# STABLE LEARNING ALGORITHM FOR LOW-DISTORTION BLIND SEPARATION OF REAL SPEECH MIXTURE COMBINING MULTISTAGE ICA AND LINEAR PREDICTION

*Tsuyoki NISHIKAWA Hiroshi SARUWATARI Kiyohiro SHIKANO*

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, JAPAN, E-mail: tsuyo-ni@is.aist-nara.ac.jp

## ABSTRACT

We propose a stable algorithm for blind source separation (BSS) combining multistage ICA (MSICA) and linear prediction. The MSICA in which frequency-domain ICA (FDICA) for a rough separation is followed by time-domain ICA (TDICA) to remove residual crosstalk. For temporally correlated signals, we must use TDICA with a nonholonomic constraint to avoid the decorrelation effect from the holonomic constraint. However, the stability cannot be guaranteed in the nonholonomic case. To solve the problem, the linear predictors estimated from the roughly separated signals by FDICA are inserted before the holonomic TDICA as a prewhitening processing, and the dewhitening is performed after TDICA. The stability of the proposed algorithm can be guaranteed by the holonomic constraint, and the pre/dewhitening processing prevents the decorrelation.

## 1. INTRODUCTION

Blind source separation (BSS) is an approach for estimating original source signals only from the information of the mixed signals observed in each input channel. This technique is applicable to high-quality hand-free speech recognition systems. Many BSS methods based on independent component analysis (ICA) [1] have been proposed [2, 3] for the acoustic signal separation. However, the performances of these methods degrade seriously, especially under heavily reverberant conditions.

In order to improve the separation performance, we have proposed multistage ICA (MSICA)[4], in which frequency-domain ICA (FDICA) [3, 5] and time-domain ICA (TDICA) [2] are combined. In this method, first, FDICA can find an approximate solution to separate the sources to a certain extent, and finally TDICA can remove the residual crosstalk components from FDICA. Therefore, the improvement of TDICA is a primary issue because the quality of resultant separated signals is determined by TDICA. In this paper, we discuss the stability of the TDICA algorithm, and newly propose a stable algorithm combining MSICA and linear prediction for temporally correlated signals, e.g., speech signals. First, the following points are explicitly noted: (1) The stability of learning in conventional TDICA with a holonomic constraint (H-TDICA) [2] is highly acceptable. However, the method cannot work well for speech signals due to the deconvolution property; i.e., the separated speech is harmfully distorted by the whitening process. (2) To decrease the whitening effect, TDICA with a nonholonomic constraint (NH-TDICA) has been proposed [6]. This method, however, includes the inherent drawback that the stability of learning cannot be guaranteed. In order to solve both problems simultaneously, we propose the novel approach in which the linear predictors estimated from the roughly separated source signals by FDICA are inserted before H-TDICA as a prewhitening processing (after TDICA, the dewhitening is also performed). The stability of the learning in TDICA can be guaranteed by the holonomic constraint, and it is still possible to separate the temporally correlated signals because the pre/dewhitening processing prevents the ICA from performing the decorrelation.

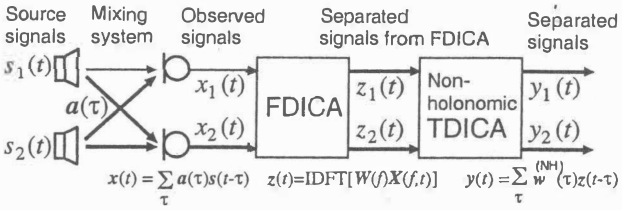

**Fig. 1**. Blind source separation procedure performed in the original MSICA [4].

## 2. SOUND MIXING MODEL

In general, the observed signals in which multiple source signals are convoluted with room impulse responses are obtained by $\boldsymbol{x}(t) = \sum_{\tau=0}^{P-1} \boldsymbol{a}(\tau)\ \boldsymbol{s}(t-\tau)$, where $\boldsymbol{x}(t) = [x_1(t), \cdots, x_K(t)]^{\mathrm{T}}$ is the observed signal vector and $\boldsymbol{s}(t) = [s_1(t), \cdots, s_L(t)]^{\mathrm{T}}$ is the source signal vector (see Fig. 1) . $K$ is the number of array elements (microphones) and $L$ is the number of multiple sound sources. In this study, we deal with the case of $K = L = 2$. Also, $\boldsymbol{a}(\tau) = [a_{ij}(\tau)]_{ij}$ ($[\cdot]_{ij}$ denotes the matrix in which $ij$-th element is $[\cdot]$) is the mixing filter matrix. $P$ is the length of the impulse response.

## 3. CONVENTIONAL ICA AND PROBLEMS

### 3.1. BSS Algorithm Based on MSICA [4]

Figure 1 shows the procedure of the original MSICA. MSICA is conducted in the following steps. First, we perform FDICA to separate the source signals to some extent with the advantage of high stability. Second, we regard the separated signals $\boldsymbol{z}(t)$ from FDICA as the input signals for TDICA, and we can remove the residual crosstalk components of FDICA by using TDICA. Finally, we regard the output signals from TDICA as the resultant separated signals. The separated signals of MSICA can be given as $\boldsymbol{y}(t) = \sum_{\tau=0}^{Q-1} \boldsymbol{w}(\tau)\ \boldsymbol{z}(t-\tau)$, where $\boldsymbol{y}(t) = [y_1(t), \cdots, y_L(t)]^{\mathrm{T}}$ is the resultant separated signal vector of MSICA and $\boldsymbol{z}(t) = [z_1(t), \cdots, z_L(t)]^{\mathrm{T}}$ is the input signal vector for the TDICA part in MSICA (i.e., the output signals from FDICA). Also, $\boldsymbol{w}(\tau) = [w_{ij}(\tau)]_{ij}$ is the separation filter matrix, and $Q$ is the length of the separation filter. In this procedure, we optimize $\boldsymbol{w}(\tau)$ so that the separated signals are mutually independent.

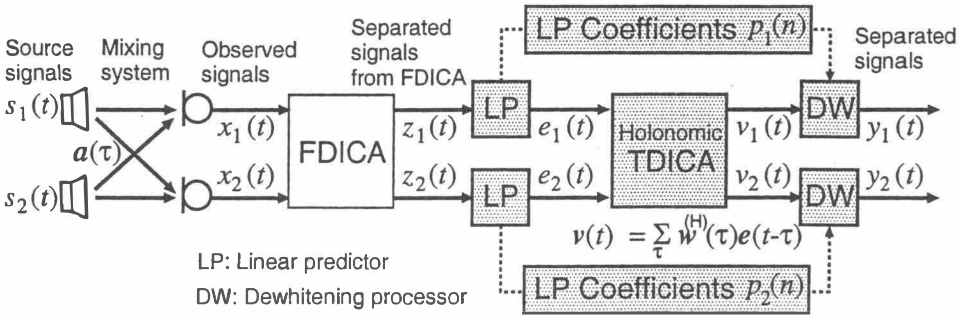

**Fig. 2**. Blind source separation procedure performed in the proposed algorithm combining MSICA and linear prediction.

The selection of TDICA is an important issue because the quality of resultant separated signals is determined by TDICA. We have two choices for TDICA algorithms; H-TDICA [2] and NH-TDICA [6]. In the next section, detailed explanations for each algorithm and their problems are described.

### 3.2. Conventional Holonomic TDICA

Amari proposed the TDICA algorithm which optimizes the separation filter by minimizing the Kullback-Leibler divergence (KLD) between the joint probability density function and the marginal probability density function of the separated signals [2]. The iterative equation of the separation filter $\boldsymbol{w}^{(\mathrm{H})}(\tau)$ to minimize the KLD is given as (hereafter we designate the iterative equation as "**H-TDICA**"):

$$\begin{aligned}\boldsymbol{w}_{i+1}^{(\mathrm{H})}(\tau) &= \boldsymbol{w}_i^{(\mathrm{H})}(\tau) + \alpha \sum_{d=0}^{Q-1} \Big\{ \boldsymbol{I}\delta(\tau - d) \\ &\quad - \langle \boldsymbol{\phi}(\boldsymbol{y}(t))\boldsymbol{y}(t-\tau+d)^{\mathrm{T}} \rangle_t \Big\} \boldsymbol{w}_i^{(\mathrm{H})}(d), \qquad (1)\end{aligned}$$

where $\langle\cdot\rangle_t$ denotes the time-averaging operator, $i$ is used to express the value of the $i$-th step in the iterations, $\alpha$ is the step-size parameter and $\boldsymbol{I}$ is the identity matrix. $\delta(\tau)$ is Dirac delta function, where $\delta(0) = 1$ and $\delta(n) = 0$ $(n \neq 0)$. Also, we define the nonlinear vector function $\boldsymbol{\phi}(\boldsymbol{y}(t)) \equiv \tanh(y_1(t)), \cdots, \tanh(y_L(t))]^{\mathrm{T}}$.

### 3.3. Conventional Nonholonomic TDICA

The H-TDICA forces the separated signals to have the characteristic that their higher-order autocorrelation is $\delta(\tau)$, i.e., the signals are temporally decorrelated. This performance might have a negative influence on the source separation. In order to solve the problem, Choi proposed a modified TDICA algorithm with a nonholonomic constraint [6]. In this algorithm, the constraint for the diagonal component of $\{\cdot\}$ part in Eq. (1), i.e., the higher-order autocorrelation of separated signals, is set to be arbitrary. The iterative equation of the separation filter $\boldsymbol{w}^{(\mathrm{NH})}(\tau)$ is given as (hereafter we designate the iterative equation as "**NH-TDICA**"):

$$\begin{aligned}\boldsymbol{w}_{i+1}^{(\mathrm{NH})}(\tau) &= \boldsymbol{w}_i^{(\mathrm{NH})}(\tau) \\ &\quad + \alpha \sum_{d=0}^{Q-1} \Big\{ \mathrm{diag}\Big( \langle \boldsymbol{\phi}(\boldsymbol{y}(t))\boldsymbol{y}(t-\tau+d)^{\mathrm{T}} \rangle_t \Big) \\ &\quad - \langle \boldsymbol{\phi}(\boldsymbol{y}(t))\boldsymbol{y}(t-\tau+d)^{\mathrm{T}} \rangle_t \Big\} \boldsymbol{w}_i^{(\mathrm{NH})}(d). \qquad (2)\end{aligned}$$

We have also introduced Eq. (2) in the original MSICA [4] to separate the mixed speech which corresponds to the temporally correlated signal by utilizing the flexibility of the nonholonomic constraint.

### 3.4. Problems in Conventional TDICAs

The advantage and disadvantage of conventional TDICAs can be summarized as follows. (1) The stability of learning in H-TDICA is satisfactory. However, the method cannot work well for speech signals due to the deconvolution property; i.e., the separated speech is harmfully distorted by the whitening process. (2) On the other hand, NH-TDICA possibly performs no deconvolution, i.e., NH-TDICA is applicable to speech signals. This method, however, includes the inherent drawback that the stability of learning cannot be guaranteed as described in Sect. 5.2. Thus, the separation of temporally correlated signals such as speech cannot be achieved only using the conventional TDICAs.

## 4. PROPOSED ALGORITHM COMBINING MSICA AND LINEAR PREDICTION

This section describes a new stable algorithm combining the linear prediction technique with an original MSICA (see Fig. 1). In the proposed algorithm, the linear predictors estimated from the roughly separated source signals by FDICA are inserted before H-TDICA as a prewhitening processing (see Fig. 2). After TDICA, the dewhitening is also performed. The stability of the learning in TDICA can be guaranteed by the holonomic constraint, and it is still possible to separate the temporally correlated signals because the pre/dewhitening processing prevents the ICA from performing the decorrelation. The detailed process using the proposed algorithm is as follows.

**[STEP 1. FDICA]**
FDICA is performed to separate sound sources to some extent. For example, the typical separation performance in FDICA is 9.4 dB under the condition that the reverberation time is 300 ms [4, 5]. Also, the mel cepstral distortion between the observed signal with the single source component at the microphone and the output signals from FDICA is about 2.5 dB. The separation filter of FDICA has spectrally flat characteristics in the direction of each sound source [5]. From this, we can estimate the approximate spectra of the sources blindly.

**[STEP 2. Prewhitening by Linear Prediction]**
In the linear prediction, the auto-regressive model of the generation process of the output signals from FDICA is given as

$$z_l(t) = -\sum_{n=1}^{N} p_l(n) z_l(t-n) + e_l(t) \quad (l = 1, \cdots, L), \qquad (3)$$

where $p_l(n)$ is a linear prediction coefficient for the $l$-th input signal, $e_l(t)$ is the input signal of this model, and $N$ is the order of the linear prediction coefficient. The linear prediction coefficient is obtained by calculating the Yule-Walker's simultaneous equations. The whitened signal $e_l(t)$ is obtained by convolving the linear prediction coefficient $p_l(n)$ with $z_l(t)$ as $e_l(t) = \sum_{n=0}^{N} p_l(n)$ $z_l(t-n)$, where $p_l(0) = 0$.

**[STEP 3. Holonomic TDICA]**

H-TDICA is performed with whitened signals. The output signals of H-TDICA can be given as $\boldsymbol{v}(t) = \sum_{\tau=0}^{Q-1} \boldsymbol{w}^{(\mathrm{H})}(\tau)\, \boldsymbol{e}(t-\tau)$, where $\boldsymbol{v}(t) = [v_1(t), \cdots, v_L(t)]^{\mathrm{T}}$ is the separated signal vector of H-TDICA, and $\boldsymbol{e}(t) = [e_1(t), \cdots, e_L(t)]^{\mathrm{T}}$ is the input signal vector whitened by the linear prediction for the H-TDICA part in MSICA. We optimize $\boldsymbol{w}^{(\mathrm{H})}(\tau)$ by the following H-TDICA:

$$\begin{aligned}\boldsymbol{w}_{i+1}^{(\mathrm{H})}(\tau) &= \boldsymbol{w}_{i}^{(\mathrm{H})}(\tau) + \alpha \sum_{d=0}^{Q-1} \Big\{ \boldsymbol{I}\delta(\tau-d) \\ &\quad - \langle \boldsymbol{\phi}(\boldsymbol{v}(t))\boldsymbol{v}(t-\tau+d)^{\mathrm{T}} \rangle_t \Big\} \boldsymbol{w}_{i}^{(\mathrm{H})}(d). \qquad (4)\end{aligned}$$

**[STEP 4. Dewhitening]**

The dewhitening process is performed by using the linear prediction coefficients $p_l(n)$ obtained in STEP 2. The resultant separated signals $y_l(t)$ can be obtained by the following IIR filtering:

$$y_l(t) = -\sum_{n=1}^{N} p_l(n) y_l(t-n) + v_l(t) \ \ (l = 1, \cdots, L). \qquad (5)$$

Note that the stability of the filtering is guaranteed because $p_l(n)$ is calculated from Levinson-Durbun's algorithm .

## 5. EXPERIMENTS AND RESULTS

### 5.1. Experimental Setup

A two-element array with the interelement spacing of 4 cm is assumed. The speech signals are assumed to arrive from two directions, $-30°$ and $40°$. The distance between the microphone array and the loudspeakers is 1.15 m. Two kinds of sentences, those spoken by two male and two female speakers are used as the original speech samples. The sampling frequency is 8 kHz and the length of speech is limited to within 3 seconds. Using these sentences, we obtain 12 combinations with respect to speakers and source directions. In these experiments, we use the following signals as the source signals: the original speech convolved with the impulse responses specified by the reverberation times of 300 ms. The impulse responses are recorded in a variable reverberation time room. In order to evaluate the performance, we used the *noise reduction rate* (NRR), defined as the output signal-to-noise ratio (SNR) in dB minus input SNR in dB. Also, in order to compare the various ICAs fairly, we perform postprocessing for the spectral compensation of the separated signals in H-TDICA. This processing is based on the utilization of the inverse of the separation filter matrix for the normalization of gain [3].

### 5.2. Experimental Results and Discussion

In this study, we compare the following MSICAs: **MSICA1:** FDICA is followed by NH-TDICA, **MSICA2:** FDICA is followed by H-TDICA, **MSICA3:** FDICA is followed by H-TDICA with spectral compensation, and **MSICA4:** FDICA is followed by the proposed method combining H-TDICA and linear prediction. The length of the separation filters, $\boldsymbol{w}^{(\mathrm{H})}(\tau)$ or $\boldsymbol{w}^{(\mathrm{NH})}(\tau)$, is 2048. In the proposed algorithm, the order $N$ in the linear predictor is 1024.

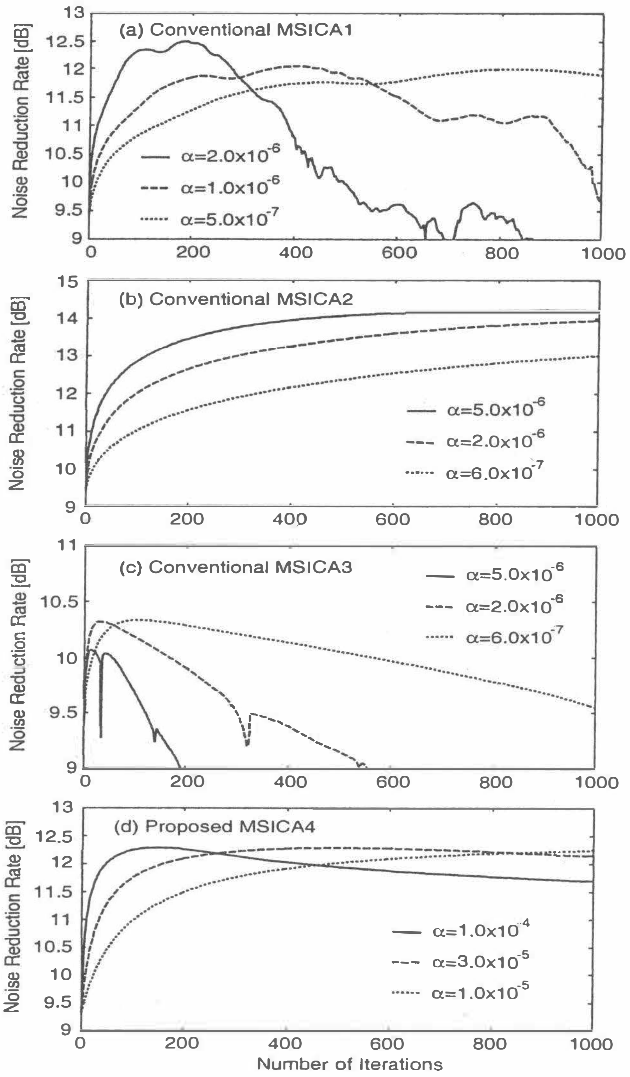

**Fig. 3**. Comparison of the noise reduction rates in (a) conventional MSICA1, (b) conventional MSICA2, (c) conventional MSICA3, and (d) proposed MSICA4.

Figures 3(a)–(d) show the NRR results of MSICA1–MSICA4 for different iteration points. These values were averages of all of the combinations with respect to speakers and source directions. The step-size parameters are chosen independently for each of the NH-TDICA, H-TDICA, and the proposed algorithm so that the NRR scores at the early iterations are almost the same in Figs. 3(a)–(d). From these results, the following are revealed. (1) In the conventional MSICA1 in which the NH-TDICA is used, the behavior of the NRR is not monotonic and there are remarkably consistent deteriorations, even when the step-size parameter is changed. (2) In the proposed algorithm, MSICA4, there are no deteriorations of NRRs. Therefore, the separation performances are almost completely retained during all of the iterations.

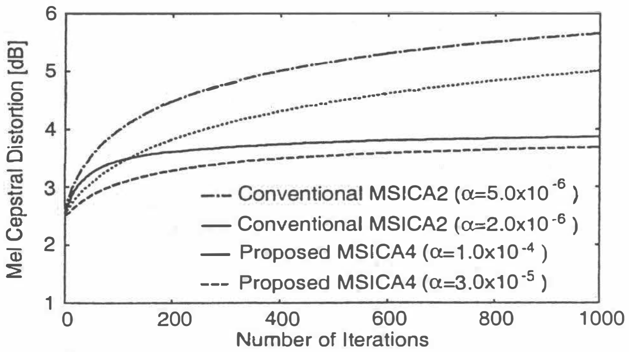

**Fig. 4**. Comparison of the mel cepstral distortions between the observed signal with the single source component at the microphone and the output signals from (a) conventional MSICA2 or (b) proposed MSICA4.

Regarding the separation performance of MSICA2 and MSICA3 in which the H-TDICA is used, the following are revealed. (1) The separation performance of MSICA2 is obviously superior to that of the proposed MSICA4. (2) However, its effective separation performance, i.e., the performance of MSICA3, is inferior to that of MSICA4. We speculate that the *specious* performance in MSICA2 is due to the exceeding emphasis of high-frequency components by the whitening effect of H-TDICA. Figure 4 shows the mel cepstral distortion between the observed signal with the single source component at the microphone and the output signals from (a) conventional MSICA2 or (b) proposed MSICA4. From these results, we can confirm the spectral distortion in MSICA2. In general, the separation in the high-frequency region is easier than that in low-frequency region [7] because the reverberation is shorter as the frequency increases. Thus, MSICA2 gains the improvement of the NRR only in the high-frequency region, and consequently we can conclude that MSICA2 is useless for separating the speech signals from the practical viewpoint. On the other hand, the distortions of the output signals from proposed MSICA4 is lower than those of MSICA4.

In order to confirm the convergence of each MSICA learning, we evaluate the frobenius norms of $\{\cdot\}$ parts on the right-hand side in Eqs. (2) and (4) . Figures 5(a) and (b) show $FN^{(\mathrm{NH})}$ of the conventional MSICA1 and $FN^{(\mathrm{H})}$ of the proposed MSICA4. These scores correspond to the stability of the iterative learning; it should be monotonically decreased. As shown in these figures, the conventional ICA loses its stability under the nonholonomic constraint. However, the proposed method can converge in every situation and consequently, we can conclude that the proposed algorithm is effective for improving the stability of the learning.

## 6. CONCLUSION

We newly proposed a stable algorithm for BSS combining MSICA and linear prediction. In the proposed algorithm, the linear predictors estimated from the roughly separated signals by FDICA are inserted before the holonomic TDICA as a prewhitening processing, and the dewhitening is performed after TDICA. The stability of the proposed algorithm can be guaranteed by the holonomic constraint, and the pre/dewhitening processing prevents the decorrelation. The experimental results under a reverberant condition revealed that the proposed algorithm results in the higher stability and higher separation performance, compared with the conventional MSICA including H-TDICA or NH-TDICA.

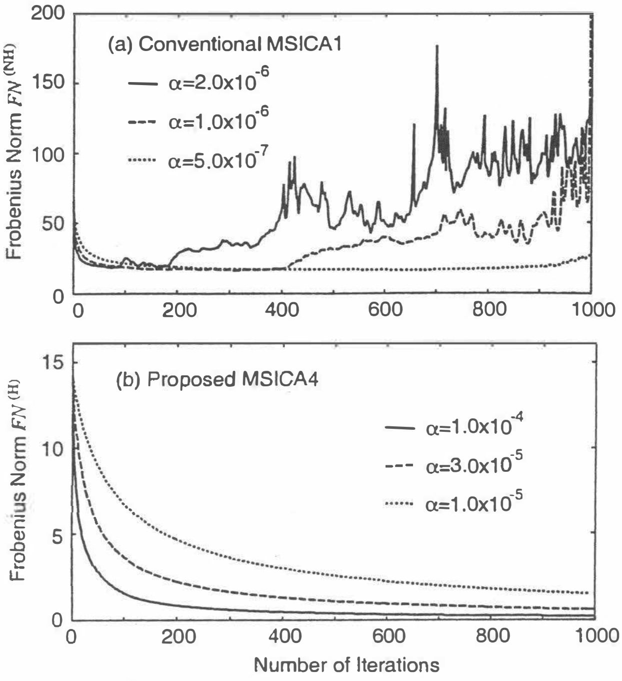

**Fig. 5**. Comparison of frobenius norms of $\{\cdot\}$ in iterative equation of (a) NH-TDICA part in conventional MSICA1 and (b) H-TDICA part in proposed MSICA4.

## 7. ACKNOWLEDGEMENT

The authors are grateful to Dr. Shoji Makino and Ms. Shoko Araki of NTT. CO., LTD for their greateful discussions. This work was partly supported by CREST (Core Research for Evolutional Science and Technology) of JST (Japan Science and Technology Corporation).

## 8. REFERENCES

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

[2] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," *Proc. SPAWC97*, pp.101–104, April 1997.

[3] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," *Proc. of 1998 International Symposium on Nonlinear Theory and Its Application (NOLTA98)*, pp.923–926, Sept. 1998.

[4] T. Nishikawa, H. Saruwatari, and K. Shikano, "Comparison of time-domain ICA, frequency-domain ICA and multistage ICA," *Proc. EUSIPCO2002*, vol.II, pp.15–18, Sept. 2002.

[5] H. Saruwatari, T. Kawamura, and K. Shikano, "Blind source separation for speech based on fast-convergence algorithm with ICA and beamforming," *Proc. Eurospeech2001*, pp. 2603–2606, Sept. 2001.

[6] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," *Proc. ICA99*, pp.371–376, January 1999.

[7] R. Aichner, S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Time domain ICA blind source separation of non-stationary convolved signals by utilizing geometric beamforming," *Proc. IEEE International Workshop on Neural Networks for Signal Processing*, pp.445–454, Sept. 2002.