adaptation consistently outperform on-line adaptation algorithms in static system.

Figure 5 shows the configuration of AEC+ABF. We use two microphone elements. At first, two-channelled AECs eliminates the response sound from the observed signals using DTD1 for the AECs. At second, the processed signals of the AECs are inputted to DTD2 for the ABF. Then, the ABF processes the outputs of the AECs to eliminate the interfering noise.

Here we describe the adaptation of the AECs. The observed signals are denoted as $x_1(\omega,t)$ and $x_2(\omega,t)$. DTD1 detects the times $t \in \mathscr{T}_1(\omega)$ when the power of the signals due to the response sound are much larger than the others. Then the filters $\hat{g}_1(\omega), \hat{g}_2(\omega)$ conducts the echo cancellation as

$$y_{\mathrm{AEC}k}(\omega,t) = x_k(\omega,t) + \hat{g}_k(\omega,t) r_{\mathrm{src}}(\omega,t) \quad \text{for} \quad k = 1,2, \tag{23}$$

where $y_{\mathrm{AEC}k}(\omega,t)$ are processed signals of the AECs. The residual echo in the detected single-talk durations can be written as

$$\varepsilon_k(\omega) = E\left[|y_{\mathrm{AEC}k}(\omega,t)|^2\right]_{t\in\mathscr{T}_1(\omega)}, \tag{24}$$

where $E[\cdot]$ denotes expectation. The optimal solution of $\hat{g}_k(\omega)$ to minimize $\varepsilon_k(\omega)$ satisfies

$$\frac{\partial \varepsilon_k(\omega)}{\partial \hat{g}_k^*(\omega)} = \frac{\partial E\left[|x_k(\omega,t) + \hat{g}_k(\omega,t) r_{\mathrm{src}}(\omega,t)|^2\right]_{t\in\mathscr{T}_1(\omega)}}{\partial \hat{g}_k^*(\omega)} = 0. \tag{25}$$

Substituting expectation by time average, the optimal solution is obtained as

$$\hat{g}_k(\omega) = -\frac{\langle x_k(\omega,t) r_{\mathrm{src}}^*(\omega,t)\rangle_{t\in\mathscr{T}_1(\omega)}}{\langle |r_{\mathrm{src}}(\omega,t)|^2\rangle_{t\in\mathscr{T}_1(\omega)}}. \tag{26}$$

Subsequently, DTD2 detects the times $t \in \mathscr{T}_2(\omega)$ when the power of the signals due to the interfering noise are much larger than the others in $y_{\mathrm{AEC1}}(\omega,t)$ and $y_{\mathrm{AEC2}}(\omega,t)$. As an adaptation method of the ABF, we adopted linear constrained minimum variance beamformer [3].

Figure 6 shows the relation between the rates of detected single-talk grids and the thresholds for DTD1 and DTD2. Simultaneous use of AEC and ABF is difficult because different DTDs are required for each of them. The difficulty can also be seen in the trade-off between the quantity of single-talk grids and the threshold which leads quality of the signals for adaptation. Among the interfering noise signals, the female utterance is the most sparse and the stationary noise is the most dense. Sparseness increases the single-talk grids in DTD1 but decreases those in DTD2. Thus the appropriate threshold varies according to the property of the signals. We adopted the threshold of 15 dB with which the best performance was obtained.

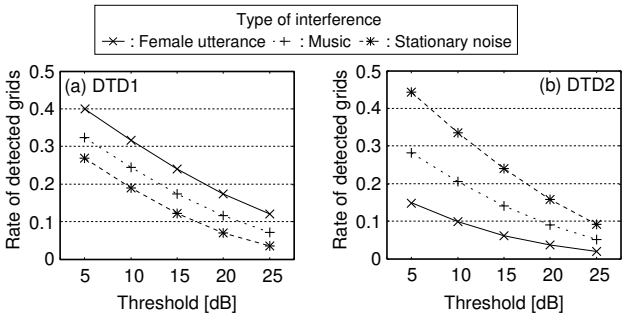

Figure 6: Rates of detected single-talk grids with various threshold.

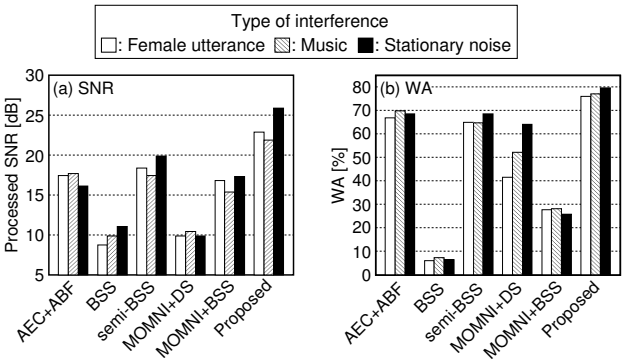

Figure 7: Experimental results.

4.3 Results and Discussions

We show signal-to-noise ratios (SNRs) of the processed signals and their speech recognition performance (word accuracy; WA [18]) in large vocabulary Japanese dictation. At first we compare the SNRs. Because of the shortage of the single-talk grids detected by the DTDs, the performance of AEC+ABF falls below 20 dB. Note that such precise DTDs cannot be implemented in practice. Because of the layout where the response sound source stand in line with the microphone array and the user, BSS cannot eliminate the response sound sufficiently. In contrast, SBSS shows the similarly high scores to those of AEC+ABF without DTDs. The performance of MOMNI+DS is not so high because DS cannot eliminate interfering noise sufficiently. MOMNI+BSS shows higher score than than MOMNI+DS, but lower than AEC+ABF and SBSS. Above all, with successful elimination of the interfering noise and the residual response sound, proposed method shows the highest performance.

As for the WAs, the scores are almost proportional to the SNRs except for the low score of MOMNI+BSS because of the distortion discussed in Sect. 3.4. The proposed method successfully eliminates the residual response sound and the interfering noise with high accuracy and low distortion, and shows the highest performance. From these results, the efficacy of the proposed method is ascertained.

5. CONCLUSION

We have proposed semi-blind source separation algorithm to separate mixture of known and unknown signals efficiently. Then we have incorporated the source separation with the spoken dialogue interface using sound field control. It is shown in the experiment that the performance of the proposed method is higher than the performance limit of the conventional combination of AEC and ABF because of difficulty in DTD. From these findings, efficacy of the proposed method is ascertained.

A. DEVIATION OF UPDATE FORMULA

In this appendix, we derive the updating formula (22) by minimizing Kullback-Leibler divergence $I_{\mathrm{KL}}(\omega)$ between the joint probability distribution $p(\boldsymbol{y}(\omega,t))$ and the product of marginal probability distributions $\prod_{l=1}^{L} p(y_l(\omega,t))$, described as

$$I_{\mathrm{KL}}(\omega) = \int p(\boldsymbol{y}(\omega,t)) \log \frac{p(\boldsymbol{y}(\omega,t))}{\prod_{l=1}^{L} p(y_l(\omega,t))} d\boldsymbol{y}(\omega,t). \qquad (27)$$

Partial differential of $I_{\mathrm{KL}}(\omega)$ by $\boldsymbol{W}(\omega)$ [7, 8] can be written as

$$\frac{\partial I_{\mathrm{KL}}(\omega)}{\partial \bar{\boldsymbol{W}}(\omega)} = -\boldsymbol{W}^{-\mathrm{H}}(\omega) + E\left[\boldsymbol{\Phi}(y(\omega,t))y^{\mathrm{H}}(\omega,t)\right]_t, \qquad (28)$$

where $\{\cdot\}^{-\mathrm{H}}$ denotes inverse of conjugate transpose and $E[\cdot]_t$ denotes expectation operator with respect to t. Similarly to Eq. (20), partial differential with respect to $\bar{\boldsymbol{W}}(\omega,t)$ can be given by

$$\begin{aligned}\frac{\partial I_{\mathrm{KL}}(\omega)}{\partial \bar{\boldsymbol{W}}(\omega)} &= -[\boldsymbol{I}_{L-1},\ \boldsymbol{0}]\frac{\partial I_{\mathrm{KL}}(\omega)}{\partial \boldsymbol{W}(\omega)} \\ &= -[\boldsymbol{I}_{L-1},\ \boldsymbol{0}]\boldsymbol{W}^{-\mathrm{H}}(\omega) + E\left[\boldsymbol{\Phi}(\bar{\boldsymbol{y}}(\omega,t))\boldsymbol{x}^{\mathrm{H}}(\omega,t)\right]_t. \qquad (29)\end{aligned}$$

By applying the natural gradient of $\boldsymbol{W}(\omega)$ [7], the update of $\bar{\boldsymbol{W}}(\omega)$ can be obtained as

$$\begin{aligned}\bar{\boldsymbol{W}}^{++}(\omega) &= \bar{\boldsymbol{W}}(\omega) - \eta\frac{\partial I_{\mathrm{KL}}(\omega)}{\partial \bar{\boldsymbol{W}}(\omega)}\boldsymbol{W}^{\mathrm{H}}(\omega)\boldsymbol{W}(\omega) \\ &= \bar{\boldsymbol{W}}(\omega) + \eta\left\{\bar{\boldsymbol{W}}(\omega) - E\left[\boldsymbol{\Phi}(\bar{\boldsymbol{y}}(\omega,t))\boldsymbol{y}^{\mathrm{H}}(\omega,t)\right]_t\boldsymbol{W}(\omega)\right\}. \qquad (30)\end{aligned}$$

Assuming the ergodicity of the sources, the expectation can be substituted by the time average and the update formula (22) is obtained.

REFERENCES

[1] B. H. Juang and F. K. Soong, "Hands-free telecommunications," *Proc. Int. Workshop on Hands-Free Speech Communication*, pp. 5–10, 2001.

[2] B. Widrow, "Adaptive Noise Cancelling: Principles and Applications," *Proc. IEEE,* vol.63, pp.1692–1716, 1975.

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

[4] 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.

[5] W. Herbordt *et al.,*"An acoustic human-machine front-end for multimedia applications," *EURASIP Journal on Advances in Signal Processing,* vol. 2003, no. 1, pp. 1–11, 2003.

[6] S. Miyabe *et al.,* "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, 13 pages, 2007.

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

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

[9] S. Miyabe *et al.,* "Double-talk free spoken dialogue interface combining sound field control with semi-blind source separation," in *Proc. Intl. Conf. Acoust., Speech, Signal Process.,* vol. 1, pp. 809–812, 2006.

[10] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 36, no. 2, pp. 145–152, 1988.

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

[12] T. Nishikawa *et al.,* "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.

[13] H. Sawada *et al.,* "Polar coordinate based on nonlinear function for frequency domain blind source separation," *IEICE Trans. Fundamentals*, vol. E86-A, no. 3, pp. 590–596, 2003.

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

[15] H. Saruwatari *et al.,* "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing,* vol. 2003, No. 11, pp. 1135–1146, 2003.

[16] H. Sawada *et al.,* "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech & Audio Process,*, vol. 12, no. 5, pp. 530–538, 2004.

[17] S. Araki *et al.,* "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.

[18] A. Lee *et al.,* "Julius — an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH2001*, vol. 3, pp. 1691–1694, 2001.