REAL-TIME IMPLEMENTATION OF TWO-STAGE BLIND SOURCE SEPARATION COMBINING SIMO-ICA AND BINARY MASKING

[1]*Y. Mori,* [1]*H. Saruwatari,* [1]*T. Takatani,* [1]*S. Ukai,* [1]*K. Shikano,* [2]*T. Hiekata* [2]*T. Morita*

[1]{yoshim-m, sawatari, tomoya-t, shikano}@is.naist.jp
[1]Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara, 630-0192, Japan
[2]Kobe Steel,Ltd., Kobe, 651-2271, Japan

ABSTRACT

A new real-time two-stage blind source separation (BSS) for convolutive mixtures of speech is proposed, in which a Single-Input Multiple-Output (SIMO)-model-based ICA and binary mask processing are combined. SIMO-model-based ICA can separate the mixed signals, not into monaural source signals but into SIMO-model-based signals from independent sources as they are at the microphones. Thus, the separated signals of SIMO-model-based ICA can maintain the spatial qualities of each sound source. Owing to the attractive property, binary mask processing can be applied to efficiently remove the residual interference components after SIMO-model-based ICA. In addition, the performance deterioration due to the latency problem in ICA can be mitigated by introducing real-time binary mask processing. The experiments using real-time BSS system reveal that the separation performance can be considerably improved by using the proposed method in comparison to the conventional BSS methods.

1. INTRODUCTION

Blind source separation (BSS) is the approach taken to estimate original source signals using only the information of the mixed signals observed in each input channel. In recent works of BSS based on independent component analysis (ICA), various methods have been proposed for acoustic-sound separation [1, 2, 3]. However the separation performance of the existing method heavily degrades especially under highly reverberant conditions. Therefore the development of high-accuracy BSS in a real-world application is a problem demanding prompt attention.

In order to address the problem, we have recently proposed a novel two-stage BSS algorithm [4] which combines (a) a Single-Input Multiple-Output (SIMO)-model-based ICA (SIMO-ICA) [5] and (b) time-frequency domain binary mask processing [6, 7, 8] applied to the SIMO-ICA's outputs. SIMO-model-based ICA can decompose the mixed signals into SIMO-model-based signals from independent sources as they are at the sensors. After the SIMO-model-based ICA, the residual components of the interference can be efficiently removed by the following binary mask processing.

It should be enhanced that the two-stage method has another important property, i.e., applicability to the *real-time* processing. In general ICA-based BSS methods require huge calculations, but binary mask processing needs very few computational complexities. Therefore, because of the introduction of binary masking into ICA, the proposed combination can function as the real-time system. In this paper, we mainly discuss the real-time implementation issue on the proposed BSS, and evaluate the "real-time" separation performance for speech mixtures under a real reverberant condition.

2. MIXING PROCESS

The number of microphones is K and the number of multiple sound sources is L, where we deal with the case of $K = L$ in this study. On the basis of the time-frequency domain signal representation, we designate the observed time series as $\boldsymbol{X}(f,t)$ $=[X_1(f,t), \cdots, X_K(f,t)]^{\mathrm{T}}$. The observed signals in which multiple source signals are mixed are given by

$$\boldsymbol{X}(f,t) = \boldsymbol{A}(f)\boldsymbol{S}(f,t), \tag{1}$$

where $\boldsymbol{S}(f,t)$ = $[S_1(f,t), \cdots, S_L(f,t)]^{\mathrm{T}}$ is the source signal vector. Also, $\boldsymbol{A}(f) = [A_{kl}(f)]_{kl}$ is the mixing matrix, where $[X]_{ij}$ denotes the matrix which includes the element X in the i-th row and the j-th column. The mixing matrix $\boldsymbol{A}(f)$ is assumed to be complex-valued in a convolutive mixture model which arises in real audio applications.

In general, the observed signal can be represented as a superposition of the SIMO-model-based signals as follows:

$$\begin{aligned}\boldsymbol{X}(f,t) = {} & [A_{11}(f)S_1(f,t), \cdots, A_{K1}(f)S_1(f,t)]^{\mathrm{T}} + \cdots \\ & +[A_{1L}(f)S_L(f,t), \cdots, A_{KL}(f)S_L(f,t)]^{\mathrm{T}}, \end{aligned} \tag{2}$$

where $[A_{1l}(f)S_l(f,t), \cdots, A_{Kl}(f)S_l(f,t)]^{\mathrm{T}}$ is a vector which corresponds to SIMO-model-based signals with respect to the l-th sound source; the k-th element corresponds to the k-th microphone's signal.

3. PROPOSED TWO-STAGE BSS

3.1. Overview

In the previous research, SIMO-ICA was proposed by, e.g., Takatani et al. [5], and they showed that SIMO-ICA can separate the mixed signals, not into monaural source signals, but into SIMO-model-based signals at the microphone points. This finding has motivated us to combine the SIMO-model-based ICA and binary mask processing. That is, the binary mask technique can be applied to the SIMO components of each source obtained from SIMO-ICA. Needless to say, the obtained SIMO components is well applicable to binary mask processing because of the spatial properties that the separated SIMO component at the specific microphone closer to the target sound still maintains the large gain.

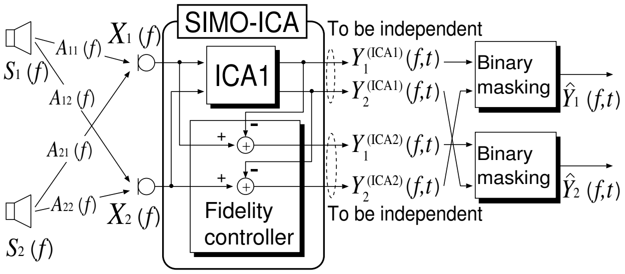

Figure 1: Input and output relations in the proposed two-stage BSS, where $K = L = 2$.

The configuration of the proposed method is depicted in Fig. 1. Binary mask processing which follows SIMO-ICA can remove the residual component of the interference effectively without adding huge computational complexities.

3.2. Algorithm

We first conduct a frequency-domain SIMO-ICA (FD-SIMO-ICA) for extracting the SIMO-model-based signals corresponding to each of sources. The FD-SIMO-ICA consists of $(L-1)$ FDICA parts and a *fidelity controller*, and each ICA runs in parallel under the fidelity control of the entire separation system (see Fig. 1). The separated signals of the l-th ICA ($l = 1, \cdots L-1$) in FD-SIMO-ICA are defined by

$$\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t) = [Y_k^{(\mathrm{ICA}l)}(f,t)]_{k1} = \boldsymbol{W}_{(\mathrm{ICA}l)}(f)\boldsymbol{X}(f,t), \quad (3)$$

where $\boldsymbol{W}_{(\mathrm{ICA}l)}(f) = [W_{ij}^{(\mathrm{ICA}l)}(f)]_{ij}$ is the separation filter matrix in the l-th ICA.

Regarding the fidelity controller, we calculate the following signal vector $\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t)$, in which the all elements are to be mutually independent,

$$\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t) \;=\; \boldsymbol{X}(f,t) - \sum_{l=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t). \quad (4)$$

Hereafter, we regard $\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t)$ as an output of a *virtual* "L-th" ICA. The reason we use the word "*virtual*" here is that the L-th ICA does not have own separation filters unlike the other ICAs, and $\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t)$ is subject to $\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ ($l = 1, \cdots, L-1$). Transposing the 2nd term $(-\sum_{l=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t))$ in the right-hand side into the left-hand side, we can show that (4) means a constraint to force the sum of all ICAs' output vectors $\sum_{l=1}^{L} \boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t)$ to be the sum of all SIMO components $[\sum_{l=1}^{L} A_{kl}(f)S_l(f,t)]_{k1}$ $(= \boldsymbol{X}(f,t))$.

If the independent sound sources are separated by (3), and simultaneously the signals obtained by (4) are also mutually independent, then the output signals converge on unique solutions, up to the permutation, as

$$\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t) = \mathrm{diag}\big[\boldsymbol{A}(f)\boldsymbol{P}_l^{\mathrm{T}}\big]\boldsymbol{P}_l\boldsymbol{S}(f,t), \quad (5)$$

where $\boldsymbol{P}_l$ ($l = 1, \cdots, L$) are exclusively-selected permutation matrices which satisfy $\sum_{l=1}^{L} \boldsymbol{P}_l = [1]_{ij}$. Regarding a proof of this, see [5] with an appropriate modification into the frequency-domain representation. Obviously the solutions given by (5) provide necessary and sufficient SIMO components, $A_{kl}(f)S_l(f,t)$, for each l-th source. Thus, the separated signals of SIMO-ICA can maintain the spatial qualities of each sound source. For example in the case of $L = K = 2$, one possibility is given by

$$\begin{aligned}
&\big[Y_1^{(\mathrm{ICA1})}(f,t),\ Y_2^{(\mathrm{ICA1})}(f,t)\big]^{\mathrm{T}} \\
&\quad = \big[A_{11}(f)S_1(f,t),\ A_{22}(f)S_2(f,t)\big]^{\mathrm{T}}, \quad (6) \\
&\big[Y_1^{(\mathrm{ICA2})}(f,t),\ Y_2^{(\mathrm{ICA2})}(f,t)\big]^{\mathrm{T}} \\
&\quad = \big[A_{12}(f)S_2(f,t),\ A_{21}(f)S_1(f,t)\big]^{\mathrm{T}}, \quad (7)
\end{aligned}$$

where $\boldsymbol{P}_1 = \boldsymbol{I}$ and $\boldsymbol{P}_2 = [1]_{ij} - \boldsymbol{I}$.

In order to obtain (5), the natural gradient of Kullback-Leibler divergence of (4) with respect to $\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ should be added to the existing nonholonomic iterative learning rule [1] of the separation filter in the l-th ICA ($l = 1, \cdots, L-1$). The new iterative algorithm of the l-th ICA part ($l = 1, \cdots, L-1$) in FD-SIMO-ICA is given as

$$\begin{aligned}
&\boldsymbol{W}_{(\mathrm{ICA}l)}^{[j+1]}(f) \\
&= \boldsymbol{W}_{(\mathrm{ICA}l)}^{[j]}(f) - \alpha \Bigg[\Bigg\{ \text{off-diag} \left\langle \boldsymbol{\Phi}\big(\boldsymbol{Y}_{(\mathrm{ICA}l)}^{[j]}(f,t)\big) \right. \\
&\quad \left. \boldsymbol{Y}_{(\mathrm{ICA}l)}^{[j]}(f,t)^{\mathrm{H}} \right\rangle_t \Bigg\} \cdot \boldsymbol{W}_{(\mathrm{ICA}l)}^{[j]}(f) \\
&\quad - \Bigg\{ \text{off-diag} \left\langle \boldsymbol{\Phi}\Big(\boldsymbol{X}(f,t) - \sum_{l=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l)}^{[j]}(f,t)\Big) \right. \\
&\quad \left. \cdot \Big(\boldsymbol{X}(f,t) - \sum_{l=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l)}^{[j]}(f,t)\Big)^{\mathrm{H}} \right\rangle_t \Bigg\} \\
&\quad \cdot \Big(\boldsymbol{I} - \sum_{l=1}^{L-1} \boldsymbol{W}_{(\mathrm{ICA}l)}^{[j]}(f)\Big)\Bigg], \quad (8)
\end{aligned}$$

where α is the step-size parameter, and we define the nonlinear vector function $\boldsymbol{\Phi}(\cdot)$ as [9]:

$$\boldsymbol{\Phi}(\boldsymbol{Y}(f,t)) \;\equiv\; \big[\tanh(|Y_l(f,t)|)e^{\jmath \cdot \arg(Y_l(f,t))}\big]_{l1}. \quad (9)$$

Also, the initial values of $\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ for all l should be different.

After FD-SIMO-ICA, binary masking processing is applied. For example in the case of (6) and (7), the resultant output signal corresponding to the source 1 is obtained as follows:

$$\hat{Y}_1(f,t) \;=\; m_1(f,t)Y_1^{(\mathrm{ICA1})}(f,t), \quad (10)$$

where $m_1(f,t)$ is the binary mask operation which is defined as $m_1(f,t) = 1$ if $Y_1^{(\mathrm{ICA1})}(f,t)$ is greater than $Y_2^{(\mathrm{ICA2})}(f,t)$; otherwise $m_1(f,t) = 0$. Also, the resultant output signal corresponding to the source 2 is given by

$$\hat{Y}_2(f,t) \;=\; m_2(f,t)Y_2^{(\mathrm{ICA1})}(f,t), \quad (11)$$

where $m_2(f,t)$ is the binary mask operation which is defined as $m_2(f,t) = 1$ if $Y_2^{(\mathrm{ICA1})}(f,t)$ is greater than $Y_1^{(\mathrm{ICA2})}(f,t)$; otherwise $m_2(f,t) = 0$. The extension to the general case of $L = K > 2$ can be easily implemented in the same manner.

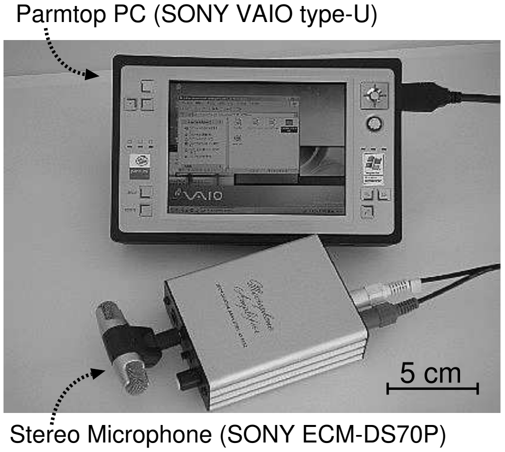
Parmtop PC (SONY VAIO type-U)
5 cm
Stereo Microphone (SONY ECM-DS70P)

Figure 2: Overview of real-time two-stage BSS system.

3.3. Real-Time Implementation

We have already built a real-time two-stage BSS system using a small stereo microphone (SONY ECM-DS70P) and a very light palmtop PC (SONY VAIO type-U with Pentium-M 1.1 GHz processor, 550 g weight) as shown in Fig. 2. Figure 3 shows a configuration of a real-time implementation for the proposed two-stage BSS. Signal processing in this implementation is performed as the following instructions.

1. Inputted signals are converted to time-frequency series by using frame-by-frame fast Fourier transform (FFT).
2. SIMO-ICA is conducted using a current 3 s-duration data for estimating the separation matrix which is applied to the next (*not current*) 3 s samples. This staggered relation is due to the fact that the filter update in SIMO-ICA requires huge computational complexities and cannot provide the optimal separation filter for the current 3 s data.
3. Binary mask processing is applied to the separated signals obtained by the previous SIMO-ICA. Unlike SIMO-ICA, binary masking can be conducted just in the current segment.
4. The output signals from binary mask processing are converted to the resultant time-domain waveforms by using an inverse FFT.

Although the separation filter update in SIMO-ICA part is not real-time processing but includes a 3 s latency, the whole two-stage system still seems real-time because the binary masking can work in the current segment with no delay. Generally the latency in the conventional ICAs is problematic and reduces the applicability of the methods to real-time systems. In the proposed method, however, the performance deterioration due to the latency problem in SIMO-ICA can be mitigated by introducing real-time binary mask processing.

4. EXPERIMENTS IN REAL-TIME APPLICATION

4.1. Conditions for Experiments

We carried out real-time sound-separation experiments using acoustical source signals recorded in the real room illustrated in Fig. 4,

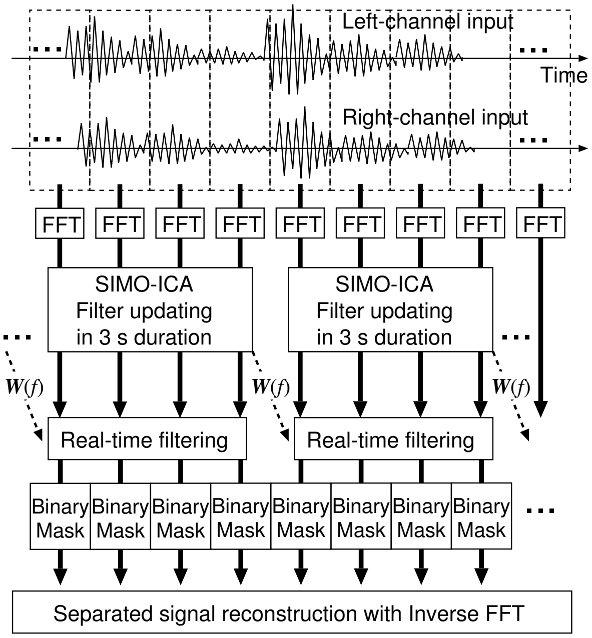

Figure 3: Signal flow in real-time implementation of proposed method.

where two sources and the real-time BSS system (Fig. 2) are set. The reverberation time in this room is 200 ms. Two acoustic signals are assumed to arrive from different directions, θ_1 and θ_2, where we prepare two kinds of source direction patterns as follows; $(\theta_1, \theta_2) = (-60^\circ, 60^\circ)$, or $(-60^\circ, 0^\circ)$. We used the speech signals spoken by two male and two female speakers as the source samples, and generated 12 speaker combinations. The sampling frequency is 8 kHz and the length of each speech sample is limited to 6 seconds. The DFT size of $\boldsymbol{W}(f)$ in each method is 1024. We use an initial value which is given by null beamformers [3] whose directions of sources are $(-40^\circ, 40^\circ)$.

4.2. Experimental Results

We compare four methods as follows: (A) the conventional binary-mask-based BSS, (B) the conventional ICA-based BSS [9], (C) simple combination of the conventional ICA and binary mask processing, and (D) the proposed two-stage BSS method. *Noise reduction rate* (NRR) [3], 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. The SNRs are calculated under the assumption that the speech signal of the undesired speaker is regarded as noise.

Figure 5 show an example of the segmental NRR which was calculated along the time axis at every 100 ms period. The first 3 s duration is spent on the initial filter learning of ICA in the methods (B), (C) and (D), and thus the valid ICA-based separation filter is absent here. Therefore, at 0.0 s–3.0 s, we simply applied binary mask processing in the methods (C) and (D). The successive 3 s duration (at 3.0 s–6.0 s) shows the separation results for *open* data sample, which is to be evaluated in this experiment. From Fig. 5, we can confirm that the proposed two-stage BSS (D) outperforms other methods at almost all the time during

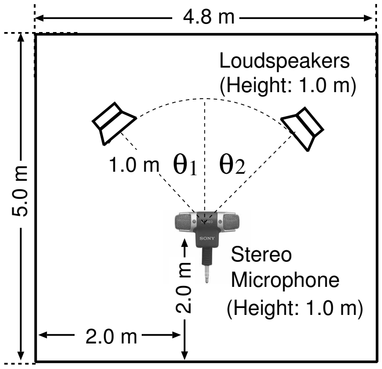

Figure 4: Layout of reverberant room used in experiments.

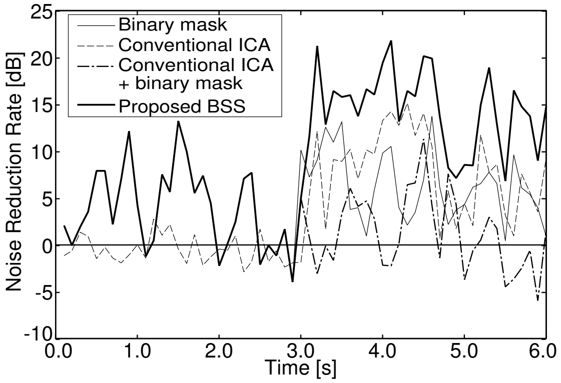

Figure 5: Example of segmental NRR for male-female separation at every 100 ms period, where $(\theta_1, \theta_2) = (-60^\circ, 0^\circ)$.

3.0 s–6.0 s.
In Figure 6, we show the time-averaged NRRs in 3.0 s–6.0 s for different speaker combinations and different source-direction patterns. As can be seen, the proposed two-stage BSS can improve the separation performance regardless of the speaker combinations as well as source directions, and the proposed BSS outperforms all of the conventional methods. It is worth noting that the simple combination of the conventional ICA and binary mask processing shows heavy deteriorations, and this method is *not* beneficial to the improvement. These facts are promising evidences on the feasibility of the proposed combination technique of SIMO-model-based ICA and binary mask processing.

5. CONCLUSION

We proposed a new BSS framework in which the SIMO-model-based ICA and binary mask processing are efficiently combined. Also we introduced the real-time implementation of the proposed method. In order to evaluate its effectiveness, a real-time BSS experiment was carried out under a reverberant condition. The experimental results revealed that the SNR can be considerably improved by using the proposed two-stage BSS algorithm. In addition, we could find the fact that the proposed method outperforms the simple ICA and binary mask processing.

6. ACKNOWLEDGMENT

This work was partly supported by CREST "Advanced Media Technology for Everyday Living" of JST in Japan.

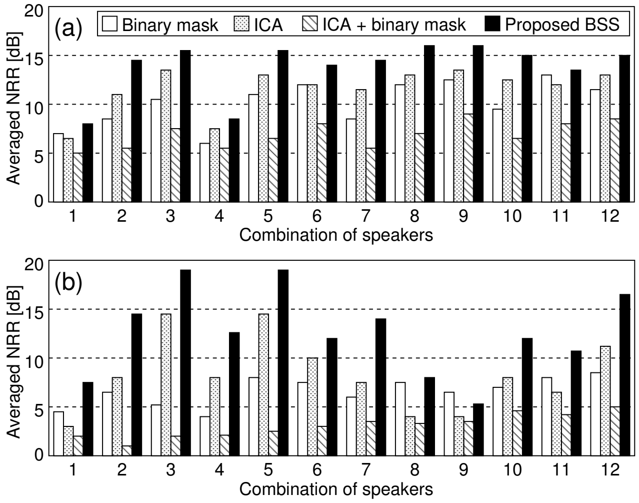

Figure 6: Time-averaged NRRs in 3.0 s–6.0 s for different speaker combinations under (a) $(\theta_1, \theta_2) = (-60^\circ, 60^\circ)$, and (b) $(\theta_1, \theta_2) = (-60^\circ, 0^\circ)$.

7. REFERENCES

[1] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," *Proc. NOLTA98*, vol.3, pp.923–926, 1998.

[2] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech & Audio Processing*, vol.8, pp.320–327, 2000.

[3] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Sig. Process.*, vol.2003, pp.1135–1146, 2003.

[4] H. Saruwatari, Y. Mori, S. Ukai, T. Takatani, K. Shikano, T. Hiekata and T. Morita, "Two-stage blind source separation using SIMO-ICA and binary masking," *Proc. HSCMA2005*, pp.d-11–d-12, 2005.

[5] T. Takatani, T. Nishikawa, H. Saruwatari and K. Shikano, "High-fidelity blind separation of acoustic signals using SIMO-model-based ICA with information-geometric learning," *Proc. IWAENC2003*, pp.251–254, 2003.

[6] R. Lyon, "A computational model of binaural localization and separation," *Proc. ICASSP83*, pp.1148–1151, 1983.

[7] N. Roman, D. Wang and G. Brown, "Speech segregation based on sound localization," *Proc. IJCNN01*, pp.2861–2866, 2001.

[8] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol.22, no.2, pp.149–157, 2001.

[9] H. Sawada, R. Mukai, S. Araki and S. Makino, "Polar coordinate based nonlinear function for frequency domain blind source separation," *IEICE Trans. Fundamentals*, vol.E86-A, no.3, pp.590–596, 2003.