PERMUTATION-ROBUST STRUCTURE FOR ICA-BASED BLIND SOURCE EXTRACTION

Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano

Nara Institute of Science and Technology, Ikoma, Nara 630-0192 JAPAN (e-mail: yuu-t@is.naist.jp)

ABSTRACT

In this paper, we investigate a new blind source separation (BSS) structure from a permutation-robustness viewpoint, to mitigate the permutation problem which commonly arises in frequency-domain independent component analysis (ICA). Permutation robustness means that how much the BSS method is not affected under a certain probability of arising permutation, unlike the conventional permutation-solving approaches. We address to analyze our previously proposed BSS architecture, so called blind spatial subtraction array (BSSA). In BSSA, source extraction is achieved by subtracting the power spectrum of the estimated noise via ICA from the power spectrum of partly speech-enhanced signal via delay-and-sum (DS) procedure. Indeed BSSA partially involves permutation problem in the ICA-based noise estimator part. However, BSSA can efficiently reduce the negative affection of the permutation owing to the over-subtraction in the spectral subtraction and defocusing properties in DS. Experiments using artificial and real-recording-based simulations reveal that the proposed method outperforms the conventional ICA.

Index Terms— Speech enhancement, acoustic signal processing, acoustic arrays

1. INTRODUCTION

Blind source separation (BSS) is the approach to estimate original sources using only information of observed signals. Recently, various BSS methods based on independent component analysis (ICA) [1] have been presented for acoustic-sound separation [2, 3]. Particularly, frequency-domain ICA (FDICA) [2] is the most popular approach to address the convolutive BSS problem. In FDICA, however, source permutation ambiguity arises in each frequency bin, and heavily decreases the resultant quality. Therefore, it is indispensable for us to align the permutation so that each separated signal contains frequency components from the same source. Although various permutation solvers, e.g., direction-of-arrival (DOA) based method, have been proposed [2, 3, 4], permutation problem cannot be solved completely. In addition, increase of the permutation-salvaging accuracy requires higher computational costs.

To mitigate the problems, in this paper, we investigate a new BSS structure from a "permutation-robustness" viewpoint, unlike the conventional permutation-solving approaches. Permutation robustness means that how much the BSS method is not affected under a certain probability of arising permutation, and such an important property has never been studied so far in the previous ICA researches. The improvement of permutation robustness with small computations is a novel and efficient way for increasing the BSS quality. How can we construct a permutation-robust BSS? The answer is within our previously proposed BSS architecture, so called blind spatial subtraction array (BSSA) [5]. In BSSA, source extraction is achieved by subtracting the power spectrum of the estimated noise via ICA from the power spectrum of partly speech-enhanced signal via delay-and-sum (DS) procedure. Indeed BSSA partially involves permutation problem in the ICA-based noise estimator part. However, BSSA can efficiently suppress the negative affection of the permutation owing to the over-subtraction in spectral subtraction and defocusing properties in DS. Efficacy of the proposed method can be revealed by artificial and real-recording-based simulations.

This work was partly supported by MEXT e-Society leading project.

2. BLIND SPATIAL SUBTRACTION ARRAY [5]

2.1. Overview of BSSA

BSSA consists of a delay-and-sum array (DS) based primary path and a reference path for the ICA-based noise estimation (see Fig. 1). The estimated noise component by ICA is efficiently subtracted from the primary path in the power-spectrum domain without phase information. The detailed signal processing is shown below.

2.2. Partial speech enhancement in primary path

First, the short-time analysis of observed signals is conducted by a frame-by-frame discrete Fourier transform (DFT). The J-channel-array's observed signal is given by

$$\boldsymbol{X}(f,\tau) = [X_1(f,\tau),\ldots,X_J(f,\tau)]^{\mathrm{T}} = \boldsymbol{A}(f)\{\boldsymbol{S}(f,\tau)+\boldsymbol{N}(f,\tau)\}, \quad (1)$$

$$\boldsymbol{S}(f,\tau) = [\underbrace{0,\ldots,0}_{U-1}, S_U(f,\tau), \underbrace{0,\ldots,0}_{K-U}]^{\mathrm{T}}, \quad (2)$$

$$\boldsymbol{N}(f,\tau) = [N_1(f,\tau),..,N_{U-1}(f,\tau),0,N_{U+1}(f,\tau),..,N_K(f,\tau)]^{\mathrm{T}}, \quad (3)$$

where f is the frequency bin and τ is the frame number, $\boldsymbol{A}(f)$ is a mixing matrix, $\boldsymbol{S}(f,\tau)$ is a target speech signal vector, $\boldsymbol{N}(f,\tau)$ is a noise signal vector, U expresses the target speech number, and K is the number of sound sources. In the primary path, the target speech signal is partly enhanced in advance by DS. This can be given as

$$\begin{aligned} Y(f,\tau) &= \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{X}(f,\tau) \\ &= \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{S}(f,\tau) + \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{N}(f,\tau), \end{aligned} \quad (4)$$

$$\boldsymbol{W}_{\mathrm{DS}}(f) = [W_1^{(\mathrm{DS})}(f),\ldots,W_J^{(\mathrm{DS})}(f)]^{\mathrm{T}}, \quad (5)$$

$$W_j^{(\mathrm{DS})}(f) = \frac{1}{J}\exp\left(-i2\pi(f/M)f_{\mathrm{s}}d_j\sin\theta_U/c\right), \quad (6)$$

where $Y(f,\tau)$ is a primary-path output which slightly enhances target speech, $\boldsymbol{W}_{\mathrm{DS}}(f)$ is a filter coefficient vector of DS, M is the DFT size, f_{s} is a sampling frequency, d_j is a microphone position, and c is sound velocity. Besides, θ_U is the estimated DOA of the target speech, which is given by ICA part in Sect. 2.3. In Eq. (4), the second term in the right-hand side expresses the remaining noise in the output of the primary path.

2.3. ICA-based noise estimation in reference path

The proposed BSSA includes ICA-based noise estimation in the reference path. In ICA part, we perform signal separation using a complex valued unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$, so that the output signals $\boldsymbol{O}(f,\tau) = [O_1(f,\tau),\ldots,O_K(f,\tau)]^{\mathrm{T}}$ become mutually independent;

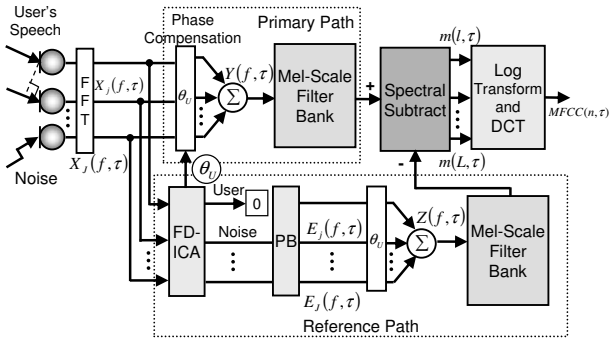

Fig. 1. Block diagram of proposed BSSA.

this procedure can be represented by

$$\boldsymbol{O}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{X}(f,\tau) \quad (7)$$

$$\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f) = \mu\left[\boldsymbol{I} - \langle\boldsymbol{\Phi}(\boldsymbol{O}(f,\tau))\boldsymbol{O}^{\mathrm{H}}(f,\tau)\rangle_\tau\right]\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f), \quad (8)$$

where μ is the step-size parameter, $[p]$ is used to express the value of the p-th step in the iterations, and $\boldsymbol{I}$ is an identity matrix. Besides, $\langle\cdot\rangle_\tau$ denotes a time-averaging operator, $\boldsymbol{M}^{\mathrm{H}}$ denotes conjugate transpose of matrix $\boldsymbol{M}$, and $\boldsymbol{\Phi}(\cdot)$ is the appropriate nonlinear vector function [3]. At the same time, we can estimate DOAs by looking at null directions in the directivity pattern which is shaped by $\boldsymbol{W}_{\mathrm{ICA}}(f)$ [3], and we designate DOA of the target speech signal as θ_U. In the reference path, target signal is not required because we want to estimate only the noise component. Accordingly we remove the separated speech component $O_U(f,\tau)$ from ICA outputs $\boldsymbol{O}(f,\tau)$, and construct the following "noise-only vector," $\boldsymbol{Q}(f,\tau)$;

$$\boldsymbol{Q}(f,\tau) = [O_1(f,\tau), ..., O_{U-1}(f,\tau), 0, O_{U+1}(f,\tau), ..., O_K(f,\tau)]^{\mathrm{T}}. \quad (9)$$

Next, we apply the projection back (PB) [2] method to remove the ambiguity of amplitude. This procedure can be represented as

$$\boldsymbol{E}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}^{+}(f)\boldsymbol{Q}(f,\tau), \quad (10)$$

where $\boldsymbol{M}^{+}$ denotes Moore-Penrose pseudo inverse matrix of $\boldsymbol{M}$. Here, $\boldsymbol{Q}(f,\tau)$ is composed of only noise components. Therefore, $\boldsymbol{E}(f,\tau)$ is a good estimation of the received noise signals at the array;

$$\boldsymbol{E}(f,\tau) \simeq \boldsymbol{A}(f)\boldsymbol{N}(f,\tau). \quad (11)$$

Finally, we obtain the estimated noise signal $Z(f,\tau)$ by performing DS as follows:

$$Z(f,\tau) = \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{E}(f,\tau) \simeq \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{N}(f,\tau). \quad (12)$$

Equation (12) is expected to be equal to the noise term of Eq. (4) in the primary path.

2.4. Source extraction processing

In the proposed BSSA, source extraction is carried out by subtracting the estimated noise power spectrum (Eq. (12)) from the partly enhanced target speech power spectrum (Eq. (4)); thus

$$Y_{\mathrm{BSSA}}(f,\tau) = \begin{cases} \left\{|Y(f,\tau)|^2 - \beta\cdot|Z(f,\tau)|^2\right\}^{\frac{1}{2}} \\ \qquad (\text{ if } \ |Y(f,\tau)|^2 - \beta\cdot|Z(f,\tau)|^2 \geq 0\ ), \\ \gamma\cdot|Y(f,\tau)| \quad (\text{otherwise}), \end{cases} \quad (13)$$

where $Y_{\mathrm{BSSA}}(f,\tau)$ is the output of BSSA, β is an over-subtraction parameter, and γ is a flooring parameter. The appropriate setting, e.g., $\beta > 1$ and $1 \gg \gamma > 0$, give an efficient noise reduction. Finally, we perform mel-scale filter bank analysis, log transform and discrete cosine transform to obtain mel-frequency cepstrum coefficient for speech recognizer [5].

3. PERMUTATION-ROBUSTNESS ANALYSIS IN BSSA

3.1. Overview

In this section, we present a permutation-robustness analysis in BSSA architecture. In the conventional ICA, when the permutation arises, we directly suffer from the permuted noise component which is wrongly regarded as the target signal. Thus the conventional ICA has no robustness against the permutation. On the other hand, in BSSA, adverse effect by the permutation is mitigated because spectral-subtraction-based source extraction technique reduces the permuted component, and DS defocuses the component arriving from out of look direction. Therefore, we can say that BSSA architecture is a permutation-robust structure. The detailed analysis is shown below.

3.2. Permutation robustness by over-subtraction

Here, we assume that source separation was performed perfectly by FDICA except for arising permutation in the frequency bin f_{p}. Under this assumption, the estimated target speech signal in the frequency bin f_{p} by ICA (including PB processing) can be described as

$$\boldsymbol{Y}_{\mathrm{ICA}}(f_{\mathrm{p}},\tau) = \boldsymbol{A}(f_{\mathrm{p}})\boldsymbol{N}_{\boldsymbol{e}}(f_{\mathrm{p}},\tau), \quad (14)$$

$$\boldsymbol{N}_{\boldsymbol{e}}(f_{\mathrm{p}},\tau) = [\underbrace{0,\ldots,0}_{n-1}, N_n(f_{\mathrm{p}},\tau), \underbrace{0,\ldots,0}_{K-n}]^{\mathrm{T}}, \quad (15)$$

where $\boldsymbol{Y}_{\mathrm{ICA}}(f_{\mathrm{p}},\tau)$ is the output signal vector as a target by ICA, $\boldsymbol{N}_{\boldsymbol{e}}(f_{\mathrm{p}},\tau)$ is a noise signal vector estimated as target speech signal vector by mistake, $N_n(f_{\mathrm{p}},\tau)$ is a noise component estimated as target speech component by mistake, and $n(\neq U)$ expresses the component number of noise. Moreover, since $\boldsymbol{N}_{\boldsymbol{e}}(f_{\mathrm{p}},\tau)$ is composed of zero components except the specific noise component $N_n(f_{\mathrm{p}},\tau)$, $\boldsymbol{Y}_{\mathrm{ICA}}(f_{\mathrm{p}},\tau)$ can be rewritten as

$$\boldsymbol{Y}_{\mathrm{ICA}}(f_{\mathrm{p}},\tau) = \hat{\boldsymbol{A}}(f_{\mathrm{p}})N_n(f_{\mathrm{p}},\tau), \quad (16)$$

$$\hat{\boldsymbol{A}}(f_{\mathrm{p}}) = [A_{1n}(f_{\mathrm{p}}),\ldots,A_{Jn}(f_{\mathrm{p}})]^{\mathrm{T}}, \quad (17)$$

where $\hat{\boldsymbol{A}}(f_{\mathrm{p}})$ is a transfer function vector of the noise component $N_n(f_{\mathrm{p}},\tau)$, and $A_{ij}(f)$ expresses an element of the mixing matrix $\boldsymbol{A}(f)$.

On the other hand, the estimated noise signal in the reference path of BSSA can be represented by

$$Z(f_{\mathrm{p}},\tau) = \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\boldsymbol{L}(f_{\mathrm{p}},\tau), \quad (18)$$

$$\boldsymbol{L}(f_{\mathrm{p}},\tau) = [L_1(f_{\mathrm{p}},\tau), ..., L_{n-1}(f_{\mathrm{p}},\tau), 0, L_{n+1}(f_{\mathrm{p}},\tau), ..., L_K(f_{\mathrm{p}},\tau)]^{\mathrm{T}}, \quad (19)$$

where $\boldsymbol{L}(f_{\mathrm{p}},\tau)$ is the estimated noise component vector including the target signal by mistake. Note that the observed signal $\boldsymbol{X}(f_{\mathrm{p}},\tau)$ can be rewritten as $\boldsymbol{X}(f_{\mathrm{p}},\tau) = \boldsymbol{A}(f_{\mathrm{p}})\{\boldsymbol{L}(f_{\mathrm{p}},\tau) + \boldsymbol{N}_{\boldsymbol{e}}(f_{\mathrm{p}},\tau)\}$. When $|Y(f_{\mathrm{p}},\tau)|^2 - \beta\cdot|Z(f_{\mathrm{p}},\tau)|^2 \geq 0$, using Eqs. (4) and (18), we can write the expectation of the power spectrum of BSSA output as

$$\begin{aligned} E\left[|Y_{\mathrm{BSSA}}(f_{\mathrm{p}},\tau)|^2\right] &= E\left[|Y(f_{\mathrm{p}},\tau)|^2 - \beta\cdot|Z(f_{\mathrm{p}},\tau)|^2\right] \\ &= E\left[|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{X}(f_{\mathrm{p}},\tau)|^2 - \beta\cdot|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\boldsymbol{L}(f_{\mathrm{p}},\tau)|^2\right] \\ &= E\left[|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\left\{\boldsymbol{L}(f_{\mathrm{p}},\tau) + \boldsymbol{N}_{\boldsymbol{e}}(f_{\mathrm{p}},\tau)\right\}|^2\right] \\ &\quad - E\left[\beta\cdot|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\boldsymbol{L}(f_{\mathrm{p}},\tau)|^2\right] \\ &\simeq (1-\beta)\cdot E\left[|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\boldsymbol{L}(f_{\mathrm{p}},\tau)|^2\right] \\ &\quad + E\left[|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\boldsymbol{N}_{\boldsymbol{e}}(f_{\mathrm{p}},\tau)|^2\right], \end{aligned} \quad (20)$$

where $E[\cdot]$ denotes the expectation operator, and we use the relation that the cross-terms among the distinct noise components are negligible with taking expectation. Since we usually set over-subtraction

parameter to $\beta > 1$, it is obvious that the first term in the right-hand side of Eq. (20) is a negative quantity and the following relation holds:

$$E\left[|Y_{\mathrm{BSSA}}(f_{\mathrm{p}},\tau)|^2\right] < E\left[|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\boldsymbol{N}_{\boldsymbol{e}}(f_{\mathrm{p}},\tau)|^2\right] = E\left[|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\hat{\boldsymbol{A}}(f_{\mathrm{p}})N_n(f_{\mathrm{p}},\tau)|^2\right]. \quad (21)$$

3.3. Permutation robustness by defocusing in DS

Under reverberant conditions, $\hat{\boldsymbol{A}}(f_{\mathrm{p}})$ can be expressed by superposition of all of reflection components. Therefore $\hat{\boldsymbol{A}}(f_{\mathrm{p}})$ can be rewritten as

$$\hat{\boldsymbol{A}}(f_{\mathrm{p}}) = \sum_q r^{(q)}\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q)}), \quad (22)$$

$$\boldsymbol{a}(f,\theta) = [a_1(f,\theta),\ldots,a_J(f,\theta)]^{\mathrm{T}}, \quad (23)$$

$$a_j(f,\theta) = \exp\left(i2\pi(f/M)f_s d_j \sin\theta/c\right), \quad (24)$$

where (q) is used to express the number of q-th reflection component, $r^{(q)}$ is a reflection coefficient, $\theta^{(q)}$ is a DOA of the reflection component of the permuted noise $N_n(f_{\mathrm{p}},\tau)$, and $\boldsymbol{a}(f,\theta)$ is a steering vector which expresses phase information of the sound source arriving from direction θ. Using Eq. (22), we can obtain the following equation,

$$|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\hat{\boldsymbol{A}}(f_{\mathrm{p}})N_n(f_{\mathrm{p}},\tau)|^2 = \Big|\sum_q r^{(q)}\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q)})N_n(f_{\mathrm{p}},\tau)\Big|^2 = \sum_q \left|r^{(q)}\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q)})N_n(f_{\mathrm{p}},\tau)\right|^2 + C_1, \quad (25)$$

where C_1 is a term which contains all of cross-terms among reflection components. Also, the power of the conventional ICA's output in the specific microphone j, $Y_{\mathrm{ICA}}^{[j]}(f_{\mathrm{p}},\tau)$, can be written as

$$|Y_{\mathrm{ICA}}^{[j]}(f_{\mathrm{p}},\tau)|^2 = \Big|\sum_q r^{(q)}a_j(f_{\mathrm{p}},\theta^{(q)})N_n(f_{\mathrm{p}},\tau)\Big|^2 = \sum_q \left|r^{(q)}a_j(f_{\mathrm{p}},\theta^{(q)})N_n(f_{\mathrm{p}},\tau)\right|^2 + C_2, \quad (26)$$

where C_2 also expresses all of cross-terms among reflection components. Here, the directivity gain of DS-filter $\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)$ is unity only when θ equals the focus direction of DS, θ_U, and it is less than one (i.e., defocused) in the other directions. This is represented by

$$\left|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{a}(f,\theta)\right| \le 1. \quad (27)$$

Thus, the power of each reflection component satisfies

$$|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{a}(f_{\mathrm{p}},\theta)|^2|r^{(q)}N_n(f_{\mathrm{p}},\tau)|^2 \le |a_j(f_{\mathrm{p}},\theta)|^2|r^{(q)}N_n(f_{\mathrm{p}},\tau)|^2 \quad (28)$$

because $|a_j(f,\theta)| = 1$ as in Eq. (24). Using the assumption that almost all the reflection components of $N_n(f_{\mathrm{p}},\tau)$ come from around the noise DOA and outside of θ_U, we can modify Eq. (28) as

$$|r^{(q)}\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q)})N_n(f_{\mathrm{p}},\tau)|^2 < |r^{(q)}a_j(f_{\mathrm{p}},\theta^{(q)})N_n(f_{\mathrm{p}},\tau)|^2. \quad (29)$$

If the interference with each reflection component is arising statistically at random, it can be expected that C_1 in Eq. (25) and C_2 in Eq. (26) become statistically the same. Therefore, the following equation holds:

$$\sum_q |r^{(q)}\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q)})N_n(f_{\mathrm{p}},\tau)|^2 + C_1 < \sum_q |r^{(q)}a_j(f_{\mathrm{p}},\theta^{(q)})N_n(f_{\mathrm{p}},\tau)|^2 + C_2. \quad (30)$$

This equation can be replaced by the following,

$$|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}\hat{\boldsymbol{A}}(f_{\mathrm{p}})N_n(f_{\mathrm{p}},\tau)|^2 < |Y_{\mathrm{ICA}}^{[j]}(f_{\mathrm{p}},\tau)|^2. \quad (31)$$

From Eqs. (21) and (31), the following relation is approved:

$$E\left[|Y_{\mathrm{BSSA}}(f_{\mathrm{p}},\tau)|^2\right] < E\left[|\boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\hat{\boldsymbol{A}}(f_{\mathrm{p}})N_n(f_{\mathrm{p}},\tau)|^2\right] < E\left[|Y_{\mathrm{ICA}}^{[j]}(f_{\mathrm{p}},\tau)|^2\right]. \quad (32)$$

This relation indicates that the power of BSSA output is less than that of ICA output in the permutation-arising frequency bin f_{p}.

On the other hand, when $|Y(f_{\mathrm{p}},\tau)|^2 - \beta\cdot|Z(f_{\mathrm{p}},\tau)|^2 < 0$, the resultant power spectrum of BSSA is floored by flooring parameter γ. If flooring parameter γ is sufficiently small, $Y_{\mathrm{BSSA}}(f_{\mathrm{p}},\tau)$ becomes smaller than the error component of the permutation.

From the above-mentioned fact, we can conclude that BSSA is permutation-robust rather than ICA. However, we must pay attention to the setting of over-subtraction parameter β. Although the oversized over-subtraction parameter β can suppress the permutation perfectly, such a parameter reduces not only noise components but also the target component in other innocent (non-permuted) frequency bins. Therefore, we should use an appropriate over-subtraction parameter β because such an oversized parameter causes an artificial distortion, so called musical noise.

4. EXPERIMENTS AND RESULT

4.1. Evaluation of permutation-robustness in BSSA

First, we compare ICA and BSSA on the basis of noise reduction rate (NRR) [3], which is defined as the output signal-to-noise ratio (SNR) minus the input SNR in dB. In this experiments, we assume that source separation is performed perfectly except for the permutation which is generated artificially in the randomly selected frequency bins. We increase permutation-arising frequency bins to evaluate the robustness against the permutation problem. Figure 2 illustrates a layout of the reverberant room in this experiment. We use speech signals (male and female) as an original speech, and input SNR is set to 0 dB at the array. Target signal is male's speech, noise is female's speech, and noise direction is 50 degrees. A four-element or eight-element array with the interelement spacing of 2 cm is used, and DFT size is 512. Over-subtraction parameter β is 1.2 and flooring coefficient γ is 0.0. Figure 3 shows the resultant curve of NRRs of ICA and BSSA with increasing permutation-arising frequency bins. From these results, we can confirm that NRR of BSSA outperforms that of ICA even if the percentage of permutation-arising increases. These results obviously indicate that BSSA involves the permutation-robust structure.

Although the previous NRR results are positive for BSSA, one might speculate that the sound distortion increases; certainly we can see the musical noise in the resultant output of the propose BSSA. Unfortunately we cannot provide distortion assessment results due to the limitation of the paper's space, but instead we show results of speech recognition which is the final goal of BSSA, where the separated sound quality is totally considered. We compare ICA and BSSA on the basis of word accuracy under the same experimental conditions. We use an eight-element array, and we generate 5% or 10% permutations artificially. We use 46 speakers (200 sentences) as the original source and we use male's speech (1 sentence) as an interference noise source. Noise direction is 50 or 80 degrees. Speech recognition task is 20 k-word dictation, acoustic model is phonetic tied mixture [7], we use 260 speakers (150 sentences / 1 speaker) as training data for acoustic model, and we use Julius [7] 3.5.1 for speech decoder. Figure 4 shows the word accuracy under each condition. From these results, we can see that the word accuracy of the proposed BSSA is superior to that of ICA under all conditions.

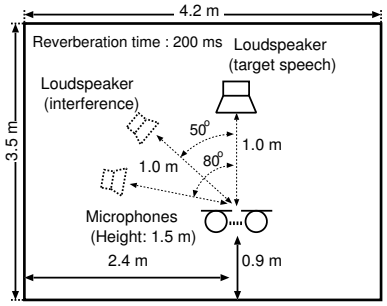

Fig. 2. Layout of reverberant room used in experiment which simulates permutation problem.

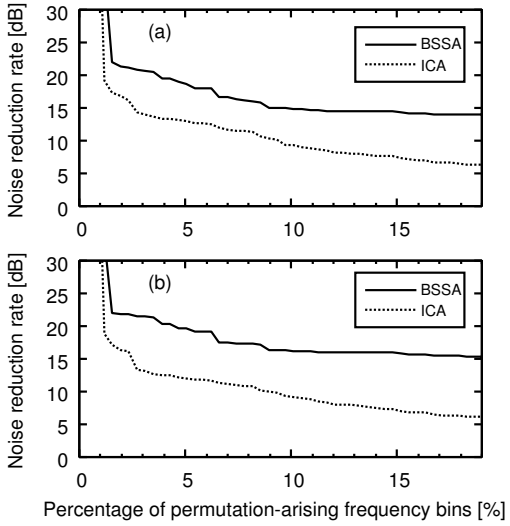

Fig. 3. Curves of NRR with increasing permutation-arising frequency bins by (a) 4-element and (b) 8-element arrays.

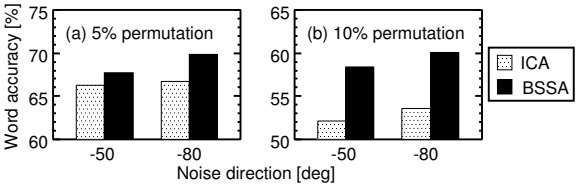

Fig. 4. Word accuracy in experiment which simulate permutation problem artificially for (a) 5%, and (b) 10% permutation.

4.2. Speech recognition test in real environment

Next, we conduct real BSS experiments, and compare DS, ICA, the conventional single-channel spectral subtraction [6] cascaded with ICA (ICA+SS), and BSSA in a real environment. In this scenario, there is not only the permutation problem but also target or noise estimation error because ICA cannot work perfectly. Figure 5 illustrates a layout of reverberant room in this experiment. Conditions and task for speech recognition are the same as those of Sect. 4.1. We use male's speech which was recored in the real environment as an interference including background noise. Input SNR is set to 10 dB. Besides, over-subtraction parameter β is 2.0 and flooring parameter γ is 0.2. Moreover, we use DOA-based permutation solver [3] in ICA. Figure 6 shows NRR and the word accuracy in each method. These results reveal that the word accuracy of the proposed BSSA are remarkably superior to those of the conventional methods. It should be mentioned that the proposed BSSA can still outperform

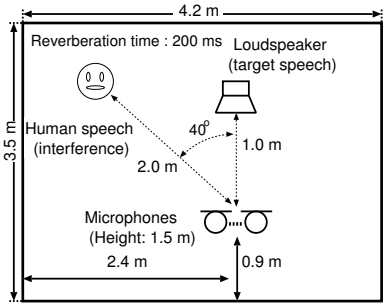

Fig. 5. Layout of reverberant room for speech recognition test in real environment.

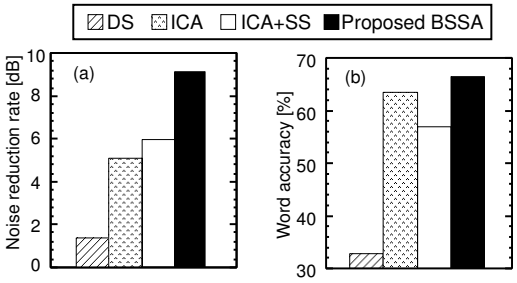

Fig. 6. (a) Result of noise reduction rate in real separation, and (b) word accuracy score of each method.

the simple combination of existing ICA and SS. This is a promising evidence that the proposed BSSA has an applicability to noise (including permutation) robust speech recognition.

5. CONCLUSIONS

In this paper, we theoretically and experimentally show that BSSA is a blind source extraction method with permutation-robust structure. BSSA is permutation-robust because over-subtraction and defocusing properties can reduce the adverse effect of permutation problem. It was confirmed that NRR and word accuracy of BSSA overtake those of the conventional ICA in the experiment which simulates permutation problem artificially. Moreover, we revealed that the word accuracy of the proposed BSSA exceeds those of DS, ICA and ICA+SS in the real environment.

6. REFERENCES

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

[2] S. Ikeda et al., "A method of ICA in the frequency domain," *Proc. Intern. Workshop on ICA and BSS*, pp.365–371, 1999.

[3] H. Saruwatari et al., "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Applied Signal Proc.*, vol.2003, no.11, pp.1135–1146, 2003.

[4] H. Sawada et al., "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech and Audio Processing*, vol.12, pp.530–538, 2004.

[5] Y. Takahashi et al., "Blind spatial subtraction array with independent component analysis for hands-free speech recognition," *Proc. of IWAENC*, 2006.

[6] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol.ASSP-27, no.2, pp.113–120, 1979.

[7] A. Lee et al., "Julius - An open source real-time large vocabulary recognition engine," *Proc. Eurospeech*, pp.1691–1694, 2001.