UNDERDETERMINED BLIND SPEECH SEPARATION WITH DIRECTIVITY PATTERN BASED CONTINUOUS MASK AND ICA

Shoko Araki *Shoji Makino* *Hiroshi Sawada* *Ryo Mukai*
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: shoko@cslab.kecl.ntt.co.jp

ABSTRACT

We propose a method for separating speech signals when sources outnumber the sensors. In this paper we mainly concentrate on the case of three sources and two sensors. Some existing methods employ binary masks to extract the signals, and therefore, the extracted signals contain loud musical noise. To overcome this problem, we propose the utilization of a directivity pattern based continuous mask, which removes a single source from the observations, and independent component analysis (ICA) to separate the remaining mixtures. Experimental results show that our proposed method can separate signals with little distortion even in a real reverberant environment of T_R=130 ms.

1. INTRODUCTION

Blind source separation (BSS) is an approach that estimates original source signals $s_i(n)$ only from observations $x_j(n)$ without source or mixing process information.

In this paper, we consider the BSS of speech signals observed in a real environment, i.e., the BSS of convolutive mixtures of speech. Recently, many methods have been proposed to solve the BSS problem of convolutive mixtures (e.g., [1]). However, most of these methods consider the determined or overdetermined case. In contrast, we focus on the underdetermined BSS problem where source signals outnumber the sensors, especially the case of three sources and two sensors.

There are two approaches with which to realize underdetermined BSS. Both approaches rely on the sparseness of source signals. One involves the clustering of time-frequency points with binary masks [2], and the other is based on ML estimation, where the sources are estimated after mixing matrix estimation [3–5]. Since separation in a real environment has already been tried with the former method, we decided to watch a binary mask approach [2]. If the signals are sufficiently sparse, that is, most of the samples of a signal are almost zero, we can assume that the sources rarely overlap. Rickard and Yilmaz [2] employ this assumption and extract each signal using a time-frequency binary mask (BM). However, the use of binary masks means that there is too much discontinuous zero-padding to the extracted signals, and they contain loud musical noise.

To overcome this, we have proposed combining binary masks and ICA (BMICA) to solve the underdetermined BSS problem [6]. In the paper, we estimate the time points when only one source is active using the sparseness assumption. Then, we remove this single source from the observations with a binary mask and apply ICA to the remaining mixtures to obtain separated signals. The single source removal with this method causes less discontinuous zero-padding than with the BM method, because the BMICA extracts more time-frequency points than the BM method. Therefore, we have been able to use the BMICA to obtain separated signals with little distortion. However, as the BMICA still employs a binary mask for one source removal, the zero-padding to the separated signals still remains. Moreover, heuristic parameters are used when designing the binary mask shapes not only for the BMICA but also for the BM method.

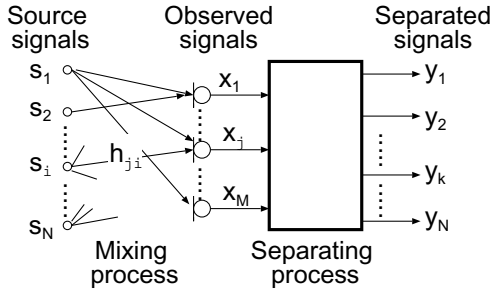

Figure 1: Block diagram of underdetermined BSS. $N > M$.

In this paper, we propose a new method for one source removal that employs a directivity pattern based continuous mask (DCmask). First, using the source sparseness, we estimate the direction of arrival (DOA) of each source. We then remove the single source from the observations with a DCmask, which has a small gain for the DOA of one source and a large gain for other directions. We next apply ICA to the remaining mixtures. Since this DCmask is not binary, discontinuous zero-padding to the extracted signals does not occur. Moreover, we do not need any parameters for designing the DCmask.

Experimental results show that our method can separate signals with little distortion even in a real reverberant environment of T_R=130 ms.

2. PROBLEM DESCRIPTION

In real environments, N signals observed by M sensors are modeled as convolutive mixtures $x_j(n) = \sum_{i=1}^{N} \sum_{l=1}^{L} h_{ji}(l)\ s_i(n-l+1)$ $(j = 1, \cdots, M)$, where s_i is the signal from a source i, x_j is the signal observed by a sensor j, and h_{ji} is the L-taps impulse response from a source i to a sensor j (see Fig. 1). Our objective is to obtain separated signals $y_k(n)$ $(k = 1, \cdots, N)$ using only the information provided by observations $x_j(n)$. Here, we consider the underdetermined case $N > M$. In this paper $N = 3$ and $M = 2$. Moreover, the sources are speech signals, i.e., the sources are assumed to be mutually independent and sufficiently sparse in the time-frequency domain.

This paper employs a time-frequency domain approach because speech signals are more sparse in the time-frequency domain than in the time-domain [5] and convolutive mixture prob-

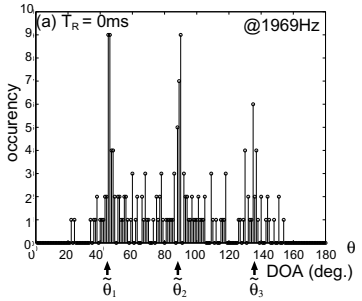

Figure 2: Example histogram. A male-male-female combination with STFT frame size $T = 512$. $T_R = 0$ ms.

lems can be converted into instantaneous mixture problems at each frequency. In the time-frequency domain, mixtures are modeled as $\mathbf{X}(\omega,m) = \mathbf{H}(\omega)\mathbf{S}(\omega,m)$, where $\mathbf{H}(\omega)$ is a 2×3 mixing matrix whose *j-i* component is a transfer function from a source *i* to a sensor *j*, $\mathbf{S}(\omega,m) = [S_1(\omega,m), S_2(\omega,m), S_3(\omega,m)]^T$ and $\mathbf{X}(\omega,m) = [X_1(\omega,m), X_2(\omega,m)]^T$ denote short-time Fourier transformed source and observed signals, respectively. ω is the frequency and m is the time-dependence of the short-time Fourier transformation (STFT). Moreover, we write short-time Fourier transformed separated signals as $\mathbf{Y}(\omega,m) = [Y_1(\omega,m), Y_2(\omega,m), Y_3(\omega,m)]^T$.

3. CONVENTIONAL METHODS

Several methods have been proposed (e.g., [2–7]) for solving the underdetermined BSS problem, and they all utilize source sparseness. If most of the samples of a signal are almost zero, we say that this signal is sparse. For a detailed experimental analysis of sparseness, see [7].

3.1 Conventional method 1: with only binary masks (BM)

Some methods use the sparseness assumption and extract each signal using time-frequency binary masks (e.g.,[2]). When signals are sufficiently sparse, it can be assumed that sources do not overlap very often. This condition is closely discussed in [2]. Using this non-overlap assumption, they extract each source by selecting the time points at which there is only one signal. One way of estimating such time points is to use the level difference and the phase difference between the observations. In this paper, we utilize omni-directional microphones, therefore we use the phase difference $\varphi(\omega,m) = \angle \frac{X_1(\omega,m)}{X_2(\omega,m)}$ between the observations.

Using $\varphi(\omega,m)$, we estimate the direction of arrival (DOA) for each time point m by calculating $\theta(\omega,m) = \cos^{-1}\frac{\varphi(\omega,m)c}{\omega d}$, where c is the speed of sound and d is the microphone spacing, and draw a histogram of the DOA $\theta(\omega,m)$ (see Fig. 2). Each peak corresponds to each source in the histogram for each frequency. Let these peaks be $\tilde{\theta}_1, \tilde{\theta}_2$ and $\tilde{\theta}_3$ where $\tilde{\theta}_1 \le \tilde{\theta}_2 \le \tilde{\theta}_3$ (Fig. 2), and the signal from $\tilde{\theta}_r$ be $\tilde{S}_r$ ($r = 1,2,3$).

We can extract each signal with a binary mask

$$M^r_{\mathrm{BM}}(\omega,m) = \begin{cases} 1 & \tilde{\theta}_r - \Delta \le \theta(\omega,m) \le \tilde{\theta}_r + \Delta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

by calculating $Y_r(\omega,m) = M^r_{\mathrm{BM}}(\omega,m)X_j(\omega,m)$ where j=1 or 2. Here, Δ is an extraction range parameter.

Although we can extract each signal using this binary mask (1), such extracted signals are discontinuously zero-padded by the binary masks, and therefore, we hear musical noise in the outputs. Moreover, the performance depends on the heuristic parameter, Δ.

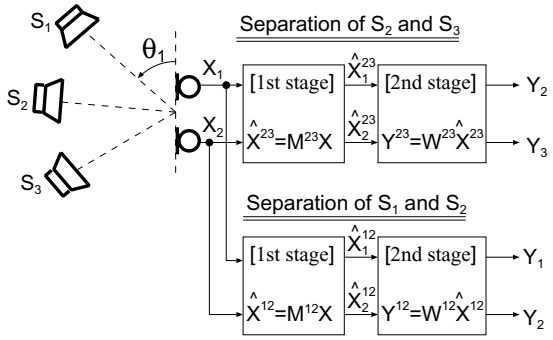

Figure 3: System setup

3.2 Conventional method 2: with binary masks and ICA (BMICA)

To overcome the musical noise problem, we have proposed using both a binary mask and ICA (BMICA) [6]. The BMICA has two stages (Fig. 3).

In the first stage, unlike the conventional method 1 where single source is extracted, we *remove* one source from the mixtures using a binary mask

$$M^{pq}_{\mathrm{BMICA}}(\omega,m) = \begin{cases} 1 & \theta_{min} \le \theta(\omega,m) \le \theta_{max} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

by calculating

$$\hat{\mathbf{X}}^{pq}(\omega,m) = M^{pq}_{\mathrm{BMICA}}(\omega,m)\mathbf{X}(\omega,m). \quad (3)$$

In (2), θ_{min} and θ_{max} are extraction range parameters that cover the direction range of two sources, and in (3), $\hat{\mathbf{X}}^{pq}(\omega,m) = [\hat{X}^{pq}_1(\omega,m), \hat{X}^{pq}_2(\omega,m)]$ are expected to be mixtures of $\tilde{S}_p$ and $\tilde{S}_q$. Then, as we can expect the remaining mixtures $\hat{\mathbf{X}}^{pq}$ to consist of only two signals, we apply a standard ICA to these remaining mixtures in the second stage.

In the 1st stage, we expect that the zero-padding of the extracted signals to cause less trouble because we extract more time-frequency points than conventional method 1. Therefore we obtained a result with less musical noise [6].

However, as BMICA still employed a binary mask for one source removal, the zero-padding of the separated signals still remained. Moreover, we have to find a reasonable θ_{min} and θ_{max} in (2). This is not an easy problem and we relied on manual setting.

4. PROPOSED METHOD

[1st stage] One source removal with new DC mask:

In this paper, we propose a new method (Fig. 4) which employs a directivity pattern based continuous mask (DCmask) in the 1st stage. The DCmask has a small gain for the DOA of one source and preserves signals from other directions. As shown in Fig. 2 and Sec. 3, DOA is a powerful clue with which to estimate the active source in each frame and to extract or remove one signal. Therefore the DCmask is a natural alternative. Because this new mask is not binary, it completely avoids any zero-padding to the extracted signals.

One way to obtain such a mask is to utilize the gain of the directivity pattern of a null beamformer (NBF) which makes a null toward the estimated direction $\tilde{\theta}_r$ of one source.

First we estimate the source directions with a histogram such as that in Fig. 2. Then, we form a matrix

$$\mathbf{H}_{\mathrm{NBF}}(\omega) = \begin{bmatrix} \exp(j\omega\tau_{11}) & \exp(j\omega\tau_{12}) \\ \exp(j\omega\tau_{21}) & \exp(j\omega\tau_{22}) \end{bmatrix}, \quad (4)$$

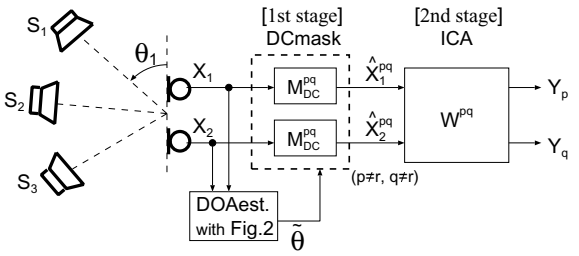

Figure 4: Block diagram of proposed method. Only one path is drawn for visibility.

where $\tau_{ji} = \frac{d_j}{c}\cos\tilde{\theta}_i$, d_j is the position of the j-th microphone and c is the speed of sound. Where, one of the $\tilde{\theta}_i$ values should be the estimated DOA $\tilde{\theta}_r$ of signal to be removed, and the other should be a different direction from which the signal's gain and phase are constrained at 1.

The directivity pattern of the NBF, $\mathbf{W}(\omega) = \mathbf{H}_{\mathrm{NBF}}^{-1}(\omega)$, is obtained by

$$F_r(\omega,\theta) = \sum_{k=1}^{M} W_{rk}(\omega)\exp(j\omega d_k\cos\theta/c). \quad (5)$$

Here, we use the gain of the directivity pattern of NBF as our mask,

$$M_{\mathrm{DC}}^{pq}(\omega,m) = |F_r(\omega,\theta(\omega,m))| \quad (p\neq r, q\neq r). \quad (6)$$

This is our new mask, the DCmask. Figure 5 shows an example of the shape of a DCmask.

Finally, the one source removal is achieved by

$$\hat{\mathbf{X}}^{pq}(\omega,m) = M_{\mathrm{DC}}^{pq}(\omega,m)\mathbf{X}(\omega,m). \quad (7)$$

It should be noted that the DCmask is applied to both channels (Fig. 4), because ICA in the 2nd stage needs two inputs that maintain the mixing matrix information. Here, our new mask $M_{\mathrm{DC}}^{pq}(\omega,m)$ changes only the gains of inputs X_1 and X_2 and preserves their phases. Therefore, we can preserve to some extent the spatial information on the sources, and a standard ICA can work in the 2nd stage.

[2nd stage] Separation of remaining sources by ICA:

Because the remaining signals $\hat{\mathbf{X}}^{pq}$ are expected to be mixtures of two signals, we can use 2×2 ICA to separate $\hat{\mathbf{X}}^{pq}$. The separation process can be formulated as

$$\mathbf{Y}^{pq}(\omega,m) = \mathbf{W}^{pq}(\omega)\hat{\mathbf{X}}^{pq}(\omega,m), \quad (8)$$

where $\hat{\mathbf{X}}^{pq}$ is the masked observed signal obtained by (7), $\mathbf{Y}^{pq}(\omega,m) = [Y_p(\omega,m), Y_q(\omega,m)]^T$ is the separated output signal, and $\mathbf{W}^{pq}(\omega)$ represents a (2×2) separation matrix. $\mathbf{W}^{pq}(\omega)$ is determined so that $Y_p(\omega,m)$ and $Y_q(\omega,m)$ become mutually independent by ICA.

Note that we need two paths to obtain three separated signals (see Fig. 3) because our system has only two outputs.

Although we consider the case of two sensors ($M=2$) and three sources ($N=3$) here, we can expand our DCmask method to the underdetermined case of $N \leq$ (the number of nulls formed by M sensors) + (the number of outputs of a standard ICA) $= (M-1)+M$.

5. EXPERIMENTS

5.1 Experimental conditions

We conducted anechoic tests and reverberant tests. For the anechoic tests (T_R = 0 ms), we simulated the recording in an anechoic room using the mixing matrix $H_{ji}(\omega) = \exp(j\omega\tau_{ji})$, where

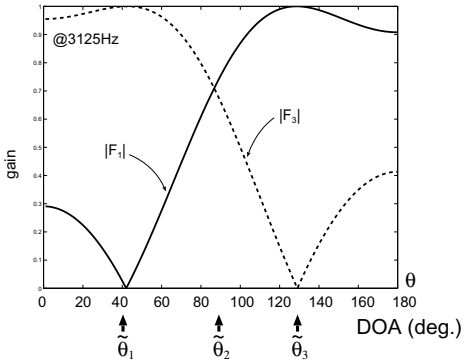

Figure 5: Example mask pattern.

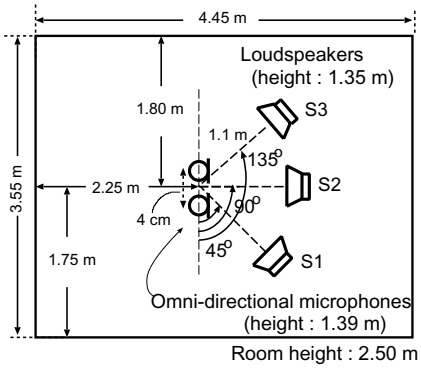

Figure 6: Room for reverberant tests. T_R = 130 ms.

$\tau_{ji} = \frac{d_i}{c}\cos\theta_i$, d_j is the position of the j-th microphone, and θ_i is the direction of the i-th source. The source directions were 45°, 90° and 135°.

For the reverberant tests, we used speech data convolved with impulse responses recorded in a real room (Fig. 6) whose reverberation time was T_R = 130 ms.

As the original speech, we used three Japanese sentences spoken by three male and three female speakers. We investigated three combinations of speakers: male-male-female (m-m-f), male-male-male (m-m-m), and female-female-female (f-f-f).

The STFT frame size T was 512 and the frame shift was 256 at a sampling rate of 8 kHz. We used $\theta_{min} = \tilde{\theta}_2 - \Delta$ and $\theta_{max} = 180°$ for M_{BMICA}^{23} (BMICA 23), and $\theta_{min} = 0°$ and $\theta_{max} = \tilde{\theta}_2 + \Delta$ for M_{BMICA}^{12} (BMICA 12). The Δ value in the conventional methods' binary masks was 15° in DOA.

The adaptation rule of ICA that we used in our experiments was $\mathbf{W}_{i+1}(\omega) = \mathbf{W}_i(\omega) + \eta\left[\mathrm{diag}\left(\langle\Phi(\mathbf{Y})\mathbf{Y}^{\mathrm{H}}\rangle\right) - \langle\Phi(\mathbf{Y})\mathbf{Y}^{\mathrm{H}}\rangle\right] \cdot \mathbf{W}_i(\omega)$, where $\Phi(\mathbf{y}) = \phi(|\mathbf{y}|)\cdot e^{j\cdot\angle(\mathbf{y})}$, $\phi(x) = \tanh(gx)$ and $g = 100$. To solve the permutation problem of frequency domain ICA, we employed the DOA and correlation approach [8], and to solve the scaling problem of frequency domain ICA, we used the minimum distortion principle [9].

5.2 Performance measures

We used the signal to interference ratio (SIR) as a measure of separation performance, and the signal to distortion ratio (SDR) as a measure of sound quality:

$$\mathrm{SIR}_i = 10\log\frac{\sum_n y_{is_i}^2(n)}{\sum_n(\sum_{i\neq j} y_{is_j}(n))^2} \quad (9)$$

Table 1: Percentage of each source power. T_R=0ms.

combi.s	DCmask $M_{\mathrm{DC}}^{12}=\|F_3\|$	DCmask $M_{\mathrm{DC}}^{23}=\|F_1\|$
m-m-f	S_1:75, S_2:23, S_3:2	S_1:2, S_2:22, S_3:76
m-m-m	S_1:71, S_2:24, S_3:5	S_1:4, S_2:24, S_3:72
f-f-f	S_1:76, S_2:22, S_3:2	S_1:2, S_2:23, S_3:75

Table 2: Results of T_R=0 ms simulations. 'Conv. 1': with conventional method **1**, 'BMICA 12' and 'BMICA 23': with conventional method **2**, 'DCICA 12' and 'DCICA 23': with proposed method.

	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Conv. 1	18.0	8.9	18.4	7.9	11.5	8.3
BMICA 12	12.6	5.9	-	18.1	15.2	-
BMICA 23	-	6.1	13.0	-	13.6	17.4
DCICA 12	16.1	4.8	-	15.2	12.5	-
DCICA 23	-	4.6	16.2	-	12.8	15.4

[dB]

$$\mathrm{SDR}_i = 10\log \frac{\sum_n x_{ks_i}^2(n)}{\sum_n (x_{ks_i}(n) - \alpha y_{is_i}(n-D))^2}, \qquad (10)$$

where y_i is the estimation of s_i, and y_{is_j} is the output of the whole separating system at y_i when only s_j is active, and $x_{ks_i} = h_{ki} * s_i$ ($*$ is a convolution operator). α and D are parameters to compensate for the amplitude and phase difference between x_{ks_i} and y_{is_i}.

5.3 Experimental results

5.3.1 Applicability of ICA at the 2nd stage

Before trying to separate signals with our method, we investigated the applicability of ICA at the 2nd stage, because if the contribution of the *third* signal is large in $\hat{\mathbf{X}}^{pq}$, we cannot use a standard ICA algorithm at the 2nd stage. Table 1 shows the percentage of each source power extracted by M_{DC}^{12} and M_{DC}^{23}. Two signals are dominant and the contributions of the *third* signal are small. Therefore, we can say that we can use ICA in the 2nd stage.

5.3.2 Separation results

Table 2 shows the experimental results we obtained for T_R = 0 ms. The SIR and SDR values were averaged over three speaker combinations. The first row shows the results obtained by conventional method 1; the rows labeled 'BMICA' show the results obtained by conventional method 2, and the rows labeled 'DCICA' show the results obtained with our proposed method.

With conventional method 1, the SDR values were unsatisfactory, and a large musical noise was heard. In contrast, with our proposed method (DCICA), we were able to obtain high SDR values without any serious deterioration in the separation performance SIR. Compared with conventional method 2 (BMICA), although SDR values were a bit degraded, we hear no musical noise with DCICA. Some sound samples can be found at our web site [10].

Tables 3 and 4 show the results of reverberant tests for T_R = 130 ms. In the reverberant case, due to the decline of sparseness, the performance with all methods was worse than when T_R = 0 ms. However, we were able to obtain higher SDR values with DCICA than with conventional method 1 even in a reverberant environment without musical noise.

It should be noted that it remains difficult to separate signals at the center position with any method.

Table 3: Percentage of each source power. T_R=130ms.

combi.s	DCmask $M_{\mathrm{DC}}^{12}=\|F_3\|$	DCmask $M_{\mathrm{DC}}^{23}=\|F_1\|$
m-m-f	S_1:69, S_2:23, S_3:8	S_1:8, S_2:38, S_3:54
m-m-m	S_1:65, S_2:23, S_3:12	S_1:7, S_2:29, S_3:64
f-f-f	S_1:66, S_2:29, S_3:5	S_1:4, S_2:41, S_3:55

Table 4: Results of reverberant tests. T_R=130 ms.

	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Conv. 1	12.3	6.3	11.0	5.0	13.9	5.8
BMICA 12	9.8	5.5	-	7.8	15.9	-
BMICA 23	-	5.5	9.2	-	14.5	9.3
DCICA 12	11.4	3.2	-	6.7	6.3	-
DCICA 23	-	4.0	11.7	-	12.3	8.4

[dB]

6. CONCLUSION

We proposed utilizing a directivity pattern based continuous mask and ICA for BSS when speech signals outnumber the sensors. Our method avoids discontinuous zero-padding, and therefore, can separate the signals with no musical noise. Moreover, our new mask does not involve any heuristic parameters.

REFERENCES

[1] S. Haykin, *Unsupervised adaptive filtering*, John Wiley & Sons, 2000.

[2] S. Rickard and O. Yilmaz, "On the W-Disjoint orthogonality of speech," *Proc. ICASSP2002*, vol.1, pp. 529-532, 2002.

[3] F. J. Theis, C. G. Puntonet, E. W. Lang, "A histogram-based overcomplete ICA algorithm," *ICA2003*, pp. 1071-1076, 2003.

[4] L. Vielva, D. Erdogmus, C. Pantaleon, I. Santamaria, J. Pereda and J. C. Principe, "Underdetermined blind source separation in a time-varying environment," *Proc. ICASSP2002*, pp. 3049-3052, 2002.

[5] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," *Proc. ICA2000*, pp. 87-92, 2000.

[6] S. Araki, S. Makino, A. Blin, R. Mukai and H. Sawada, "Blind separation of more speech than sensors with less distortion by combining sparseness and ICA," *Proc. IWAENC2003*, pp. 271-274, 2003.

[7] A. Blin, S. Araki and S. Makino, "Blind source separation when speech signals outnumber sensors using a sparseness-mixing matrix combination," *Proc. IWAENC2003*, pp. 211-214, 2003.

[8] H. Sawada, R. Mukai, S. Araki and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *Proc. ICA2003*, pp. 505-510, 2003.

[9] K. Matsuoka and S. Nakashima, "A robust algorithm for blind separation of convolutive mixture of sources," *Proc. ICA2003*, pp. 927-932, 2003.

[10] http://www.kecl.ntt.co.jp/icl/signal/araki/nbficademo.html