4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

PERMUTATION CORRECTION AND SPEECH EXTRACTION BASED ON SPLIT SPECTRUM THROUGH FASTICA

H. Gotanda, K. Nobu, T. Koya, K. Kaneda, T. Ishibashi and N. Haratani

Graduate School of Advanced Technology, Kinki University
11–6 Kayanomori, Iizuka-shi, Fukuoka 820–8555, Japan
{gotanda haratani}@fuk.kindai.ac.jp
{nobu koya keiichi takaaki}@glab.elec.fuk.kindai.ac.jp

ABSTRACT

A blind source deconvolution method without indeterminacy of permutation and scaling is proposed by using notable features of split spectrum and locational information on signal sources. A method for extracting human speech exclusively is also proposed by taking advantage of the rule, the property of FastICA separates sources in order of large non-Gaussianity from their mixtures and the fact that human speeches are usually larger in non-Gaussianity than noises. The proposed methods have been verified by several experiments in a real room.

1. INTRODUCTION

Independent component analysis (ICA) can separate unknown signal sources from their mixtures if they are statistically independent [1, 2]. For the instantaneous mixtures, the original sources can be completely recovered in the time domain except for indeterminacy of scaling and permutation. In a real environment, however, the signals observed at microphones are not instantaneous mixtures but are convolved mixtures: delayed, attenuated and reverberated versions of the original sources. On account of this convolution, there have been reported many trials to separate the convolved mixtures in the frequency domain. However the indeterminacy of scaling and permutation appears at every frequency bin. In order to recover the sources in the time domain, this indeterminacy problem must be essentially solved before making an inverse transformation from the frequency to the time domain.

In the presence of noise, it is further desired to extract exclusively human speech. In this case, there still remains another problem even if the indeterminacy problem is successfully settled and the sources are properly estimated; of the estimated signals, which is the human speech and which is the undesired noise?

In the present paper, at first, the blind deconvolution in the frequency domain is reviewed. Ikeda *et al.* [3, 4] have suggested that the split spectrum defined by themselves is determined uniquely in terms of a signal source and its propagating characteristics to a microphone and has no scaling inderminacy. By applying their suggestion to the case of two sources and two microphones, a permutation correction rule is proposed under the condition that, of the two sources, the first source is in the side of a microphone and the second source is in the side of the other microphone. This rule also provides proper estimates for the sources. Further a method for extracting exclusively a human speech is proposed by taking advantage of the rule and the following two facts: FastICA proposed by Hyvärinen [5, 6] yields a separated signal in order of large non-Gaussianity and human speeches are usually larger in non-Gaussianity than noises. From several experiments in a real room, it has been confirmed that the proposed rule and method are valid.

2. BLIND DECONVOLUTION IN FREQUENCY DOMAIN

Consider the case that two statistically independent sound sources are observed by two microphones as

$$\mathbf{x}(t) \quad = \quad G(t) * \mathbf{s}(t) \tag{1}$$

where $\mathbf{s}(t) = [s_1(t), s_2(t)]^T$ denotes the source signals, $\mathbf{x}(t) = [x_1(t), x_2(t)]^T$ the convolved mixtures, $*$ the convolutional operator, $G(t)$ a matrix whose elements are transfer functions from the sources to the microphones. The mixtures are transformed into the short time spectra by the discrete Fourier transform:

$$x_j(\omega,k) = \sum_t e^{-\sqrt{-1}\omega t} x_j(t) w(t - k\tau) \quad (j = 1, 2; k = 0, 1, \cdots, K-1) \tag{2}$$

where $\omega (= 0, 2\pi/M, \cdots, 2\pi(M-1)/M)$ denotes a frequency bin, M the number of samples in a frame, k the frame number, τ the frame shift, $w(t)$ a window function.

In the frequency domain, the mixtures are expressed by

$$\mathbf{x}(\omega, k) \quad = \quad G(\omega)\mathbf{s}(\omega, k) \tag{3}$$

where $\mathbf{s}(\omega, k)$ and $G(\omega)$ are the Fourier transformed representations of the windowed sources and the transfer function matrix, respectively. The mixtures are generally whitened so that its covariance matrix becomes equal to the identity matrix. This whitening process can be expressed as $\widetilde{\mathbf{x}}(\omega) = Q(\omega)\mathbf{x}(\omega)$ where $Q(\omega)$ is a whitening matrix.

The separated spectra $\mathbf{u}(\omega, k) = [u_1(\omega, k) u_2(\omega, k)]^T$ can be obtained as

$$\mathbf{u}(\omega, k) \quad = \quad H(\omega)\widetilde{\mathbf{x}}(\omega, k) \tag{4}$$

where $H(\omega) = [\overline{\mathbf{h}}_1^T(\omega), \overline{\mathbf{h}}_2^T(\omega)]^T$ is a demixing matrix.

2.1. FastICA Algorithm

Under the assumption that all the spectra of whitened mixtures $\widetilde{\mathbf{x}}(\omega,k)$ are zero-mean and they have unit variances and uncorrelated real and imaginary parts of equal variances, FastICA algorithm [7] is formulated in the frequency domain as follows.

$$\mathbf{h}_n^+(\omega) = \frac{1}{K}\sum_{k=0}^{K-1}\{\widetilde{\mathbf{x}}(\omega,k)\overline{u}_n(\omega,k)f(|u_n(\omega,k)|^2) - [f(|u_n(\omega,k)|^2)+|u_n(\omega,k)|^2 f'(|u_n(\omega,k)|^2)]\mathbf{h}_n(\omega)\} \quad (5)$$

$$\mathbf{h}_n(\omega) = \mathbf{h}_n^+(\omega)/\parallel \mathbf{h}_n^+(\omega)\parallel \quad (6)$$

where $\mathbf{h}_n(\omega)(n=1,2)$ a demixing weight vector, $f(\cdot)$ is a nonlinear function, $f'(\cdot)$ is its differential and $\overline{\ }$ denotes the conjugation. At each frequency ω, the demixing weights are updated by the FastICA algorithm until a convergence condition $\overline{\mathbf{h}}_n^T(\omega)\mathbf{h}_n^+(\omega) \simeq 1$ can be satisfied. In addition, $\mathbf{h}_2(\omega)$ is orthogonalized with $\mathbf{h}_1(\omega)$ such as $\mathbf{h}_2(\omega)=\mathbf{h}_2(\omega)-\mathbf{h}_1(\omega)\overline{\mathbf{h}}_1^T(\omega)\mathbf{h}_2(\omega)$, and $\mathbf{h}_2(\omega)$ is again regularized by Eq.(6). The separated spectra $\mathbf{u}(\omega,k)$ are yielded by substituting $\mathbf{h}_n(\omega)$ to Eq.(4). The separated signals $\mathbf{u}(t)$ can be obtained by the inverse discrete Fourier transform of these spectra $\mathbf{u}(\omega,k)$ to the time domain.

2.2. Indeterminacy of Scaling and Permutation

In the frequency domain, the indeterminacy of scaling and permutation occur at every frequency ω:

$$H(\omega)Q(\omega)G(\omega) = PD(\omega) \quad (7)$$

where P is a permutation matrix of which all the elements of each column and row are 0 except for one element with value 1, and $D(\omega)=\mathrm{diag}[d_1(\omega),d_2(\omega)]$ is a diagonal matrix, of which elements $d_n(\omega)$ denote scaling factors at frequency ω and take arbitrary complex values resulting from the whitening process. This means that not only amplitude but also phase are indeterminate. Therefore, the indeterminacy of permutation, amplitude and phase must be settled to get a meaningful signal $u_n(t)$ before inversely transforming $\mathbf{u}(\omega,k)$ from the frequency domain to the time domain.

3. SPLIT SPECTRUM AND ITS UNIQUENESS

A flow from the source $s_i(\omega,k)$ to the separated spectra $u_n(\omega,k)$ is shown in Fig.1, where the nodes yielding $u_1(\omega,k)$ and $u_2(\omega,k)$ are represented by Node A and B, respectively, for convenience of the later discussion. Split spectra $\mathbf{v}_A(\omega,k)=[v_{A1}(\omega,k),v_{A2}(\omega,k)]^T$ at Node A and $\mathbf{v}_B(\omega,k)=[v_{B1}(\omega,k),v_{B2}(\omega,k)]^T$ at Node B are defined by Ikeda *et al.*, respectively, as follows [3, 4]:

$$\begin{bmatrix}v_{A1}(\omega,k)\\ v_{A2}(\omega,k)\end{bmatrix} = (H(\omega)Q(\omega))^{-1}\begin{bmatrix}u_1(\omega,k)\\ 0\end{bmatrix} \quad (8)$$

$$\begin{bmatrix}v_{B1}(\omega,k)\\ v_{B2}(\omega,k)\end{bmatrix} = (H(\omega)Q(\omega))^{-1}\begin{bmatrix}0\\ u_2(\omega,k)\end{bmatrix}. \quad (9)$$

First, we consider the case where there occur no permutations but there occurs the scaling problem. In this case,

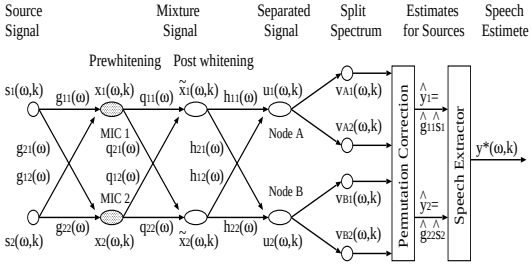

Figure 1: Signal flow: sound sources, convolutive mixtures, separated signals, split spectra.

the separated spectra $\mathbf{u}(\omega,k)$ are given by

$$\begin{bmatrix}u_1(\omega,k)\\ u_2(\omega,k)\end{bmatrix} = \begin{bmatrix}d_1(\omega)s_1(\omega,k)\\ d_2(\omega)s_2(\omega,k)\end{bmatrix} \quad (10)$$

and the split spectra are derived to be expressed as a product of the source signal $s_i(t)$ and the transfer function $g_{ji}(\omega)$ from the i-th source to the j-th microphone (See Appendix).

$$\begin{bmatrix}v_{A1}(\omega,k)\\ v_{A2}(\omega,k)\end{bmatrix} = \begin{bmatrix}g_{11}(\omega)s_1(\omega,k)\\ g_{21}(\omega)s_1(\omega,k)\end{bmatrix} \quad (11)$$

$$\begin{bmatrix}v_{B1}(\omega,k)\\ v_{B2}(\omega,k)\end{bmatrix} = \begin{bmatrix}g_{12}(\omega)s_2(\omega,k)\\ g_{22}(\omega)s_2(\omega,k)\end{bmatrix}. \quad (12)$$

As seen in Eq.(11), Node A generates a pair of split spectra originated from the first source $s_1(t)$; v_{A1} $(=g_{11}s_1)$ is the observation of $s_1(t)$ through Mic 1 and v_{A2} $(=g_{21}s_1)$ is the observation of $s_1(t)$ through Mic 2. Node B also generates another pair of split spectra originated from the second source $s_2(t)$; v_{B1} $(=g_{12}s_2)$ is through Mic 1 and v_{B2} $(=g_{22}s_2)$ is through Mic 2. Note that the split spectra have no interdeterminacy of scaling because, at frequency bins, the scaling factors do not take arbitrary values anymore but take the values equal to the transfer functions $g_{ji}(\omega)$.

Next, we consider the case where both permutation and scaling problems occur. In this case, the separated spectra are given by

$$\begin{bmatrix}u_1(\omega,k)\\ u_2(\omega,k)\end{bmatrix} = \begin{bmatrix}d_1(\omega)s_2(\omega,k)\\ d_2(\omega)s_1(\omega,k)\end{bmatrix}. \quad (13)$$

The split spectra are also derived to be expressed as a product of $s_i(t)$ and $g_{ji}(\omega)$ (See Appendix).

$$\begin{bmatrix}v_{A1}(\omega,k)\\ v_{A2}(\omega,k)\end{bmatrix} = \begin{bmatrix}g_{12}(\omega)s_2(\omega,k)\\ g_{22}(\omega)s_2(\omega,k)\end{bmatrix} \quad (14)$$

$$\begin{bmatrix}v_{B1}(\omega,k)\\ v_{B2}(\omega,k)\end{bmatrix} = \begin{bmatrix}g_{11}(\omega)s_1(\omega,k)\\ g_{21}(\omega)s_1(\omega,k)\end{bmatrix}. \quad (15)$$

In this case, as seen in Eq.(14), a pair of split spectra generated at Node A is not originated from $s_1(t)$ but from $s_2(t)$; v_{A1} $(=g_{12}s_2)$ and v_{A2} $(=g_{22}s_2)$ are the observations of s_2, respectively, through Mic 1 and Mic 2. Another pair of split spectra at Node B is originated from $s_1(t)$; v_{B1} $(=g_{11}s_1)$ is the observation through Mic 1 and v_{B2} $(=g_{21}s)$ is the one through Mic 2. These split spectra have also no interdeterminacy of scaling any longer.

Table 1: Decision rule for permutation.

Permutation	Split Spectral Difference	
	Node A $v^A_{1-2}(\omega,k)$	Node B $v^B_{1-2}(\omega,k)$
Not occur	positive	negative
Occur	negative	positive

From the above discussion, it is clarified that every split spectrum is uniquely determined as a product of a source spectrum and its transfer function; although the combination of the source and transfer function differ depending on whether permutations occur or not. It is also clarified that every split spectrum has no ambiguity of scaling in that the scaling factor is a transfer function itself; while the scaling factor $d_n(\omega)$ for the separated spectrum $u_n(\omega)$ varies arbitrarily, depending on the whitening result at individual frequency bin ω.

4. PERMUTATION CORRECTION RULE

4.1. Decision Rule for Permutation

Hereafter we assume that the first source $s_1(t)$ is closer to Mic 1 than to Mic 2 and the second source $s_2(t)$ is closer to Mic 2 than to Mic 1. From this assumption, gain inequalities on the transfer functions from the source 1 and 2 are obtained, respectively, by

$$|g_{11}(\omega)| > |g_{21}(\omega)| \quad (16)$$
$$|g_{12}(\omega)| < |g_{22}(\omega)|. \quad (17)$$

We also define the difference between two split spectra pairwisely generated at Node A and B, respectively, by

$$v^A_{1-2}(\omega,k) = |v_{A1}(\omega,k)| - |v_{A2}(\omega,k)| \quad (18)$$
$$v^B_{1-2}(\omega,k) = |v_{B1}(\omega,k)| - |v_{B2}(\omega,k)|. \quad (19)$$

Under the above preparation, we derive a decision rule whether the permutation occurs or not [8]. In the case of no permutation, it is derived by substituting Eqs.(11) and (12) to Eqs.(18) and (19) that the difference $v^A_{1-2}(\omega,k)$ at Node A is positive while the difference $v^B_{1-2}(\omega,k)$ at Node B is negative. On the other hand, in the case of permutation, it is derived by substituting Eqs.(14) and (15) to Eqs.(18) and (19) that $v^A_{1-2}(\omega,k)$ is negative while $v^B_{1-2}(\omega,k)$ is positive. These results are summarized in Table 1 as a decision rule for permutation. If the spectral difference is positive at Node A and negative at Node B, the rule gives the decision that no permutations occur. On the contrary, if the spectral difference is negative at Node A and positive at Node B, the rule gives the decision that permutation occurs.

4.2. Permutation Correction Rule and Recovered Signal in Time Domain

There exist two candidate estimates for one source; *e.g.*, for the first source $s_1(t)$, $v_{A1}(\omega,k)$ and $v_{A2}(\omega,k)$ in the case of no permutation, and $v_{B1}(\omega,k)$ and $v_{B2}(\omega,k)$ in the case of permutation. Here, in the case of no permutation, we adopt $v_{A1}(\omega,k)$ as the spectral estimate $y_1(\omega,k)$ for the first source $s_1(t)$; this is because $v_{A1}(\omega,k)$, the split spectrum observed at Mic 1, is larger and thus may be less disturbed by ambient noise than $v_{A2}(\omega,k)$, the split spectrum observed at Mic 2. For the same reason, in the case of permutation, we adopt $v_{B1}(\omega,k)$ as the spectral estimate $y_1(\omega,k)$ for the first source $s_1(t)$. Then, for the first source $s_1(t)$, we can formulate a permutation correction rule as

$$y_1(\omega,k) = \begin{cases} v_{A1}(\omega,k) \text{ if } v^A_{1-2}>0, v^B_{1-2}<0 \\ v_{B1}(\omega,k) \text{ if } v^A_{1-2}<0, v^B_{1-2}>0. \end{cases} \quad (20)$$

Similarly, as the spectral estimate $y_2(\omega,k)$ for the second source $s_2(t)$, we adopt $v_{B2}(\omega,k)$ in the case of no permutation and $v_{A2}(\omega,k)$ in the case of permutation:

$$y_2(\omega,k) = \begin{cases} v_{A2}(\omega,k) \text{ if } v^A_{1-2}<0, v^B_{1-2}>0 \\ v_{B2}(\omega,k) \text{ if } v^A_{1-2}>0, v^B_{1-2}<0. \end{cases} \quad (21)$$

The recovered signal $y_i(t)$ $(i=1,2)$ in the time domain for the signal source $s_i(t)$ can be obtained by applying the inverse Fourier transform of sepectrograms $\{y_i(\omega,k)|k = 0,1,\cdots,K-1\}(i=1,2)$.

$$y_i(t) = \frac{1}{2\pi}\frac{1}{W(t)}\sum_k\sum_\omega e^{\sqrt{-1}\omega(t-k\tau)}y_i(\omega,k) \quad (22)$$

where $W(t)=\sum_k w(t-k\tau)$.

5. SOURCE ESTIMATION AND SPEECH EXTRACTION

5.1. Source Estimation

It should be noted that the signs of split spectral difference $v^A_{1-2}(\omega,k)$ and $v^B_{1-2}(\omega,k)$ depend on the frequency bin ω but are independent of k; in a single frequency bin, the difference at a node is calculated by using two split spectra originated from the same source signal.

Thus, instead of correcting permutation at every frame k as in Eqs.(20) and (21), we correct the permutation at every sequence: $v_{nl}(\omega) = \{v_{nl}(\omega,k)|k=0,1,\cdots,K-1\}(n = A,B; l = 1,2)$. To this end, we define a power of split spectrum sequence as

$$P_{nl}(\omega) = \frac{1}{K}\sum_{k=0}^{K-1}|v_{nl}(\omega,k)|^2. \quad (23)$$

Then, the permutation correction rules of Eqs. (20) and (21) are modified as follows.

$$y_1(\omega) = \begin{cases} v_{A1}(\omega) \text{ if } P^A_{1-2}(\omega)>0, P^B_{1-2}(\omega)<0 \\ v_{B1}(\omega) \text{ if } P^A_{1-2}(\omega)<0, P^B_{1-2}(\omega)>0 \end{cases} \quad (24)$$

$$y_2(\omega) = \begin{cases} v_{A2}(\omega) \text{ if } P^A_{1-2}(\omega)<0, P^B_{1-2}(\omega)>0 \\ v_{B2}(\omega) \text{ if } P^A_{1-2}(\omega)>0, P^B_{1-2}(\omega)<0 \end{cases} \quad (25)$$

where $P^A_{1-2}(\omega) = P_{A1}(\omega) - P_{A2}(\omega)$ and $P^B_{1-2}(\omega) = P_{B1}(\omega) - P_{B2}(\omega)$ are the power differences at Node A and B.

For some trial experiments, at high frequencies over 3.1[KHz] there appeared two exceptional instances such that

both of the power differences at the two nodes take the same signs: one is the instance of $P^A_{1-2}(\omega) > 0$ and $P^B_{1-2}(\omega) > 0$ and the other is the instance of $P^A_{1-2}(\omega) < 0$ and $P^B_{1-2}(\omega) < 0$. On close inspection, the reason has been found to be because the power $P_{nl}(\omega)$ is extremely small and is disturbed by ambient noises.

Note that the permuation correction rules described in Eqs.(24) and (25) can be performed in accordance with the power difference at either of the two nodes, instead of using the two power differences at both of the two nodes. Thus, we first define a node-power at Node A and B, respectively, by $PA(\omega) = P_{A1}(\omega) + P_{A2}(\omega)$ and $PB(\omega) = P_{B1}(\omega) + P_{B2}(\omega)$. Then we select the node of larger node-power as a significant node for permutation correction. Finally, the permutation correction rule is modified as follows: If Node A is significant ($PA(\omega) > PB(\omega)$), then

$$y_1(\omega) = \begin{cases} v_{A1}(\omega) \text{ if } P^A_{1-2}(\omega) > 0 \\ v_{B1}(\omega) \text{ if } P^A_{1-2}(\omega) < 0 \end{cases} \tag{26}$$

$$y_2(\omega) = \begin{cases} v_{A2}(\omega) \text{ if } P^A_{1-2}(\omega) < 0 \\ v_{B2}(\omega) \text{ if } P^A_{1-2}(\omega) > 0. \end{cases} \tag{27}$$

If Node B is significant ($PA(\omega) < PB(\omega)$), then

$$y_1(\omega) = \begin{cases} v_{A1}(\omega) \text{ if } P^B_{1-2}(\omega) < 0 \\ v_{B1}(\omega) \text{ if } P^B_{1-2}(\omega) > 0 \end{cases} \tag{28}$$

$$y_2(\omega) = \begin{cases} v_{A2}(\omega) \text{ if } P^B_{1-2}(\omega) > 0 \\ v_{B2}(\omega) \text{ if } P^B_{1-2}(\omega) < 0. \end{cases} \tag{29}$$

5.2. Speech Extractor

Assume the situation that, of the two sources, one is a human speech and the other is a noise, and that it is uncertain whether the speech source is located in the side of Mic 1 or Mic 2. In other words, there is no prior information on the location of the two sources. Under this situation, for extracting the speech exclusively, there still remains another problem even if the two sources are successfully estimated; of the two estimated spectra, which is the human speech and which is the undesired noise?

FastICA proposed by Hyvärinen [6] generates the separated signal in order of large non-Gaussianity. Human speech is usually larger in non-Gaussianity than noises. Under the above situation, therefore, the separated spectra for a speech are yielded more frequently at Node A than at Node B. In other words, Node A gives the spectral estimate for a speech at higher probability than Node B. Therefore it is concluded that permutations do not occur at many frequency bins if the source 1 is a speech, and permutations occur at many frequency bins if the source 2 is a speech. This conclusion is formulated as a method for extracting only the spectral estimate $y^*(\omega)$ of the human speech:

$$Y^* = \begin{cases} Y_1 \text{ if } \#NoPerm > \#Perm \\ Y_2 \text{ if } \#NoPerm < \#Perm \end{cases} \tag{30}$$

where $Y^* = \{y^*(\omega_m) \| m = 1, 2, \cdots, M\}$, $Y_i = \{y_i(\omega_m) \| m = 1, 2, \cdots, M\}$ $(i = 1, 2)$, M denotes the number of all the frequency bins, and $\#NoPerm$ and $\#Perm$ denote the number of frequency bins, respectively, with permutation and without permutation.

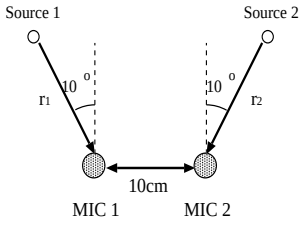

Figure 2: Locations of sources and microphones in experiments.

Table 2: No permutation rates[%].

Noise Location	30[cm]	60[cm]
Proposed Correction	93.26	97.92
FastICA	89.77	90.93

6. EXPERIMENTAL RESULTS

In order to verify our proposals, several experiments were carried out in a laboratory($L7.3m \times W6.5m \times H2.9m$) where the reverberation time was 0.5[sec] and the ambient noise was 48.0[dB]. As shown in Fig. 2, Source 1 was located at a distance of r_1 = 10cm with 10° angle from Mic 1, and Source 2 was located at a distance of r_2 = 30, 60cm with 10° angle from Mic 2. Mic 1 and 2 with 10cm spacing were the condenser microphones with band width 200-5000[Hz].

The microphone signals were sampled at the rate of 8KHz with 16[Bit] resolution. The sampled data were windowed by Hamming window of frame length 16[msec] with frame period 8[msec]. In the FastICA algorithm, the nonlinear function was chosen as $f(|u_n(\omega,k)|^2) = 1 - 2/(e^{2|u_n(\omega,k)|^2} + 1)$ and the weight were initialized by random numbers (-1,1). The algorithm was iterated until the convergence criterion $|\overline{\mathbf{h}}^T(\omega)\mathbf{h}^+(\omega)| > 0.999999$ was satisfied.

Over a loudspeaker located at Source 2, we play a roaring train noise recorded at a station premises [9]; departure and arrival announcements and passengers' conversations were also included in the noise. The noise level was in average 82.1[dB] at a distance of r_2=30cm and 76.3[db] at a distant of r_2=60cm from the loudspeaker. Under these conditions, 4 kinds of Japanese words (Tokyo, Sin Iizuka, Kinki Daigaku, Sangyo Gijyutu Kenkyuhka) were uttered by 4 male and 4 female speakers at the position of r_1=10cm. In total, 64 data sets were obtained and the utterance duration was from 2.3 to 6.9 seconds.

6.1. Permutation Correction

Table 2 shows no permutation rate which is defined by

$$NoPermRate = \frac{\#NoPerm}{\#NoPerm + \#Perm}. \tag{31}$$

From the results, it is found that the permutation problem are well resolved by applying only FastICA. This reflects

Table 3: Speech extraction rates[%].

Noise Location	30[cm]	60[cm]
Proposed Method	100	100
Kurtosis	87.50	96.88

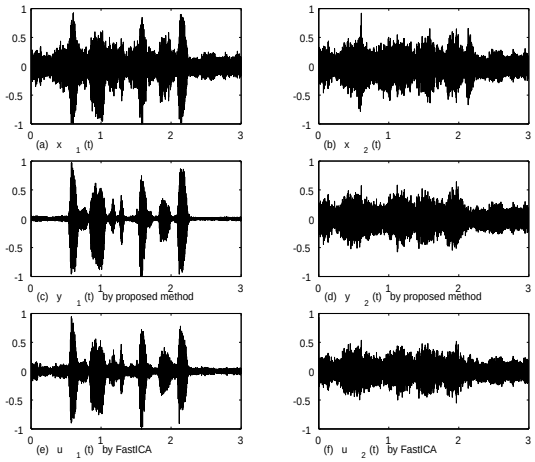

Figure 3: Experimental results when a male speaker (target sound source) uttered "Sangyou gijyutu kenkyuhka" under the station premise noise from the loudspeaker ($r_2 = 30$[cm]) : (a), (b) Mixed signals recorded by Mic 1 and 2 respectively. (c), (d) Restored signals of the male utterance and the station premise noise, respectively, by the proposed method. (e), (f) Restored signals by FastICA.

the property that FastICA separates a signal in order of high non-Gaussianity. Much better rates are obtained by our proposed correction rule given by Eqs.(26) to (29). The further the noise source is located from the microphones and thus the lower the noise level becomes, the higher no permutation rates become in the case of our rule, while only little improvement can be seen in the case of FastICA.

We examined what frequency bins the permutation occurred at. In the case of our rule, the permutation occurred at frequency bins with extremely small power spectrum. In the case of FastICA, contrastedly, the permutation occurred even at frequency bins with large power spectrum and its occurrence was independent of magnitude of power spectrum. This discrepancy in permutation occurrence at frequency bins can be considered to result in the fact that the restored waveform in Fig. 3(c) is much less noisy than that in Fig. 3(e), although the no-permutation rate by our proposal is only 3.5% higher than that by FastICA at $r_2 = 30$cm. By articulation tests, it was confirmed that the restored utterances by our rule are much better than those by FastICA. This test result can be also attributed to the discrepancy of permutation occurrence at frequency bins.

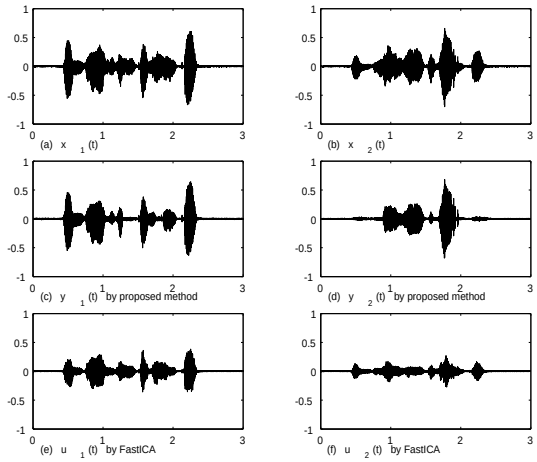

Figure 4: Experimental results when a male speaker uttered "Sangyou gijyutu kenkyuhka" and a female speaker "Shin Iizuka" : (a), (b) Mixing signals recorded by Mic 1 and 2 respectively. (c), (d) Restored signals of the male and female utterance, respectively, by the proposed method. (e), (f) Restored signals by FastICA.

6.2. Speech Extraction

In trial experiments for the FastICA algorithm, it was confirmed that, at 80% frequency bins, the separated spectrum for a speaker's utterance appeared at Node A, irrespective of the speaker position.

Table 3 shows the speech extraction rates for the 64 data sets; the rate is defined by $c/64$ where c is the number of times the estimate for the speaker's utterance is properly selected. As seen in the results, the method described in Eq.(30) extracted only the human utterance without exception in both cases of $r_2 = 30$ and 60[cm]. For comparison, we computed the kurtosis of $y_i(t)$ given by inverse Fourier transform of the spectral estimate $y_i(\omega)$ and adopted $y_i(t)$ of larger kurtosis as the human utterance. The rates by this kurtosis based approach is 87.50% at $r_2 = 30$[cm] and 96.88% at $r_2 = 60$[cm]. These results mean that proposed method is robust to the noise level, while the performance of the kurtosis based approach depends on the noise level.

6.3. Source Estimation

For experiments for the source estimation, we acquired another set of mixture data in the following way. One of two speaker uttered a word at r_1=10cm and the other uttered another word at r_2=10cm in Fig. 2. These words were again uttered after changing the speaker's position. Then we acquired 4 sets of mixtures for a pair of words. This data acquisition was repeated for 6 speakers (3 males and 3 females) and for 3 pairs of words. In total, 180 sets of mixtures were obtained and their utterance duration was from 2.3 to 4.1 seconds.

No permutation rate obtained by our permutation correction was 99.08% while the rate obtained by applying only

FastICA was 50.60%. This fact assures that the correction rule is valid even if both of the two sources are speeches. Fig. 4 shows the source estimates $y_i(t)$ recovered in the time domain. From the results, it is found that there exist less cross talk in Fig. 4(c) and (d) by our proposal while in Fig. 4(e) and (f) by FastICA there exist much cross talk because of the unresolved permutation described above. It is also found that the waveforms in Fig. 4(c) and (d) are properly scaled while those in Fig. 4(e) and (f) not.

7. CONCLUSIONS

In the present paper, it has been clarified that the split spectrum is uniquely determined in terms of the sound source and its propagating characteristics to the microphone. Taking advantage of this fact, the permutation correction rule has been proposed under the condition that, of the two sources, the source 1 is in the side of a microphone and the source 2 is in the side of the other microphone. This rule can give the estimates for the source signals, too. The method for extracting only a human speech has been also proposed by taking advantage of the rule, the FastICA's useful property that the signal is separated in order of large non-Gaussianity and the fact that human speech signals are usually larger in non-Gaussianity than noises. The proposed rule and method have been verified by several experiments in a real room. The assumption on source and microphone location can lead to phase inequalities, from which another permutation correction can be developed. The results will be reported later.

8. REFERENCES

[1] J. Bell and T. J. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," Neural Computation, vol. 7, pp.1129-1159, 1995.

[2] S. Amari, "Natural gradient works efficiently in learning," Neural Computation, vol. 10, pp.251-276, 1998.

[3] S. Ikeda and N. Murata, "A method of ICA in time frequency domain", Proc. ICA '99, pp.365-371, Aussois, France, Jan. 1999.

[4] N. Murata, S. Ikeda and A. Ziehe, "An method of blind separation based on temporal structure of Signals", Neurocomputing, vol. 41,Issue 1-4, pp.1-24, 2001.

[5] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," IEEE Trans. Neural Networks, vol. 10 (3), pp.626-634, 1999.

[6] A. Hyvärinen, and E. Oja, "Independent component analysis: algorithms and applications," Neural Networks, vol. 13 (4-5), pp.411-430, 2000.

[7] E. Bingham, and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," Int. J. of Neural Systems, vol. 10 (1), pp.1-8, 2000.

[8] K. Nobu, T. Ishibashi, T. Koya, K. Kaneda, N. Haratani and H. Gotanda, "Extraction of target sounds using their locational information," Technical Report of IEICE, NC2002-33, pp.19-24, 2002.

[9] NTT Advanced Technology Corporation, "Ambient Noise Database for Telephonometry 1996," 1996.

Appendix: Derivation of Eqs.(20),(21),(24) and (25)

First, consider the case where both permutation and scaling occur:

$$HQG = PD \tag{32}$$

where P is the permutation matrix and $D = \mathrm{diag}[d_1, d_2]$ is the scaling matrix. Then, the separated signal $\mathbf{u} = [u_1, u_2]^T$ is developed as $\mathbf{u} = H\tilde{\mathbf{x}} = HQG\mathbf{s} = PD\mathbf{s}$ where $\tilde{\mathbf{x}}(= Q\mathbf{x})$ is the convolved mixture after whitening. Hence,

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} d_1 s_2 \\ d_2 s_1 \end{bmatrix} \tag{33}$$

Note that the separated signal is the permuted version of source with indeterminate scaling. The split spectra $\mathbf{v}_A = [v_{A1}, v_{A2}]^T$ generated at Node A is developed as

$$\begin{bmatrix} v_{A1} \\ v_{A2} \end{bmatrix} = (HQ)^{-1} \begin{bmatrix} u_1 \\ 0 \end{bmatrix} =$$
$$GD^{-1}P \begin{bmatrix} u_1 \\ 0 \end{bmatrix} = \frac{1}{d_1} \begin{bmatrix} g_{12} u_1 \\ g_{22} u_1 \end{bmatrix} \tag{34}$$

Therefore, from $u_1 = d_1 s_2$ in Eq.(33), the split spectrum is found to be expressed as a product of the second source s_2 and its transfer function g_{j2} $(j = 1, 2)$ to the j-th microphone:

$$\begin{bmatrix} v_{A1} \\ v_{A2} \end{bmatrix} = \begin{bmatrix} g_{12} s_2 \\ g_{22} s_2 \end{bmatrix} \tag{35}$$

The other split spectrum $\mathbf{v}_B = [v_{B1}, v_{B2}]^T$ generated at Node B is expanded as

$$\begin{bmatrix} v_{B1} \\ v_{B2} \end{bmatrix} = (HQ)^{-1} \begin{bmatrix} 0 \\ u_2 \end{bmatrix}$$
$$= GD^{-1}P \begin{bmatrix} 0 \\ u_2 \end{bmatrix} = \frac{1}{d_2} \begin{bmatrix} g_{11} u_2 \\ g_{21} u_2 \end{bmatrix} \tag{36}$$

and, from $u_2 = d_2 s_1$ in Eq.(33), is found to be expressed as a product of the first source s_1 and its transfer function g_{j1} $(j = 1, 2)$ to the j-th microphone:

$$\begin{bmatrix} v_{B1} \\ v_{B2} \end{bmatrix} = \begin{bmatrix} g_{11} s_1 \\ g_{21} s_1 \end{bmatrix} \tag{37}$$

Next, consider the case where the scaling is indeterminate but no permutations occur:

$$HQG = D \tag{38}$$

Hence, the separated signal is developed as $\mathbf{u} = HQG\mathbf{s} = D\mathbf{s}$, and is given in the same ordering to the source, *i.e.*, $[u_1, u_2]^T = [d_1 s_1, d_2 s_2]^T$. Therefore, from $u_1 = d_1 s_1$, the split spectra generated at Node A is found to be expressed as

$$\begin{bmatrix} v_{A1} \\ v_{A2} \end{bmatrix} = \begin{bmatrix} g_{11} s_1 \\ g_{21} s_1 \end{bmatrix} \tag{39}$$

From $u_2 = d_2 s_2$, the split spectra generated at Node B is found to be expressed as

$$\begin{bmatrix} v_{B1} \\ v_{B2} \end{bmatrix} = \begin{bmatrix} g_{12} s_2 \\ g_{22} s_2 \end{bmatrix} \tag{40}$$