MLSP 2007 DATA ANALYSIS COMPETITION: FREQUENCY-DOMAIN BLIND SOURCE SEPARATION FOR CONVOLUTIVE MIXTURES OF SPEECH/AUDIO SIGNALS

Hiroshi Sawada *Shoko Araki* *Shoji Makino*

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

This paper describes the frequency-domain approach to the blind source separation of speech/audio signals that are convolutively mixed in a real room environment. With the application of short-time Fourier transforms, convolutive mixtures in the time domain can be approximated as multiple instantaneous mixtures in the frequency domain. We employ complex-valued independent component analysis (ICA) to separate the mixtures in each frequency bin. Then, the permutation ambiguity of the ICA solutions should be aligned so that the separated signals are constructed properly in the time domain. We propose a permutation alignment method based on clustering the activity sequences of the frequency bin-wise separated signals. We achieved the overall winner status of MLSP 2007 Data Analysis Competition based on the presented method.

1. INTRODUCTION

MLSP 2007 Data Analysis Competition [1] was organized by the 2007 IEEE International Workshop on Machine Learning for Signal Processing (MLSP). The data analysis task was the blind separation of audio sources that had been mixed and then captured by multiple microphones in a real-room environment (e.g., solving a cocktail party problem). Such a task has been well recognized as a major application of blind source separation (BSS) or independent component analysis (ICA) techniques [2–5]. The difficulty of this problem lies in the fact that the mixing system is not simply instantaneous but convolutive, with delay and reflections. Let $\mathrm{s}_1, \ldots, \mathrm{s}_N$ be source signals and $\mathrm{x}_1, \ldots, \mathrm{x}_M$ be sensor observations. The convolutive mixture model is formulated as

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \sum_{l=0}^{P} \mathrm{h}_{jk}(l \cdot t_s)\, \mathrm{s}_k(t - l \cdot t_s), \quad j = 1, \ldots, M, \qquad (1)$$

where t represents time (a multiple of $t_s = 1/f_s$ with f_s being the sampling rate) and h_{jk} is the impulse response from source k to sensor j with $P+1$ samples. In a practical room situation, P would be some thousands even with $f_s = 8$ kHz sampling rate, and this makes the convolutive problem difficult to solve.

Many approaches have been proposed to the convolutive BSS problem. Among them, we consider the frequency-domain approach [6–14] where we apply a short-time Fourier transform (STFT) to the sensor observations $\mathrm{x}_j(t)$. If we use a sufficiently long frame for STFT to cover the main part of the impulse responses h_{jk}, the convolutive mixture (1) can be approximated well with an instantaneous mixture at each frequency f:

$$x_j(n, f) = \sum_{k=1}^{N} h_{jk}(f) s_k(n, f), \quad j = 1, \ldots, M, \qquad (2)$$

where n represents the time frame index, h_{jk} is the frequency response from source k to sensor j, and $s_k(n, f)$ is the time-frequency representation of a source signal s_k. Consequently, we can employ any complex-valued instantaneous ICA algorithm to separate the frequency bin-wise mixtures. Section 3 of this paper presents an efficient ICA procedure for frequency-domain speech/audio signals.

The drawback of frequency-domain BSS is that the permutation ambiguity of an ICA solution becomes a serious problem. The ambiguities should be aligned properly so that the separated signals that originate from the same source are grouped together. This problem is known as the permutation problem of frequency-domain BSS, and various methods [6–14] have been proposed for its solution. Section 4 discusses a strategy that exploits the mutual dependence of bin-wise separated signals across frequencies [8–10, 14]. The advantage of this strategy is that it is less affected by bad mixing conditions, such as severe reverberations or closely located sources, than another popular strategy based on time-difference-of-arrival estimations [10, 11].

Section 5 reports experimental results. We have our own experimental condition (3 sources and 3 observations) to validate the effectiveness of the proposed method. As regards the MLSP data analysis competition [1], our submitted results obtained the overall winner status, and we were then invited to submit this paper to present our approach.

2. FREQUENCY-DOMAIN BSS

Let us start with an overview of frequency-domain BSS. Figure 1 shows the system structure. First, sensor observations (1) are converted into frequency-domain time-series signals (2) by a short-time Fourier transform (STFT) with an L-sample frame and its S-sample shift:

$$x_j(n, f) \leftarrow \textstyle\sum_t \mathrm{x}_j(t)\, \mathit{win}_a(t - nSt_s)\, e^{-\imath 2\pi f t}, \qquad (3)$$

for all discrete frequencies $f \in \{0, \frac{1}{L}f_s, \ldots, \frac{L-1}{L}f_s\}$ and for frame indexes n. The analysis window win_a is defined as non-zero only in the L-sample interval $[-\frac{L}{2}t_s, (\frac{L}{2} - 1)\, t_s]$ and preferably tapers smoothly to zero at each end of the interval.

Next, separation is performed in each frequency bin f:

$$\mathbf{y}(n, f) = \mathbf{W}(f)\, \mathbf{x}(n, f), \qquad (4)$$

1-4244-1566-7/07/$25.00 ©2007 IEEE.

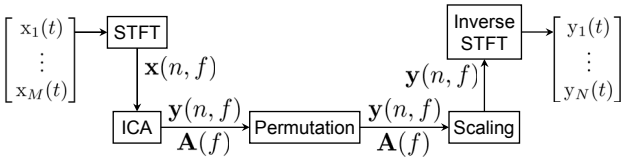

Fig. 1. System structure for frequency-domain BSS

where $\mathbf{x} = [x_1, \ldots, x_M]^T$ is the vector of observations, $\mathbf{y} = [y_1, \ldots, y_N]^T$ is the vector of separated signals, and $\mathbf{W}$ is an $N \times M$ separation matrix. We apply the complex-valued instantaneous ICA algorithm described in Sect. 3 for the calculation of $\mathbf{W}$. If ICA works well, we expect $y_1, \ldots, y_N$ to be close to the original source frequency components $s_1, \ldots, s_N$. However, the correspondence is up to the scaling and permutation ambiguities that an ICA solution inherently has. Even if we permute the elements of $\mathbf{y} = [y_1, \ldots, y_N]^T$ or multiply an element by a constant, it is still an ICA solution. In other words,

$$\mathbf{W}(f) \leftarrow \mathbf{\Lambda}(f)\,\mathbf{P}(f)\,\mathbf{W}(f) \tag{5}$$

is also an ICA solution for any permutation $\mathbf{P}(f)$ and diagonal $\mathbf{\Lambda}(f)$ matrices.

To align such ambiguities, it is advantageous to calculate basis vectors $\mathbf{a}_i = [a_{1i}, \ldots, a_{Mi}]^T$, $i = 1, \ldots, N$, and to represent the vector $\mathbf{x}$ by a linear combination of the basis vectors:

$$\mathbf{x}(n,f) = \sum_{i=1}^{N} \mathbf{a}_i(f) y_i(n,f) = \mathbf{A}(f)\,\mathbf{y}(n,f)\,, \tag{6}$$

where $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_N]$. If $\mathbf{W}$ has the inverse, the matrix is given simply by $\mathbf{A} = \mathbf{W}^{-1}$. Otherwise it is calculated as a least-mean-square estimator [15]

$$\mathbf{A} = \mathrm{E}\{\mathbf{x}\mathbf{y}^H\}(\mathrm{E}\{\mathbf{y}\mathbf{y}^H\})^{-1}\,,$$

which minimizes $\mathrm{E}\{||\mathbf{x} - \mathbf{A}\mathbf{y}||^2\}$.

In the next step, the permutation matrix $\mathbf{P}(f)$ is determined for each frequency f so that separated frequency components $y_i(n,f)$ are grouped together for the same source. Section 4 presents a method for permutation alignment. After $\mathbf{P}$ is determined, the vector of separated components $\mathbf{y}$ and the matrix $\mathbf{A}$ of basis vectors is updated as

$$\mathbf{y}(n,f) \leftarrow \mathbf{P}(f)\,\mathbf{y}(n,f)\,, \quad \forall\, n \tag{7}$$

$$\mathbf{A}(f) \leftarrow \mathbf{A}(f)\,[\mathbf{P}(f)]^T\,, \tag{8}$$

for each frequency f. Equation (6) is not changed by the update because a permutation matrix is an orthogonal matrix $\mathbf{P}^T\mathbf{P} = \mathbf{I}$.

Then, the scaling ambiguity is aligned [9] by adjusting a separated frequency component $y_i(n,f)$ to the observation $x_J(n,f)$ of an arbitrary selected reference sensor $J \in \{1, \ldots, M\}$:

$$y_i(n,f) \leftarrow a_{Ji}(f) y_i(n,f), \quad \forall\, i,n,f.$$

In a vector notation,

$$\mathbf{y}(n,f) \leftarrow \mathbf{\Lambda}(f)\,\mathbf{y}(n,f)\,, \quad \forall\, n \tag{9}$$

with a diagonal matrix

$$\mathbf{\Lambda}(f) = \mathrm{diag}[a_{J1}(f), \ldots, a_{JN}(f)] \tag{10}$$

aligns the scaling ambiguity for each frequency f.

At the end of the flow, time-domain output signals $\mathrm{y}_i(t)$ are obtained by the inverse operation of the STFT:

$$\mathrm{y}_i(t) = \sum_n \mathit{win}_s(t - nSt_s) \sum_{f \in \{0,\, \frac{1}{L}f_s,\, \ldots,\, \frac{L-1}{L}f_s\}} y_i(n,f)\, e^{\imath 2\pi f t}$$

where win_s is a synthesis window defined as non-zero only in the L-sample interval $[-\frac{L}{2}t_s,\, (\frac{L}{2}-1)\,t_s]$. The summation over the frame index n is with those that satisfy $-\frac{L}{2}t_s \le t - nSt_s \le (\frac{L}{2}-1)\,t_s$. To realize a perfect reconstruction, the analysis and synthesis windows should satisfy the condition

$$\textstyle\sum_n \mathit{win}_s(t - nSt_s)\, \mathit{win}_a(t - nSt_s) = 1$$

for any time t. A synthesis window that tapers smoothly to zero at each end is preferred in terms of mitigating the edge effect.

The procedure described above and depicted in Fig. 1 separates the mixtures in the frequency domain. However, in the MLSP 2007 Data Analysis Competition [1], time-domain filters were to be submitted for linear systems. The modification made for this purpose will be described in Sect. 5.2.

3. COMPLEX-VALUED ICA

This section presents a criterion and procedure for complex-valued ICA to separate the mixtures $\mathbf{x}(n,f)$ in the frequency domain. For a simpler notation, let us omit the frequency dependency f of the separation formula (4):

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n)\,. \tag{11}$$

For the calculation of the separation matrix $\mathbf{W}$, we maximize the log-likelihood function

$$\mathcal{J} = \mathrm{E}\{\log p(\mathbf{x}|\mathbf{W})\} = \log|\det \overline{\mathbf{W}}| + \sum_{i=1}^{N} \mathrm{E}\{\log p(y_i)\} \tag{12}$$

where

$$\overline{\mathbf{W}} = \begin{bmatrix} \mathrm{Re}(\mathbf{W}) & -\mathrm{Im}(\mathbf{W}) \\ \mathrm{Im}(\mathbf{W}) & \mathrm{Re}(\mathbf{W}) \end{bmatrix}$$

is a real-valued $2N \times 2M$ matrix introduced to transform the probability density function of a complex-valued vector [16].

An ICA algorithm generally assumes the source signal model with a probability density function. For a speech or audio signal y_i in the frequency domain, we employ the following density function

$$p(y_i) \propto \exp\left(-\frac{\sqrt{|y_i|^2 + \alpha}}{b}\right) \tag{13}$$

where $b > 0$ specifies the variance, and a small nonnegative parameter $\alpha \ge 0$ controls the smoothness around the origin $y_i = 0$. Figure 2 shows the assumed density functions (13) together with a complex-valued Gaussian distribution. The variance is normalized to 1 for all the cases. We see that (13) provides a sharper peaked distribution than the Gaussian distribution. Such a distribution models a speech/audio signal in the frequency domain very well [17]. As regards the parameter α, a smaller value gives a more sharply peaked distribution. However, α should be non-zero if the second order derivative of $\log p(y_i)$ is used in the ICA algorithm. Otherwise, α can be set to zero to make the density function simpler: $p(y_i) \propto \exp(-\frac{|y_i|}{b})$.

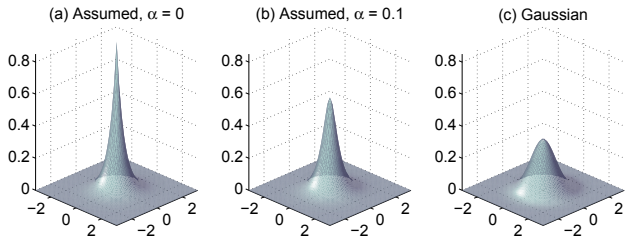

Fig. 2. Probability density functions of complex variables. (a) and (b) are assumed by (13), and (c) is a Gaussian distribution.

For computational efficiency and better separation performance, we adopt a 3-step procedure to maximize (12). The first step performs a whitening operation $\mathbf{V}$ which serves as a preprocessing for the FastICA algorithm [4] in the next step. The second step (FastICA) is constrained to learn only a unitary matrix $\mathbf{U}$. This makes the algorithm fast to converge by employing Newton's method. The third step has no constraint for the separation matrix $\mathbf{W}$. In this sense, it might improve the initial solution $\mathbf{W} = \mathbf{UV}$ obtained by the first and second steps.

The first step performs the whitening operation with an $M \times M$ matrix $\mathbf{V}$ by

$$\mathbf{z}(n) = \mathbf{V}\mathbf{x}(n)$$

such that the correlation matrix of the output vector $\mathbf{z} = [z_1, \ldots, z_M]^T$ becomes an identity matrix $\mathrm{E}\{\mathbf{z}\mathbf{z}^H\} = \mathbf{I}$. The whitening matrix $\mathbf{V}$ is simply given by $\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^H$ if we have an eigenvalue decomposition $\mathrm{E}\{\mathbf{x}\mathbf{x}^H\} = \mathbf{E}\mathbf{D}\mathbf{E}^H$.

The second step separates the whitened mixture $\mathbf{z}$ with a unitary matrix $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]^H$:

$$\mathbf{y}(n) = \mathbf{U}\mathbf{z}(n)\,. \tag{14}$$

If we consider a log-likelihood function similar to one in (12), the $\log|\det \mathbf{U}|$ term disappears since $\mathbf{U}$ is a unitary matrix. Thus, the criterion here is to maximize $\sum_{i=1}^{N} \mathrm{E}\{\log p(y_i)\}$. The FastICA algorithm [4] efficiently maximizes this criterion. The core part of the algorithm is designed to extract a separated signal y_i one by one with a unit-norm vector $\mathbf{u}_i$:

$$y_i(n) = \mathbf{u}_i^H \mathbf{z}(n)\,.$$

The vector $\mathbf{u}_i$ is updated by

$$\mathbf{u}_i \leftarrow \mathrm{E}\{g'(y_i)\}\mathbf{u}_i - \mathrm{E}\{g(y_i)\,\mathbf{z}\}, \tag{15}$$

where $g(y_i)$ and $g'(y_i)$ are the first and second derivatives of $\log p(y_i)$. With $b = 1$ for the assumed density function (13), we have

$$g(y_i) = \frac{\partial \log p(y_i)}{\partial y_i} = -\frac{y_i^*}{2\sqrt{|y_i|^2 + \alpha}}\,,$$

$$g'(y_i) = \frac{\partial g(y)}{\partial y_i^*} = -\frac{1}{2\sqrt{|y_i|^2 + \alpha}}\left[1 - \frac{1}{2}\frac{|y_i|^2}{|y_i|^2 + \alpha}\right],$$

where y_i^* is the complex conjugate of y_i. After every update, the vector $\mathbf{u}_i$ is normalized to unit-norm and made orthogonal to already found other unit-norm vectors by the Gram-Schmidt orthogonalization method.

The third step maximizes the log-likelihood function $\mathcal{J}$ in (12) with a general matrix $\mathbf{W}$ [18, 19]. We can start with a good initial solution $\mathbf{W} = \mathbf{UV}$ obtained by the first and the second steps, and then iteratively optimize $\mathbf{W}$ with a small step-size η

$$\mathbf{W} \leftarrow \mathbf{W} + \eta \cdot \Delta\mathbf{W}\,. \tag{16}$$

Therein the natural gradient [5, 20]

$$\Delta\mathbf{W} = \frac{\partial \mathcal{J}}{\partial \mathbf{W}^*}\mathbf{W}^H\mathbf{W} = \left[\mathbf{I} - \mathrm{E}\{\mathbf{\Phi}(\mathbf{y})\mathbf{y}^H\}\right]\mathbf{W}$$

is commonly used for its nice properties. The nonlinear functions $\mathbf{\Phi}$ are defined as

$$\mathbf{\Phi}(\mathbf{y}) = [\Phi(y_1), \ldots, \Phi(y_N)]^T\,,$$

$$\Phi(y_i) = -\frac{\partial \log p(y_i)}{\partial y_i^*} = \frac{y_i}{2b\sqrt{|y_i|^2 + \alpha}}\,,$$

in which we assume the density function (13). By setting $\alpha = 0$ and $b = 1/2$, the nonlinear function becomes simpler as

$$\Phi(y_i) = \frac{y_i}{|y_i|}\,.$$

4. PERMUTATION ALIGNMENT

In this section, we discuss a permutation alignment strategy based on the signal activity sequence $v_i^f(n)$ of a bin-wise separated signal $y_i(n, f)$. The signal activity sequence can be calculated in various ways as discussed in Sect. 4.1. We expect the correlation coefficient $\rho(v_i^f, v_j^g)$ of two activity sequences $v_i^f(n)$, $v_j^g(n)$ to be high if they originate from the same source. The rationale behind this is that the active time frames of bin-wise separated signals are likely to coincide among frequencies for the same source.

The correlation coefficient ρ between two real-valued sequences $v_i(n)$ and $v_j(n)$ is defined as

$$\rho(v_i, v_j) = \frac{r_{ij} - \mu_i\mu_j}{\sigma_i\sigma_j} \tag{17}$$

where

$$r_{ij} = \mathrm{E}\{v_i v_j\},\;\; \mu_i = \mathrm{E}\{v_i\},\;\; \sigma_i = \sqrt{\mathrm{E}\{v_i^2\} - \mu_i^2}$$

are the correlation, the mean, and the standard deviation, respectively. For any two sequences v_i and v_j, the correlation coefficient is bounded by $-1 \le \rho(v_i, v_j) \le 1$, and becomes 1 if the two sequences are identical.

4.1. Activity Sequence

The activity sequence of a separated signal y_i has commonly been represented by its envelope [8–10]

$$v_i^f(n) \leftarrow |y_i(n, f)|\,. \tag{18}$$

Envelopes usually result in high correlation coefficients for the same source among adjacent frequencies or harmonic frequencies [10]. However, they may end up with almost zero correlation coefficients for the same source among frequencies that have no specific relation. Figure 3 shows the envelopes of two separated signals at two such frequencies. The correlation coefficients are

$$\begin{bmatrix} \rho(v_1^f, v_1^g) & \rho(v_1^f, v_2^g) \\ \rho(v_2^f, v_1^g) & \rho(v_2^f, v_2^g) \end{bmatrix} = \begin{bmatrix} 0.10 & -0.14 \\ -0.19 & 0.06 \end{bmatrix},$$

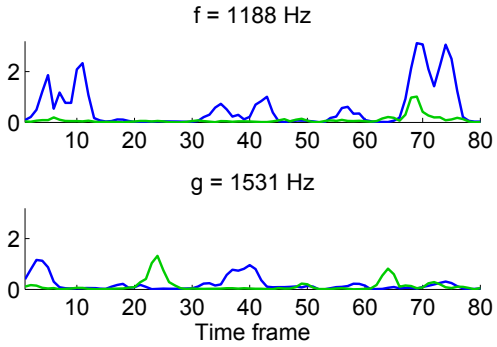

Fig. 3. Envelopes of two separated signals at two frequencies that have no specific relation

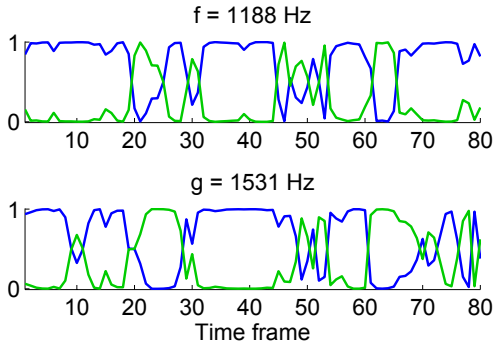

Fig. 4. Dominance measures of two separated signals at the same two frequencies as in Fig. 3

which are very low even for the same source.

Recently in [14], we proposed to use a dominance measure for the activity sequence. It represents how dominant the i-th separated signal is in the observations (6). An example of such a measure is the power ratio between the i-th separated signal and the total power sum of all the separated signals:

$$v_i^f(n) \leftarrow \mathit{powRatio}_i(n,f) = \frac{||\mathbf{a}_i(f)\, y_i(n,f)||^2}{\sum_{k=1}^N ||\mathbf{a}_k(f)\, y_k(n,f)||^2}\,. \quad (19)$$

It is in the range $0 \le \mathit{powRatio}_i \le 1$ by definition. It is close to 1 if the i-th signal term $\mathbf{a}_i(f)y_i(n,f)$ is dominant in the decomposition (6) of the mixtures $\mathbf{x}(n,f)$. In contrast, it is close to 0 if other signals $\mathbf{a}_{i'}(f)y_{i'}(n,f)$ are dominant. For speech/audio signals, there are many cases where only one signal is dominant due to their sparseness property [21]. Figure 4 shows the dominance measure of two separated signals at the same two frequencies as in Fig. 3. The correlation coefficients are

$$\begin{bmatrix} \rho(v_1^f, v_1^g) & \rho(v_1^f, v_2^g) \\ \rho(v_2^f, v_1^g) & \rho(v_2^f, v_2^g) \end{bmatrix} = \begin{bmatrix} 0.54 & -0.54 \\ -0.54 & 0.54 \end{bmatrix},$$

which are sufficiently high for the same source.

As summarized in [14], dominance measures (19) experimentally resulted in higher correlation coefficients than envelopes (18) for the same source among many frequencies. One reason for this is that envelopes have wide dynamic range, and active time frames are represented with various values. On the other hand, dominance measures have the limited dynamic range (from 0 to 1), and active time frames are uniformly represented with values close to 1 as long as the sources roughly satisfy the sparseness property.

4.2. Permutation Optimization

In the optimization procedure described here, we determine a permutation $\Pi_f : \{1,\ldots,N\} \rightarrow \{1,\ldots,N\}$ for each frequency f such that the output y_i,

$$i = \Pi_f(k)\,, \quad (20)$$

is grouped together for the k-th source. Although the notation Π_f used here is different from the matrix notation $\mathbf{P}(f)$ in (7), they represent the same permutation if we describe the matrix as

$$\mathbf{P}(f) = \begin{bmatrix} \mathbf{e}_{\Pi_f(1)} \\ \vdots \\ \mathbf{e}_{\Pi_f(N)} \end{bmatrix} \quad (21)$$

where $\mathbf{e}_i$ is a row vector in which the i-th element is 1 and all the other elements are 0.

The permutations Π_f in (20) of all frequency bins f can be determined if the activity measures are clustered for each source by maximizing the correlation coefficients ρ. However, calculating all the possible pair-wise correlation coefficients is computationally heavy. Thus, we practically perform rough global optimization, where the centroid c_k of each cluster is explicitly identified and accordingly the cost function

$$\mathcal{J}(\{c_k\},\{\Pi_f\}) = \sum_{f\in\mathcal{F}}\sum_{k=1}^N \rho(v_i^f, c_k)\,\Big|_{i=\Pi_f(k)} \quad (22)$$

is maximized. The set $\mathcal{F}$ consists of all frequency bins. The centroid c_k is calculated for each source as the average value of the activity measures with the current permutation Π_f:

$$c_k(n) \leftarrow \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}} v_i^f(n)\,\Big|_{i=\Pi_f(k)}, \quad \forall\, k,n\,, \quad (23)$$

where $|\mathcal{F}|$ is the number of elements in the set $\mathcal{F}$. The permutation Π_f is optimized to maximize the correlation coefficients ρ between the activity sequences v_i^f and the current centroid:

$$\Pi_f \leftarrow \mathrm{argmax}_\Pi \sum_{k=1}^N \rho(v_i^f, c_k)\,\Big|_{i=\Pi(k)}\,. \quad (24)$$

These two operations (23) and (24) are iterated until convergence.

According to the cost function (22), one centroid c_k is identified for each source k. This means that we expect similar activity sequences for all the frequencies. However, if we increase the sampling rate, for example up to 16 kHz, the activity sequences are significantly different between low and high frequency ranges. In order to precisely model such source signals, we introduce multiple centroids for a source, and extend the cost function as

$$\mathcal{J}(\{c_{k,m}\},\{\Pi_f\}) = \sum_{f\in\mathcal{F}}\sum_{k=1}^N \max_m \rho(v_i^f, c_{k,m})\,\Big|_{i=\Pi_f(k)}\,, \quad (25)$$

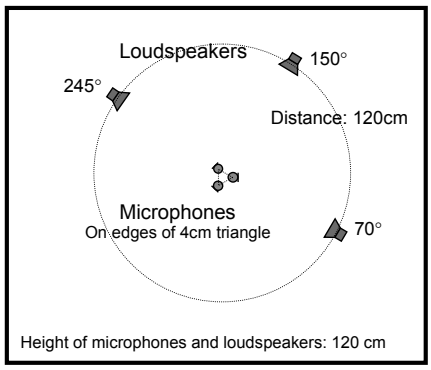

Fig. 5. Experimental condition

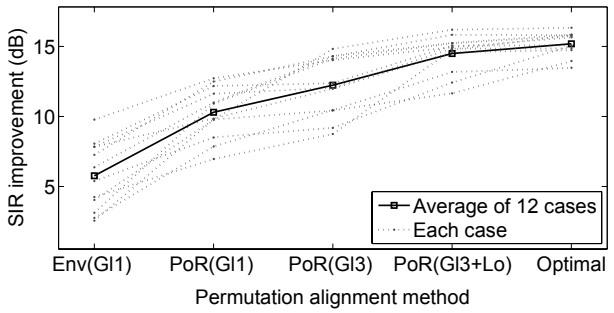

Fig. 6. Signal-to-interference ratio (SIR) improvements with several permutation alignment methods

where $c_{k,m}$ is the m-th centroid for source k. Practically, we have two or three centroids ($m = 1, 2$ or $m = 1, 2, 3$) for each source.

In this multiple centroid version, the centroids $c_{k,m}$ are obtained as another level of clustering centroids. The clustering is performed for the activity sequences of all frequencies that belong to the k-th source

$$v_i^f(n) \,\big|_{i=\Pi_f(k)}, \forall f \in \mathcal{F}\,. \tag{26}$$

Typically, k-means algorithm [22] or EM algorithm can be employed for this level of clustering. As regards the permutation of each frequency, it is optimized by

$$\Pi_f \leftarrow \mathrm{argmax}_\Pi \sum_{k=1}^{N} \mathrm{max}_m \rho(v_i^f, c_{k,m}) \,\big|_{i=\Pi(k)}\,, \tag{27}$$

which is also extended from (24). We interleave the multiple centroid calculation for (26) and the permutation optimization by (27) until convergence.

After the rough global optimization just described, we can perform fine local optimization [14] for better permutation alignment.

5. EXPERIMENTS

5.1. Private Experimental Condition

We performed experiments to separate three sources (male speech, female speech and music) with three microphones. We measured impulse responses h_{jk} under the condition shown in Fig. 5. Mixtures at the microphones were made by convolving the impulse responses and 3-second source signals. The separation performance was evaluated in terms of signal-to-interference ratio (SIR) improvement. The improvement was calculated by $\mathsf{OutputSIR}_i - \mathsf{InputSIR}_i$ for each output i, and we took the average over all outputs. These two types of SIRs are defined by

$$\mathsf{InputSIR}_i = 10\log_{10} \frac{\sum_t |\sum_l \mathrm{h}_{Ji}(l\cdot t_s)\mathrm{s}_i(t-l\cdot t_s)|^2}{\sum_t |\sum_{k\neq i}\sum_l \mathrm{h}_{Jk}(l\cdot t_s)\mathrm{s}_k(t-l\cdot t_s)|^2} \quad \text{(dB)},$$

$$\mathsf{OutputSIR}_i = 10\log_{10} \frac{\sum_t |\mathrm{y}_{ii}(t)|^2}{\sum_t |\sum_{k\neq i} \mathrm{y}_{ik}(t)|^2} \quad \text{(dB)},$$

where $J \in \{1, \ldots, M\}$ is the index of one selected reference sensor, and $\mathrm{y}_{ik}(t)$ is the component of s_k that appears at output $\mathrm{y}_i(t)$, i.e. $\mathrm{y}_i(t) = \sum_{k=1}^N \mathrm{y}_{ik}(t)$.

Table 1. SIR improvement and computational time with several ICA procedures (1st step: whitening, 2nd step: FastICA, 3rd step: maximum likelihood). The number of iterations for (16) in the 3rd step was 30 unless otherwise specified.

	1st+2nd+3rd	1st+2nd	1st+3rd	1st+3rd (250 ite.)
SIR impr.	14.50 dB	14.29 dB	10.28 dB	14.51 dB
Comp. time	4.3 s	1.8 s	2.7 s	20.2 s

Experiments were conducted with 12 combinations of 3 sources. Figure 6 shows the SIR improvements obtained with several permutation alignment methods. The abbreviations "Env" and "PoR" indicate the methods using envelopes $|y_i|$ in (18) and dominance measures $powRatio_i$ in (19), respectively. The abbreviations "Gl1" and "Gl3" correspond to the rough global optimization, where the number of centroids for each source in (25) is described by the number 1 or 3. "Lo" corresponds to the local optimization presented in [14]. The entry "Optimal" represents the results with the optimal permutations calculated with a knowledge of the original source signals. By comparing the results of "Env(Gl1)" and "PoR(Gl1)", the advantage of using dominance measures $powRatio_i$ instead of envelopes $|y_i|$ is clearly observed. If we increased the number of centroids to precisely model the activity of a source "PoR(Gl3)", the results were then improved in most cases. Together with the local optimization "PoR(Gl3+Lo)", the proposed method achieved good results which were very close to the "Optimal" results.

Table 1 compares several ICA procedures to see the effectiveness of the 3-step procedure described in Sect. 3. For permutation alignment, the method "PoR(Gl3+Lo)" was employed. The combination of all the three steps "1st+2nd+3rd" attained a good separation with an efficient computational cost. The total computational time (including STFT, permutation and so on) was around 8 seconds with this ICA procedure. The program was coded in Matlab and run on Athlon 64 FX-53. Without the 3rd step, the computational time was smaller but there was slightly less SIR improvement. When we combined the 1st and 3rd steps, 30 iterations of the update (16), starting from the whitening matrix as an initial solution, were insufficient for us to obtain a good result. We needed many iterations (250 in this case) to obtain a good separation, which results in a large computational time.

Table 2. Settings for MLSP 2007 Competition

Numbers of sources and sensors	$N = 2$, $M = 2$
Source signals	Male speech and Music
Sampling rate	f_s = 11.025 kHz
Duration of observations	30000 points (2.72 s)
STFT frame size L	2048 points (185.76 ms)
STFT frame shift S	256 points (23.22 ms)
Time-domain filter length L	2048 points
Permutation alignment method	PoR(GI3+Lo)
Computational time	6.14 s

5.2. MLSP 2007 Data Analysis Competition

Table 2 summarizes the settings for the MLSP 2007 Data Analysis Competition [1]. Time-domain filters w_{ij} were to be submitted for linear systems to separate the mixtures by

$$\mathrm{y}_i(t) = \sum_{j=1}^{M} \sum_{l=0}^{L-1} \mathrm{w}_{ij}(l \cdot t_s)\, \mathrm{x}_j(t - l \cdot t_s), \quad i = 1, \ldots, N. \qquad (28)$$

For this purpose, we needed to slightly modify the system flow described in Sect. 2. After the permutation and scaling ambiguities have been aligned by (5), (21), (10), separation filters w_{ij} were obtained by applying inverse DFT to $w_{ij}(f) = [\mathbf{W}(f)]_{ij}$:

$$\mathrm{w}_{ij}(l \cdot t_s) = \sum_{f \in \{0, \frac{1}{L} f_s, \ldots, \frac{L-1}{L} f_s\}} w_{ij}(f)\, e^{j2\pi f (l - \frac{L}{2}) t_s},$$

for $l = 0, \ldots, L-1$. Then, a synthesis window that tapers smoothly to zero at each end was multiplied to mitigate the edge effect: $\mathrm{w}_{ij}(l \cdot t_s) \leftarrow win_s(l \cdot t_s - \frac{L}{2} t_s) \cdot \mathrm{w}_{ij}(l \cdot t_s)$.

6. CONCLUSION

This paper presented the basic scheme of frequency-domain BSS, and then explained a complex-valued ICA procedure and a permutation alignment method in detail. As regards the permutation alignment method, we newly introduced multiple centroids for modeling the activity of separated signals more precisely. The experimental results showed the effectiveness of the new method for separating speech/music mixtures.

7. REFERENCES

[1] "MLSP 2007 Data Analysis Competition," 2007, http://mlsp2007.conwiz.dk/index.php?id=43.

[2] T.-W. Lee, *Independent Component Analysis - Theory and Applications*, Kluwer Academic Publishers, 1998.

[3] S. Haykin, Ed., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*, John Wiley & Sons, 2000.

[4] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[5] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley & Sons, 2002.

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

[7] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[8] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. ICA 2000*, June 2000, pp. 215–220.

[9] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, Oct. 2001.

[10] 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," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[11] 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 Signal Processing*, vol. 2003, no. 11, pp. 1135–1146, Nov. 2003.

[12] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. ICA 2006 (LNCS 3889)*. Mar. 2006, pp. 601–608, Springer.

[13] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech and Language Processing*, pp. 70–79, Jan. 2007.

[14] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. ISCAS 2007*, 2007, pp. 3247–3250.

[15] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice Hall, 2000.

[16] T. Adali and H. Li, "A practical formulation for computation of complex gradients and its application to maximum likelihood ICA," in *Proc. ICASSP 2007*, Apr. 2007, vol. II, pp. 633–636.

[17] T. Lotter, "Single- and multi-microphone spectral amplitude estimation using a super-gaussian speech model," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds., pp. 67–95. Springer, 2005.

[18] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[19] J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Letters*, vol. 4, no. 4, pp. 112–114, Apr. 1997.

[20] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems*. 1996, vol. 8, pp. 757–763, The MIT Press.

[21] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[22] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley Interscience, 2nd edition, 2000.