DOUBLE-TALK FREE SPOKEN DIALOGUE INTERFACE COMBINING SOUND FIELD CONTROL WITH SEMI-BLIND SOURCE SEPARATION

Shigeki Miyabe†*, Tomoya Takatani*†*, Yoshimitsu Mori*†*,*
Hiroshi Saruwatari†*, Kiyohiro Shikano*†*, Yosuke Tatekura*‡

† Nara Institute of Science and Technology
{shige-m, tomoya-t, yoshim-m, sawatari, shikano}@is.naist.jp
‡ Shizuoka University
tytatek@ipc.shizuoka.ac.jp

ABSTRACT

In this paper we introduce a new double-talk free spoken dialogue interface combining sound field control and a source separation technique based on independent component analysis (ICA). First, sound field control provides silent zones on the microphone elements and prevents the response sound from being observed. In the second step, we propose a novel semi-blind source separation algorithm to suppress the error caused by fluctuation of the room transfer function. By using a direct input of response sound signal to ICA, a source separation problem can be converted to a supervised learning problem. Since the problem becomes easier, the proposed method showed higher performances than the method using blind source separation.

1. INTRODUCTION

In human-machine communication based on a spoken dialogue system, it is desirable that a user can input his speech without wearing special equipment. In addition, the system should be ready for receiving the user's speech input anytime to set the user free from waiting, even in the moment when the system outputs a message to the user by sound (response sound). However, in such a situation when the user and the system utter simultaneously, the user's speech utterance is observed mixed with the response sound and its speech recognition performance degrades. For a successful realization of the hands-free spoken dialogue system, a mechanism to eliminate the response sound is necessary.

To eliminate the response sound from the system, an acoustic echo canceller (AEC) is commonly used. Many types of AECs have been proposed, e.g., single channel, stereophonic, wave synthesis, and beamformer-integrated types [1, 2, 3]. However, the AEC has an inherent problem in which an accurate adaptation is difficult in the duration when both the user and the system utter simultaneously (double-talk). Because of this problem, the conventional AEC should adapt filter coefficients when only the system utters, and detect the double-talk duration and stop adaptation; this implies that the elimination performance is likely to degrade when a change in room transfer function arises during double-talk.

To solve the problem of the AEC, one of the authors has proposed Multiple-Output and Multiple-No-Input (MOMNI) method [4], which combines sound field control and beamforming. The MOMNI method controls the sound field around microphones to be silent and prevents the response sound from being observed. In the second step, the observed signals are applied to delay-and-sum array signal processing to improve the robustness of the elimination of the response sound. The elimination performance of the MOMNI method can be improved by increasing the numbers of loudspeakers and microphone elements.

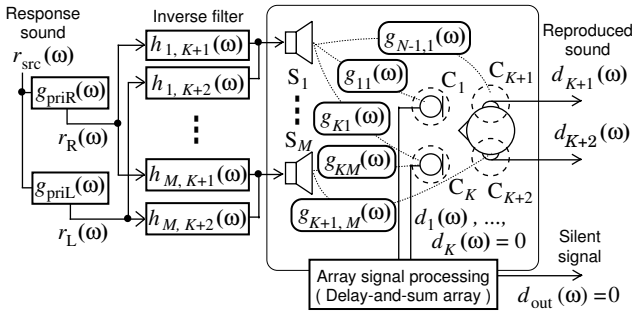

Fig. 1. Configuration of conventional MOMNI method.

Though the control of the MOMNI method is highly robust against a fluctuation of the transfer function, there is still room for improvement in its performance. By applying an adaptive process to update the filter coefficients of microphone array, the MOMNI method obtains an adaptation faculty to fluctuation of the room transfer functions. In many types of array signal processing, one of the most powerful candidate is blind source separation (BSS) based on independent component analysis (ICA) because BSS doesn't require double-talk detection. In this paper we extend BSS and propose semi-blind source separation which is a supervised learning. By giving ICA a direct input of the response sound signal as an answer, we make other output signals statistically independent of the response sound, or in other words, only the response sound is eliminated.

2. CONVENTIONAL MOMNI METHOD

2.1. Algorithm

The configuration of the MOMNI method is shown in Fig. 1. We set M loudspeakers S_m, $(m = 1, \ldots, M)$ and $K + 2$ control points C_k $(k = 1, \ldots, K + 2)$ to satisfy the condition $M > K + 2$. K control points C_k $(k = 1, \ldots, K)$ are set on the microphone elements to observe the response sound, and C_{K+1} and C_{K+2} are set on the user's right and left ears. The vector $\boldsymbol{r}(\omega) = [r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]^{\mathrm{T}}$ where $\{\}^{\mathrm{T}}$ describes transposition and ω shows angular frequency, is a set of signals intended to be reproduced at the control points C_{K+1} and C_{K+2}, and the vector

$$\boldsymbol{d}(\omega) = [d_1(\omega), \ldots, d_K(\omega), d_{K+1}(\omega), d_{K+2}(\omega)]^{\mathrm{T}} \qquad (1)$$

is a set of signals at the control points. The room transfer functions between the loudspeakers S_m $(m = 1, \ldots, M)$ and the control

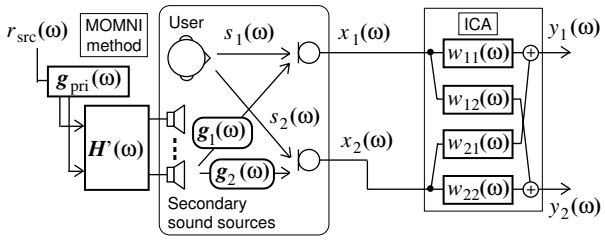

Fig. 2. Configuration of the simple connection of BSS with MOMNI method.

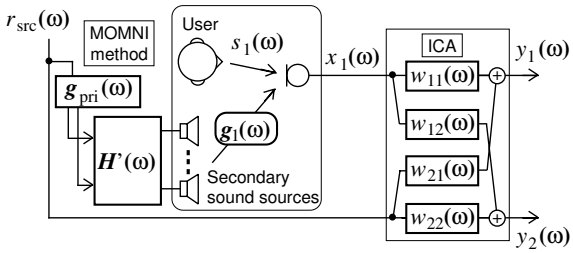

Fig. 3. Configuration of the proposed method.

points $\mathrm{C}_k(\omega)$ $(k = 1, \ldots, K+2)$ are described by $M \times (K+2)$ matrix $\boldsymbol{G}(\omega)$ whose entries are the room transfer functions $g_{km}(\omega)$. To reproduce the input signals $\boldsymbol{r}(\omega)$ on the control points $\mathrm{C}_k(\omega)$, we design an $M \times (K+2)$ inverse filter matrix $\boldsymbol{H}(\omega)$ by calculating Moore-Penrose generalized inverse matrix of $\boldsymbol{G}(\omega)$ composed of h_{mk} $(m = 1, \ldots, M, k = 1, \ldots, K+2)$. Then, we truncate the matrix $\boldsymbol{H}(\omega)$ into $\boldsymbol{H}'(\omega)$ which is an $M \times 2$ filter matrix composed of the filter components $h_{mk'}(\omega)$ $(m = 1, \ldots, M, k' = K+1, K+2)$ of $\boldsymbol{H}(\omega)$. With this filter matrix, the following equation holds;

$$\boldsymbol{d}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{H}'(\omega)\boldsymbol{r}(\omega) = [\underbrace{0, \ldots, 0}_{K}, r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]^{\mathrm{T}}. \quad (2)$$

Therefore, on one hand, the response sound signals equal the signals at the user's ears ($[d_{K+1}(\omega), d_{K+2}(\omega)] = [r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]$) and reproduced strictly. On the other hand, silent zones are realized at microphone elements ($d_k(\omega) = 0$ for $k = 1, \ldots, K$) and the response sound is prevented from being observed at the microphone elements. Then, delay-and-sum array signal processing is applied to the observed signals.

Since the MOMNI method uses an inverse filter of the room transfer function, three dimensional sound field reproduction can be presented. To make full use of this property, we make the response sound signals $(r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega))$ by multiplying the room transfer functions $\boldsymbol{g}_{\mathrm{pri}}(\omega) = [g_{\mathrm{priR}}(\omega), r_{\mathrm{priL}}(\omega)]^{\mathrm{T}}$ between a primary sound source and both of the user's ears, and a monaural source of the response sound signal $r_{\mathrm{src}}(\omega)$ as

$$[r_{\mathrm{R}}(\omega), r_{\mathrm{L}}(\omega)]^{\mathrm{T}} = \boldsymbol{g}_{\mathrm{pri}}(\omega) r_{\mathrm{src}}(\omega). \quad (3)$$

This mechanism can present the source position of an agent of dialogue system with high precision.

2.2. Response Sound Elimination Error When Changing Room Transfer Functions

The MOMNI method can make its control robust against fluctuation of the room transfer functions. Assume that the number of loudspeakers M is enough larger than the number of control points, and the condition number of the inverse filter matrix approaches to 1. Then, it is proved that the elimination error after fluctuation of room transfer function is in proportion to $1/\sqrt{MK}$ [4]. Therefore, the robustness of the MOMNI method against the room transfer functions is improved by increasing the number of the loudspeakers and the microphone elements.

3. INTRODUCING INDEPENDENT COMPONENT ANALYSIS TO MOMNI METHOD

In this section we propose an algorithm which apply ICA after the sound field control of the MOMNI method. The conventional MOMNI method adopts delay-and-sum array signal processing with fixed filter coefficients. If some adaptive array signal processing is applied, the MOMNI method can obtain not only improvement of robustness against room transfer function but also environmental noise or another talker. Though most of adaptive array signal processings require information of single-talk duration, BSS based on independent component analysis can learn its filter coefficients only from observed signals. In this paper we assume that there is no additional noise and discuss only elimination of the mixture of the response sound in the observed signal caused by fluctuation of the room transfer function. However, in case there is some additional noise, by increasing the number of microphone elements and size of the filter matrix of ICA, the proposed method obtains ability to separate the user's speech from the additional noise.

3.1. Simple Connection of BSS with MOMNI Method

The most simple idea is just to connect BSS with the MOMNI method as shown in Fig. 2. We define an M-dimensional vector $\boldsymbol{g}_k(\omega)$ $(k = 1, \ldots, K)$ composed of 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 room transfer function after fluctuation given by

$$\boldsymbol{g}'_k(\omega) = \boldsymbol{g}_k(\omega) + \Delta\boldsymbol{g}_k(\omega), \quad (4)$$

where $\Delta\boldsymbol{g}_k(\omega)$ is a differential of $\boldsymbol{g}_k(\omega)$ and $\boldsymbol{g}'_k(\omega)$. If input signals are given by (3), $\boldsymbol{g}_k(\omega)\boldsymbol{H}'(\omega) = 0$ and observed signal x_k at k-th microphone element is given by

$$\begin{aligned} x_k(\omega) &= \boldsymbol{g}'_k(\omega)\boldsymbol{H}'(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega) r_{\mathrm{src}}(\omega) + s_k(\omega) \\ &= \Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}'(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega) r_{\mathrm{src}}(\omega) + s_k(\omega), \end{aligned} \quad (5)$$

where $s_k(\omega)$ is a component of the user's utterance observed at the k-th microphone element. Equation (5) shows that the number of independent signals included in $x_k(\omega)$ is two and separation can be achieved by using two observed signals. Therefore this method uses two microphone elements ($K = 2$) and inputs observed signals of these microphone elements to frequency-domain ICA (FD-ICA). We define 2×2 separation filter matrix $\boldsymbol{W}(\omega)$ as

$$\boldsymbol{y}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega) = \begin{bmatrix} w_{11}(\omega) & w_{12}(\omega) \\ w_{21}(\omega) & w_{22}(\omega) \end{bmatrix} \boldsymbol{x}(\omega), \quad (6)$$

where two dimensional column vector $\boldsymbol{y}(\omega) = [y_1(\omega), y_2(\omega)]^{\mathrm{T}}$ describes output signals. FD-ICA updates its filter $\boldsymbol{W}(\omega)$ to make its output signals statistically independent. The update of filter coefficients are given by

$$\boldsymbol{W}_{++}(\omega) = \boldsymbol{W}(\omega) - \eta \left\{ \boldsymbol{I} - \langle \boldsymbol{\Phi}\left(\boldsymbol{y}(\omega, t)\right) \boldsymbol{y}^{\mathrm{H}}(\omega, t) \rangle_t \right\} \boldsymbol{W}(\omega), \quad (7)$$

where $\boldsymbol{W}_{++}(\omega)$ is the updated filter, $\boldsymbol{y}(\omega, t)$ is $\boldsymbol{y}(\omega)$ observed at time t, $\langle \cdot \rangle_t$ is a time average operator, η is a step-size parameter, $\boldsymbol{\Phi}$

is an activation function like polar function [5] given by

$$\boldsymbol{\Phi}(\boldsymbol{y}(\omega)) = \begin{bmatrix} \tanh(|y_1(\omega)|)\exp\left(j\arg\left(y_1\left(\omega\right)\right)\right) \\ \tanh(|y_2(\omega)|)\exp\left(j\arg\left(y_2\left(\omega\right)\right)\right) \end{bmatrix}. \quad (8)$$

Since the gain of each frequency has arbitrariness in FD-ICA, its output signals are distorted. To compensate for this, projection back [6] is applied. In this case, the output signals $\boldsymbol{p}(\omega) = [p_1(\omega), p_2(\omega)]^{\mathrm{T}}$ processed by projection back can be written as

$$\boldsymbol{p}(\omega) = \mathrm{diag}\left(\boldsymbol{W}^{-1}(\omega)\begin{bmatrix} y_1(\omega) & 0 \\ 0 & y_2(\omega) \end{bmatrix}\right), \quad (9)$$

where $\mathrm{diag}(\cdot)$ is an operator to make a vector composed of diagonal components of its argument.

In learning of FD-ICA, null-beamformer with some reasonable directivity pattern is often used as an initial filter. In addition, since filter coefficients of FD-ICA of each frequency is learned separately, permutation ambiguity occurs. To align the permutation, a directivity pattern of the separation filter is utilized [7]. However, as shown in (5), since observed response sound is multiplied by not room transfer function but difference of room transfer function, it is difficult to find reliable directivity patterns. Therefore, we cannot expect this method performs as good as ordinary BSS.

3.2. Proposed Method: Semi-Blind Source Separation with Observed Signal of a Microphone and Direct Input of Response Sound

Since the response sound signal $r_{\mathrm{src}}(\omega)$ is known for the system, we can use this signal as an input signal of ICA. Therefore, in the proposed method, we use only one microphone element as shown in Fig. 3 and learn the separation filter of (6) in which $\boldsymbol{x}(\omega) = [x_1(\omega), r_{\mathrm{src}}(\omega)]^{\mathrm{T}}$ is substituted. Then, if we try to make an output signal $y_2(\omega)$ to include only the component of $r_{\mathrm{src}}(\omega)$, that condition can be satisfied by setting $w_{21}(\omega) = 0$ because

$$\begin{aligned} y_2(\omega) &= w_{21}(\omega)x_1(\omega) + w_{22}(\omega)r_{\mathrm{src}}(\omega) \\ &= w_{22}(\omega)r_{\mathrm{src}}(\omega). \end{aligned} \quad (10)$$

Therefore, by setting $w_{21}(\omega) = 0$ as an initial value, the learning can be started from the state where one of the signals is already separated. Since the separation of one signal is finished, now this problem is not blind nor unsupervised. We call it *semi-blind* source separation.

Although the update of (8) changes the value of $w_{21}(\omega)$, the semi-blind condition can hold by substituting $w_{21}(\omega) = 0$ in every iteration. By this constraint that $w_{21}(\omega)$ to be zero, $y_1(\omega)$ is updated to be statistically independent of $y_2(\omega) = w_{21}(\omega)r_{\mathrm{src}}(\omega)$ and the independence is satisfied when and only when

$$y_1(\omega) = C(\omega)s_1(\omega), \quad (11)$$

where $C(\omega)$ is an arbitrary value. Since $y_1(\omega)$ can be given by

$$\begin{aligned} y_1(\omega) &= w_{11}(\omega)x_1(\omega) + w_{12}(\omega)r_{\mathrm{src}}(\omega) \\ &= \left(w_{11}(\omega)\Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}'(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega) + w_{12}(\omega)\right)r_{\mathrm{src}}(\omega) \\ &\quad + w_{11}(\omega)s_1(\omega), \end{aligned} \quad (12)$$

the condition (11) yields

$$\begin{aligned} &\left(w_{11}(\omega)\Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}'(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega) + w_{12}(\omega)\right)r_{\mathrm{src}}(\omega) = 0 \\ &\quad \rightleftharpoons \frac{w_{21}(\omega)}{w_{11}(\omega)} = -\Delta\boldsymbol{g}_k(\omega)\boldsymbol{H}'(\omega)\boldsymbol{g}_{\mathrm{pri}}(\omega). \end{aligned} \quad (13)$$

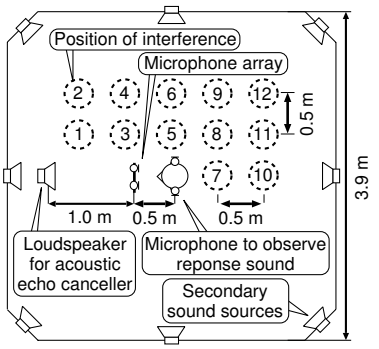

Fig. 4. Layout of acoustic environment room.

Therefore, the separation filter is optimum only when $w_{12}(\omega)/w_{11}(\omega)$ identifies the minus transfer function between the input of the inverse filter to the microphone element. In fact, the output signals $\boldsymbol{p}(\omega)$ of the projection back in (9) is given by

$$\boldsymbol{p}(\omega) = \begin{bmatrix} x_1(\omega) + \frac{w_{12}(\omega)}{w_{11}(\omega)}r_{\mathrm{src}}(\omega) \\ r_{\mathrm{src}}(\omega) \end{bmatrix} \quad (14)$$

and agree to (13). On one hand, BSS aims to make an inverse filter of the transfer system and requires more filter length than the transfer functions. To obtain a good performance with long filter length, FD-ICA requires long input signals. On the other hand, the proposed semi-blind source separation requires only equal length of filter to that of the transfer system.

In addition, since increasing the number of microphone elements in the MOMNI method lowers the stability of sound field control, decreasing one microphone element is beneficial to the MOMNI method.

4. SIMULATION

In this section, we present two experiments in which the proposed method is compared with the conventional methods, i.e., an acoustic echo canceller and the MOMNI method, and the simple connection of BSS to the MOMNI method discussed in Sect. 3.1. To validate the robustness of the proposed method against the fluctuation of the room transfer functions, we perform a response sound elimination experiment in which changes in the transfer functions are simulated. Then we evaluate the performance of each method on the basis of a speech recognition experiment to verify the applicability of the proposed method to a spoken dialogue system.

4.1. Experimental Conditions

Figure 4 shows the arrangement of the apparatuses. We placed a dummy head, which has an average human head and an upper body, at the user's position. We designed the filters used in the MOMNI and the proposed method with the room transfer functions before fluctuation. We gave the AEC the room transfer functions before fluctuation as its filter coefficients, assuming that its adaptation was performed accurately without errors before the fluctuation of the transfer functions. However, after the fluctuation, the adaptation could not be performed due to double-talk. We evaluated the performances with the average of 12 kinds of impulse responses caused by movements of a mannequin. The interelement spacing was 30 cm with the conventional the MOMNI method, and 6 cm with the simple connection of BSS. The sampling frequency was 16 kHz. In the learning of ICA, we used the input signals of early 5 seconds. The length of the separation filters is 2048 taps.

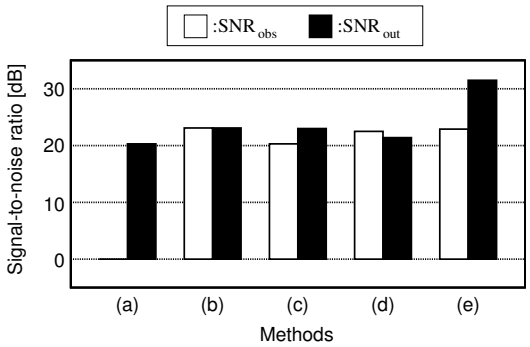

Fig. 5. Comparison of SNR_{obs} and SNR_{out} of (a)acoustic echo canceller, (b)MOMNI with 1 microphone microphone element, (c) MOMNI with Delay-and-sum with 2 elements, (d) simple connection of ICA and MOMNI with two microphone elements and (e) proposed method.

Table 1. Experimental conditions for speech recognition

Task	Newspaper dictation from JNAS [8]
Feature vector	12 MFCCs, 12 ΔMFCCs, Δpower
Language model	Newspaper dictation with 20,000 words
Phoneme model	Phonetic Tied Mixture (PTM) [8]
Decoder	Julius ver. 3.4.2 standard [8]
User's speech	200 sentences (23 males and 23 females)
Response sound	female utterance

4.2. Evaluation of Response Sound Elimination

We evaluated signal-to-noise ratios of the observed signal (SNR_{obs}) and final output signal (SNR_{out}) of the system in Fig 5. These SNRs are just the power ratios of the user's speech and the response sound. Therefore, distortion of spectrum doesn't influence these scores. When two microphone elements are used, we evaluated their average. Regarding SNR_{obs}, the result of one microphone element shows higher performance than two microphone elements. However, by the effect of delay-and-sum array signal processing, SNR_{out} of two elements is recovered to the same level of one element. This reveals that the condition of eight loudspeakers and two microphones is a hard condition for stable control of the MOMNI method, and its performance doesn't agree with the law described in Sect. 2.2 that error is proportional to $1/\sqrt{MK}$. In the simple combination of BSS and the MOMNI method, BSS cannot improve SNR_{out} from its input because of its poor initial filter and difficulty in solution of permutation. However, the proposed method improves SNR_{out} considerably. This shows that the efficacy of semi-blind source separation.

4.3. Speech Recognition Experiment

The effect of the response sound elimination is evaluated using a large vocabulary continuous speech recognition task. To evaluate the speech recognition performance, we adopt word accuracy (WA) as an evaluation score[8]. Table 1 lists the experimental conditions for the speech recognition.

Figure 6 shows the WAs with all the combinations. All the scores in the graph are almost proportional to those of SNRs except for the simple connection of BSS and the MOMNI method. Because of the permutation discussed in Sect. 3.1, simple connection has large distortion and its performance is worse than the MOMNI method. The proposed method is not so much affected by permutation and shows the highest performance.

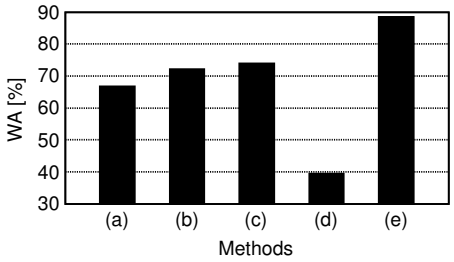

Fig. 6. Comparison of WAs of (a) acoustic echo canceller, (b) MOMNI with 1 microphone microphone element, (c) MOMNI with Delay-and-sum with 2 elements, (d) simple connection of ICA and MOMNI with two microphone elements and (e) Proposed method.

5. CONCLUSION

We proposed a semi-blind source separation algorithm and applied it to the spoken dialogue interface using sound field control. As the results of the experiment, the robustness of sound elimination and the performance of speech recognition improved with the proposed method. From these findings, the efficacy of the proposed method is ascertained.

6. REFERENCES

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

[2] S. Makino and S. Shimauchi, "Stereophonic acoustic echo cancellation — an overview and recent solutions," in *Proc. The 1999 IEEE Workshop on Acoustic Echo and Noise Control*, pp. 12–19, September 1999.

[3] W. Herbordt, J. Ying, H. Buchner, and W. Kellermann, "A real-time acoustic human-machine front-end for multimedia applications integrating robust adaptive beamforming and stereophonic acoustic echo cancellation," in *Proc. 7th International Conf. on Spoken Language Processing*, vol. 2, pp. 773–776, September 2002.

[4] Y. Hinamoto, K. Mino, H. Saruwatari, and K. Shikano, "Interface for barge-in free spoken dialogue system based on sound field control and microphone array," in *Proc. 2003 IEEE International Conf. on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 505–508, April 2003.

[5] 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, March 2003.

[6] N. Murata and S Ikeda, "An On-line Algorightm for Blind Source Separation on Speech Signals," in *Proc. 1998 International Symposium on Nonlinear Theory and its Applications*, vol. 3, pp. 923-926, September, 1998.

[7] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. 2000 IEEE International Conf. on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3140–3143, June 2000.

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