SOURCE ADAPTIVE BLIND SIGNAL EXTRACTION USING CLOSED-FORM ICA FOR HANDS-FREE ROBOT SPOKEN DIALOGUE SYSTEM

Yu Takahashi[†], *Hiroshi Saruwatari*[†], *Yuki Fujihara*[†], *Kentaro Tachibana*[†], *Yoshimitsu Mori*[†], *Shigeki Miyabe*[†], *Kiyohiro Shikano*[†], *Akira Tanaka*[‡]

[†]Nara Institute of Science and Technology, Ikoma, Nara, 630-0192, JAPAN
[‡]Hokkaido University, Kita-14, Nishi-9, Kita-ku, Sapporo, 060-0814, JAPAN

ABSTRACT

In this paper, we propose a new ICA-based BSS algorithm including estimation of sources' probability density functions (PDFs) to adapt the nonlinear activation function to various noise conditions. In the proposed method, closed-form second-order ICA is introduced as a computational-cost-efficient preprocessing to extract sources' PDFs, which is beneficial for real-time application. Compared with various type of conventional ICAs, e.g., fixed activation-function type and ML-based type, our proposed algorithm can give a faster and higher convergence. Based on the proposed source-adaptive ICA, we show a real-time noise reduction results under diffuse noise environment. Also we can demonstrate our recently developed hands-free robot spoken dialogue system via real-time ICA.

Index Terms— Separation, speech enhancement, acoustic signal processing, adaptive signal processing, robot

1. INTRODUCTION

Blind source separation (BSS) is the approach taken to estimate original source signals using only information of the mixed signals observed in each input channel. This technique is based on *unsupervised* filtering in that the source-separation procedure requires no training sequences and no a priori information on the directions-of-arrival (DOAs) of the sound sources. Owing to the attractive features of BSS, much attention has been paid to the BSS technique in many fields of signal processing. One promising example in acoustic signal processing is a humanoid robot auditory system [1], which constructs an indispensable basis for intelligent robot technology [2].

In this paper, we propose a new independent component analysis [3] (ICA)-based BSS algorithm including estimation of sources' probability density functions (PDFs) to adapt the nonlinear activate function to various noise conditions. Our previously proposed closed-form second-order ICA (SO-ICA) [4] is introduced as a computational-cost-efficient preprocessing to extract sources' PDFs. This feature is beneficial for real-time implementation of BSS. Compared with various type of conventional ICAs, e.g., fixed activation-function type [5, 6] and maximum likelihood (ML)-based type [7], our proposed algorithm can give a faster and higher convergence. Based on the proposed source-adaptive ICA, we show a real-time noise reduction results under realistic diffuse noise environment. Also we can demonstrate our recently developed hands-free robot spoken dialogue system [10] for a railway-station guidance task via real-time ICA.

This work was partly supported by the NEDO project for strategic development of advance robotics elemental technologies, and MIC Strategic Information and Communications R&D Promotion Programme in Japan.

2. MIXING PROCESS AND CONVENTIONAL METHODS

2.1. Mixing process

In this study, the number of microphones is K and the number of multiple sound sources is L, where we deal with the case of $K \geq L$.

Multiple mixed signals are observed at the microphone array, and these signals are converted into discrete-time series via an A/D converter. By applying the short-time discrete-time Fourier transform framewisely, we can express the observed signals, in which multiple source signals are linearly mixed, as follows in the time-frequency domain:

$$\boldsymbol{x}(f,t) = \boldsymbol{A}(f)\boldsymbol{s}(f,t), \tag{1}$$

where f and t represent frequency bin number and time index, respectively, $\boldsymbol{x}(f,t) = [x_1(f,t), \cdots, x_K(f,t)]^{\mathrm{T}}$ is the observed signal vector, and $\boldsymbol{s}(f,t) = [s_1(f,t), \cdots, s_L(f,t)]^{\mathrm{T}}$ is the source signal vector. Also, $\boldsymbol{A}(f)$ is the mixing matrix which is complex-valued because we introduce a model to deal with the relative time delays among the microphones and room reverberations.

2.2. ICA-based BSS

Next, we perform signal separation using the complex-valued unmixing matrix $\boldsymbol{W}(f)$, so that the L time-series output $\boldsymbol{y}(f,t) = [y_1(f,t), \cdots, y_L(f,t)]^{\mathrm{T}}$ becomes mutually independent; this procedure can be given as

$$\boldsymbol{y}(f,t) = \boldsymbol{W}(f)\boldsymbol{x}(f,t). \tag{2}$$

We perform this procedure with respect to all frequency bins.

Various ICA methods for optimizing $\boldsymbol{W}(f)$ have been proposed. In the conventional frequency-domain higher-order ICA (HO-ICA) [5, 6, 8], the optimal $\boldsymbol{W}_{\mathrm{HO}}(f)$ is obtained by the following iterative equation:

$$\boldsymbol{W}_{\mathrm{HO}}^{[m+1]}(f) = \mu\left[\boldsymbol{I} - \left\langle \boldsymbol{\Phi}(\boldsymbol{y}(f,t))\boldsymbol{y}^{\mathrm{H}}(f,t)\right\rangle_t\right] \cdot \boldsymbol{W}_{\mathrm{HO}}^{[m]}(f) + \boldsymbol{W}_{\mathrm{HO}}^{[m]}(f), \tag{3}$$

where $\boldsymbol{X}^{\mathrm{H}}$ denotes hermitian transpose of matrix $\boldsymbol{X}$, μ is the step-size parameter, $\boldsymbol{I}$ is an identity matrix, $[m]$ is used to express the value of the m-th step in the iteration, $\langle\cdot\rangle_t$ denotes a time-averaging operator.

Here $\boldsymbol{\Phi}(\boldsymbol{Y}(f,t))$ is the appropriate nonlinear vector function, a.k.a. *activation function*. Basically this function's element corresponding to each source $s_l(f,t)$ should be determined as

$$\Phi_l(\cdot) \quad = \quad -\frac{\partial}{\partial s_l(f,t)}\log(p_{s_l}), \tag{4}$$

where p_{s_l} is PDF of the source $s_l(f,t)$. Thus, generally speaking in HO-ICA, we need to determine the activation function in advance, or

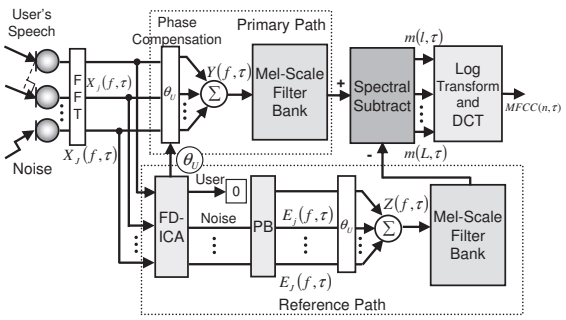

Fig. 1. Block diagram of BSSA.

estimate it via ML scheme [7]. However, we could not know a priori information of sources' PDFs in a BSS context. Also ML-based activation function estimation requires additional huge computations because this method should iteratively update the activation function form along with ICA's iterations; this results in a great drawback for real-time application. Hence typical use of ICA simply substitutes the activation function with fixed function, e.g., $\tanh(\cdot)$ for speech signal [6, 8]. This leads to a notable mismatch and bad convergence especially when we are confronted with more general acoustic signals like non-speech noises.

2.3. Blind spatial subtraction array [9]

In a hands-free system at a real environment, it is required to extract a target speech and reduce noises which cannot be regarded as point sources. Although the conventional ICA-based BSS could work especially in point sources mixing, it is difficult to apply ICA to non-point source noise reduction. BSSA has been proposed to extract a target speech in such a case. In BSSA, ICA is partly utilized as a noise estimator because of the fact that ICA is proficient in noise estimation rather than target estimation [9]. BSSA consists of two paths; a delay-and-sum (DS) array based primary path as target speech enhancing part, and ICA-based reference path as noise estimation part (see Fig. 1; FD-ICA is frequency-domain HO-ICA and PB means projection back operation using $\boldsymbol{W}(f)^{-1}$). Based on the spectral subtraction method, the BSSA's output $Y_{\mathrm{BSSA}}(f,t)$ can be given by

$$Y_{\mathrm{BSSA}}(f,t) = \begin{cases} \left\{|Y_{\mathrm{DS}}(f,t)|^2 - \gamma \cdot |Z(f,t)|^2\right\}^{\frac{1}{2}} \\ \qquad (\text{if } |Y_{\mathrm{DS}}(f,t)|^2 - \gamma \cdot |Z(f,t)|^2 \geq 0), \\ \delta \cdot |Y_{\mathrm{DS}}(f,t)| \quad (\text{otherwise}), \end{cases} \tag{5}$$

where $Y_{\mathrm{DS}}(f,t)$ is the output signal from the primary path, $Z(f,t)$ is the output signal from the reference path, γ represents over-subtraction parameter, and δ denotes the flooring parameter.

3. PROPOSED METHOD

3.1. Motivation

In a previous study, closed-form solution of SO-ICA was proposed by one of the authors [4], who showed that simple algebraic calculations enable the separation of mixed signals without iterative filter updating. This finding has motivated us to combine closed-form SO-ICA and source's PDF estimation with few computational costs.

Our algorithm consists of three stages, namely, closed-form SO-ICA for roughly separating the sources, kurtosis-based activation function estimation applied to the roughly separated signals, and post-HO-ICA with the optimized activation function for increase of the separation accuracy. This strategy is very reasonable because advance SO-ICA can separate the sources to some extent regardless of the sources' PDF (there is no activation function in SO-ICA), and then we can identify PDFs after SO-ICA. Hereinafter we describe the detailed algorithm.

3.2. First stage: closed-form SO-ICA

This subsection briefly describes the overview of signal processing in the closed-form SO-ICA (see Ref. [4] for more details). First, we obtain the correlation matrices with different time points as

$$\boldsymbol{R}_{t_i}(f) \quad = \quad \langle \boldsymbol{x}(f,t)\boldsymbol{x}(f,t)^{\mathrm{H}} \rangle_{t \in t_i}, \tag{6}$$

where $\langle\cdot\rangle_{t\in t_i}$ denotes the time-averaging operator over specific time duration t_i, and $i = 1, 2, ...$ represent indices of time-averaging block.

Next, we apply the singular value decomposition (SVD) to a superposition of $\boldsymbol{R}_{t_i}(f)$, which is represented as

$$\sum_i \boldsymbol{R}_{t_i}(f) \quad = \quad \boldsymbol{U}(f)\mathrm{diag}(\lambda_1, \lambda_2, ...)\boldsymbol{U}(f)^{\mathrm{H}}, \tag{7}$$

where λ_k are the eigenvalues, $\mathrm{diag}(\lambda_1, ...)$ denotes the diagonal matrix which includes the eigenvalues, and $\boldsymbol{U}(f)$ is the matrix consisting of the eigenvectors. Then we obtain a full-rank decomposition for pseudo-inverse of $\sum_i \boldsymbol{R}_{t_i}(f)$ as follows

$$\Big[\sum_i \boldsymbol{R}_{t_i}(f)\Big]^{+} \quad = \quad \boldsymbol{L}(f)\boldsymbol{L}(f)^{\mathrm{H}}, \tag{8}$$

$$\boldsymbol{L}(f) \quad = \quad \boldsymbol{U}(f)\mathrm{diag}(1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}, ...). \tag{9}$$

It can be proved [4] that if the covariance of the sources $\boldsymbol{s}(f,t)$ in t_i is negligible, every $\boldsymbol{L}(f)^{\mathrm{H}}\boldsymbol{R}_{t_i}(f)\boldsymbol{L}(f)$ for any i shares the same eigenvectors, and this is given via SVD form as

$$\boldsymbol{L}(f)^{\mathrm{H}}\boldsymbol{R}_{t_i}(f)\boldsymbol{L}(f) = \boldsymbol{T}(f)\mathrm{diag}(\sigma_1(t_i), \sigma_2(t_i), ...)\boldsymbol{T}(f)^{\mathrm{H}}, \tag{10}$$

where $\sigma_k(t_i)$ are the eigenvalues for a specific time block t_i, and $\boldsymbol{T}(f)$ denotes the matrix consisting of shared eigenvectors which are *independent* of time-block index i. Therefore, for any i, the simultaneous diagonalization of $\boldsymbol{R}_{t_i}(f)$ can be achieved as follows;

$$\boldsymbol{T}(f)^{\mathrm{H}}\boldsymbol{L}(f)^{\mathrm{H}}\boldsymbol{R}_{t_i}(f)\boldsymbol{L}(f)\boldsymbol{T}(f) = \mathrm{diag}(\sigma_1(t_i), \sigma_2(t_i), ...), \tag{11}$$

and this means that the optimal separation filter matrix in the 2nd-order sense is given by

$$\boldsymbol{W}_{\mathrm{SO}}(f) \quad = \quad (\boldsymbol{L}(f)\boldsymbol{T}(f))^{\mathrm{H}}. \tag{12}$$

Note that, for the calculation of $\boldsymbol{T}(f)$ in (10), it is sufficient for us to only apply a single SVD to an *arbitrary* single time-block t_i because of the eigenvector-sharing property. So the total calculation of closed-form SO-ICA is quite few, almost the same as that of one iteration in HO-ICA.

3.3. Second stage: kurtosis-based activation function estimation

After closed-form SO-ICA, the roughly separated sources can be obtained. Therefor, we can estimate PDFs of sources via the following generalized Gaussian distribution (GGD) [11] modeling and its kurtosis. This result of PDF estimation will be utilized for HO-ICA in the next stage (3rd stage). Hereafter $y(t)$ means a real part of the separated signal.

GGD is a flexible family of PDF modeling with some variable parameters, and GGD can represent various types of well-known

PDFs, e.g., Gaussian and Laplacian distributions. The definition of GGD is as follows:

$$f_{GG}(z;\alpha,\beta) = \frac{\beta}{2\alpha\Gamma\left(\frac{1}{\beta}\right)} \exp\left(-\left[\frac{|z-\bar{z}|}{\alpha} \right]^{\beta} \right), \tag{13}$$

where $\bar{z}$ is the mean of variable z, $\Gamma(x) = \int_0^{\infty} \exp(-t)t^{x-1}dt$ is a gamma function, α is *scale parameter*, and β is *shape parameter* of GGD. Figure 2 shows examples of different PDFs in GGD; note that $\beta = 2$ corresponds to Gaussian PDF and that $\beta = 1$ corresponds to Laplacian PDF.

If source's PDF obeys GGD, we can easily derive the appropriate activation function as

$$\Phi(f_{GG}(z;\alpha,\beta)) = -\frac{\partial}{\partial z}\log(f_{GG}(z;\alpha,\beta)) = \begin{cases} \frac{\beta}{\alpha^{\beta}}|z|^{\beta-1}, & (z \geq 0), \\ -\frac{\beta}{\alpha^{\beta}}|z|^{\beta-1}, & (z < 0). \end{cases} \tag{14}$$

Figure 3 depicts examples of activation functions for GGD.

The estimation of the shape parameter β is the most important issue here because β dominantly determine the activation function (see (14)) but the scale parameter α becomes negligible through scale normalization. We can introduce *kurtosis* to estimate β. In general kurtosis of signal $y(t)$ is given by

$$\mathrm{kurt}(y(t)) = \langle y^4(t)\rangle_t \langle y^2(t)\rangle_t^{-2} - 3. \tag{15}$$

The n-th order moment of GGD has the following useful relationship;

$$\langle z^n \rangle = \beta^{-n}\Gamma\left(\frac{n+1}{\beta}\right)\Gamma\left(\frac{1}{\beta}\right)^{-1}. \tag{16}$$

From (15) and (16) we have

$$\mathrm{kurt}(y(t)) = \Gamma\left(\frac{5}{\beta}\right)\Gamma\left(\frac{1}{\beta}\right)\Gamma\left(\frac{3}{\beta}\right)^{-2} - 3. \tag{17}$$

This is a monotonically decreasing function of β. Thus, we can estimate the shape parameter β by measuring kurtosis and using an inverse relationship of (17) in table-lookup manner.

In summary, we measure kurtosis of each of separated signals by closed-form SO-ICA, and then we can adaptively determine the corresponding activation function by (14) for each sound source. The required computations is only one kurtosis calculation just after SO-ICA, and consequently our method is computative efficient in comparison to ML-based activation function estimation. Possible drawback of the proposed method is PDF estimation error due to poor separation accuracy in SO-ICA. Thus, PDF estimation performance is highly related to degree of ease in source separation, e.g., reverberant conditions.

3.4. Third stage: nonclosed-form HO-ICA

The separation filter matrix $\boldsymbol{W}_{\mathrm{SO}}(f)$ obtained by SO-ICA often provides insufficient source-separation performance. To polish up the separation filter matrix and gain the further performance, we propose to combine the nonclosed-form HO-ICA after closed-form SO-ICA employing the source-PDF adapted activation function. This strategy regards the separation filter matrix $\boldsymbol{W}_{\mathrm{SO}}(f)$ as an initial value for HO-ICA's iterative learning given by (3).

In general, HO-ICA suffers from an problem of the poor and slow convergence of nonlinear optimization. In the proposed method, however, preceding closed-form SO-ICA can give a better initial state for HO-ICA, and the previous PDF estimation enables HO-ICA to use more appropriate activation function. This combination mitigates the drawbacks on the poor convergence.

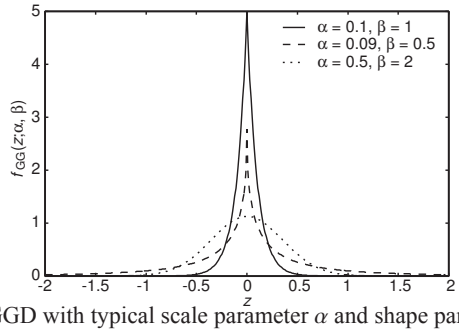

Fig. 2. GGD with typical scale parameter α and shape parameter β.

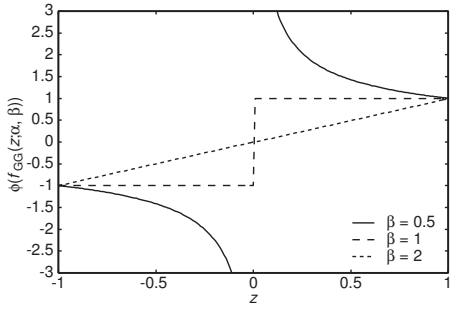

Fig. 3. Activation function for GGD with typical shape parameter β.

4. EXPERIMENTAL EVALUATION FOR ALGORITHM

To evaluate the efficacy of the proposed method, we carried out noise reduction experiments in a real reverberant room where two omnidirectional microphones are set. The reverberation time (RT) in this room is 200 ms. Target speech signal is arriving from a fixed direction, and spoken by two male and two female speakers. As for the noise, a diffuse noise recorded in an actual railway station is emitted from surrounded 36 loudspeakers. *Noise reduction rate* (NRR) [8], defined as the output signal-to-noise ratio (SNR) in dB minus the input SNR in dB, is used as the objective indication of separation performance.

Figure 4 shows the convergence curves of NRR for noise estimation part in BSSA (batch processing for 3-second data). We compare simple ICA (activation function is fixed to $\tanh(\cdot)$), ICA with ML-based activation function estimation, and the proposed method. The horizontal axis represents the total computational cost which is almost equal to the number of iterations in the HO-ICA part multiplied by the number of frequency bins, including additional computations to estimate the activation function. From the results, we can see that the proposed method has a fast and high convergence performance.

Figure 5 depicts the resultant NRR of speech extraction in BSSA output, where this BSSA is implemented to be in real-time [10]. To simulate the realistic spoken dialogue system, we made a test input consisting of noise only periods (0–35 s and 55–100 s) and noise-speech mixing parts (35–55 s; speech DOA is −20°, and 100–120 s; speech DOA is 20°). We confirm the proposed method's great efficacy still in real-time operation.

5. HANDS-FREE ROBOT SPOKEN DIALOGUE SYSTEM

Recently we develop a hands-free robot spoken dialogue system using the proposed real-time BSSA, which is mainly used for railway-station guidance in a noisy environment (see Fig. 6).

To evaluate the system, speech recognition test was conducted

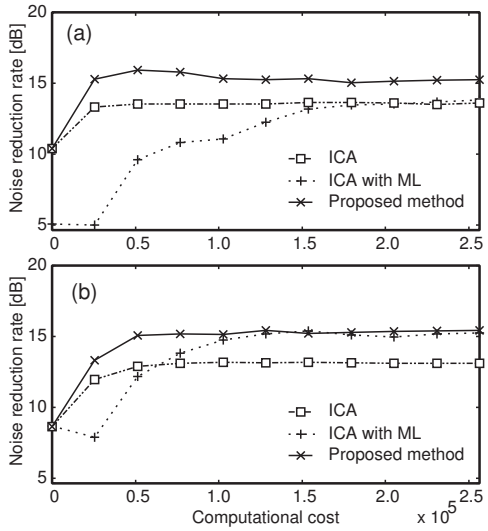

Fig. 4. Noise estimation performance where speech direction is (a) −40° and (b) 30°.

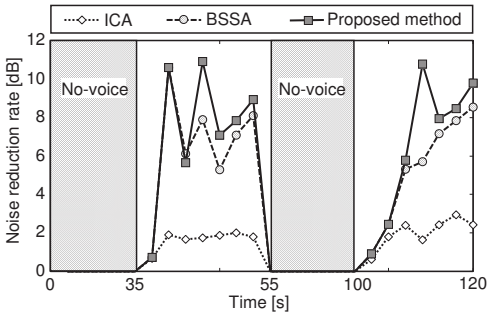

Fig. 5. Real-time implementation results, where fixed activation function is used in ICA and BSSA.

in the reverberant room where RT is more than 400 ms. The target speech is talked in front of a microphone array and 1.5 m apart. We use 5 speakers (250 words) as the target utterances. As for noise, two noises were added simultaneously. First noise is a diffuse noise recorded in an actual railway station emitted from surrounded 8 loudspeakers (it simulates railway-station noise). Second noise is an interference speech located at 50 degrees in the right direction of the microphone array, and its distance is 2.0 m. An eight-element array with the interelement spacing of 2 cm is used.

Figure 7 gives a comparative assessment example from the viewpoint of preprocessing microphone array methods, i.e., the conventional DS, ICA, or the proposed BSSA. The results reveal that both the word correct and word accuracy of the proposed BSSA are obviously superior to those of the conventional DS and ICA, and our proposed system notably sustains the recognition accuracy of more than 80%. The demonstration movie of the robot dialogue system is available in the following URL. Readers can confirm that the fluent conversation.

Demo video: **http://spalab.naist.jp/database/Demo/rtbssa/**

6. CONCLUSION

In this paper, first, we proposed a new efficient BSS algorithm combining closed-form SO-ICA and source-PDF adaptive HO-ICA,

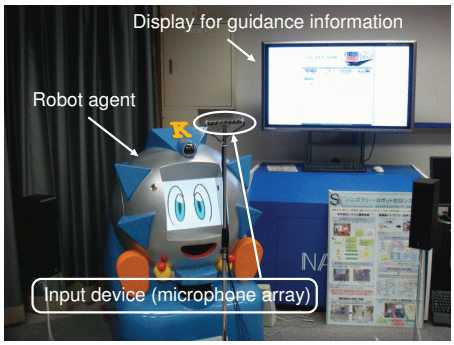

Fig. 6. Appearance of robot spoken dialogue system.

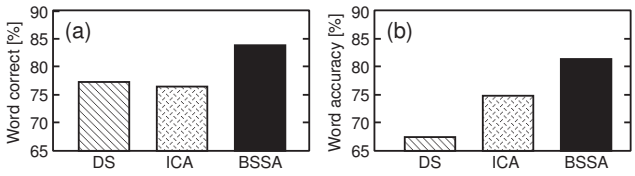

Fig. 7. Comparison of preprocessing methods in (a) word correct, and (b) word accuracy.

where the activation function in HO-ICA is optimized by using information from closed-form ICA. This enables us to improve the separation accuracy with saving the computational costs. Secondly we demonstrate our recently developed hands-free robot spoken dialogue system, and show that the proposed system can work under an advarse condition like railway station environments.

7. REFERENCES

[1] K. Nakadai, et al., "Applying scattering theory to robot audition system: robust sound source localization and extraction," *Proc. IROS-2003*, pp.1147–1152, 2003.

[2] R. Prasad, et al., "Robots that can hear, understand and talk," *Advanced Robotics*, vol.18, pp.533–564, 2004.

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

[4] A. Tanaka, et al., "Theoretical foundations of second-order-statistics-based blind source separation for non-stationary sources," *Proc. ICASSP*, pp.III-600–III-603, 2006.

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

[6] S. Ikeda and N. Murata, "A method of blind source separation based on temporal structure of signals," *Proc. ICONIP*, pp.737–742, 1998.

[7] S. Haykin (ed.), *Unsupervised Adaptive Filtering, Volume 1, Blind Source Separation*, John Wiley & Sons, 2000.

[8] H. Saruwatari et al., "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio, Speech & Language Process.*, vol.14, pp.666–678, 2006.

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

[10] Y. Takahashi, et al., "Real-time implementation of blind spatial subtraction array for hands-free robot spoken dialogue system," *Proc. IROS*, pp.1687–1692, 2008.

[11] G. Box, et al., *Bayesian Inference in Statistical Analysis*, Adison Wesley, Reading, Massachusetts, 1973.