CONVOLUTIVE BLIND SOURCE SEPARATION FOR MORE THAN TWO SOURCES IN THE FREQUENCY DOMAIN

Hiroshi Sawada *Ryo Mukai* *Shoko Araki* *Shoji Makino*

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{sawada,ryo,shoko,maki}@cslab.kecl.ntt.co.jp

ABSTRACT

Blind source separation (BSS) for convolutive mixtures can be efficiently achieved in the frequency domain, where independent component analysis is performed separately in each frequency bin. However, frequency-domain BSS involves a permutation problem, which is well known as a difficult problem, especially when the number of sources is large. This paper presents a method for solving the permutation problem, which works well even for many sources. The successful solution for the permutation problem highlights another problem with frequency-domain BSS that arises from the circularity of discrete frequency representation. This paper discusses the phenomena of the problem and presents a method for solving it. With these two methods, we can separate many sources with a practical execution time. Moreover, real-time processing is currently possible for up to three sources with our implementation.

1. INTRODUCTION

Blind source separation (BSS) [1] is a technique for estimating original source signals solely from their mixtures at sensors. Its potential audio signal applications include teleconferences, voice control and hearing aids. In such applications, signals are mixed in a convolutive manner with reverberations. This makes the BSS problem much more difficult than the instantaneous mixture problem. Let us formulate the convolutive BSS problem. Suppose that N source signals $s_k(t)$ are mixed and observed at M sensors

$$x_j(t) = \sum_{k=1}^{N} \sum_l h_{jk}(l) s_k(t-l),$$

where $h_{jk}(l)$ represents the impulse response from source k to sensor j. The goal is to obtain N output signals $y_i(t)$, each of which is a filtered version of a source $s_i(t)$. If we have enough sensors ($M \geq N$), a set of FIR filters $w_{ij}(l)$ of length L is typically used to produce separated signals

$$y_i(t) = \sum_{j=1}^{M} \sum_{l=0}^{L-1} w_{ij}(l) x_j(t-l)$$

at the outputs, and independent component analysis (ICA) [2, 3] is generally used to obtain the FIR filters $w_{ij}(l)$. We can classify the BSS methods into two categories based on how we apply ICA for convolutive mixtures.

The first is time-domain BSS, where ICA is applied directly to the convolutive mixture model [4, 5]. It provides good separation once the algorithm converges, and is easy to extend to more than two sources. However, ICA for convolutive mixtures is not so simple as ICA for instantaneous mixtures, and computationally expensive for long filters.

The other approach is frequency-domain BSS, where complex-valued ICA for an instantaneous mixture is applied in each frequency bin [6–15]. The merit of this approach is that the ICA algorithm can be performed separately at each frequency, and the convergence of each ICA is fast. However, the permutation ambiguity of an ICA solution becomes a serious problem. We need to align the permutation in each frequency bin so that a separated signal in the time domain contains frequency components from the same source. This is the well-known permutation problem. Although various methods have been proposed for the permutation problem [6–10], most of them are applicable only for two sources or their performance deteriorates as the number of sources increases. Therefore, most of the published results with frequency-domain BSS were for only two sources.

This paper presents a frequency-domain BSS method that performs well even for more than two sources. The first key technique relates to solving the permutation problem and is discussed in Sec. 3. However, just solving the permutation problem does not provide good separation performance. We need to solve another problem that originates with the circularity of discrete frequency representation. This problem is not well known since it is not serious in a two-source case but it becomes serious as the number of sources increases. We discuss the phenomena and the reason for this problem and present an approach for its solution in Sec. 4. The effectiveness of the presented methods is shown by experimental results for up to four sources in Sec. 5. We also report the result of real-time processing for three sources, where the system can track moving sources.

2. FREQUENCY-DOMAIN BSS

This section describes frequency-domain BSS whose flow is shown in Fig. 1. First, time-domain signals $x_j(t)$ at sensors are converted into frequency-domain time-series signals $X_j(f,t)$ by short-time Fourier transform (STFT), where t is now down-sampled with the distance of the frame shift. Then, to obtain the frequency responses $W_{ij}(f)$ of filters $w_{ij}(l)$, complex-valued ICA $\mathbf{Y}(f,t) = \mathbf{W}(f)\mathbf{X}(f,t)$ is solved, where $\mathbf{X}(f,t) = [X_1(f,t), \ldots, X_M(f,t)]^T$, $\mathbf{Y}(f,t)$

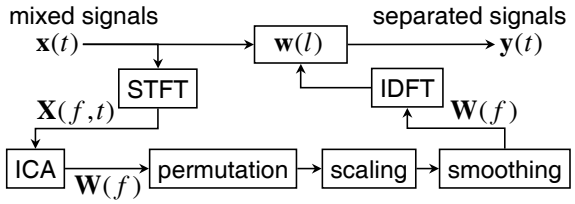

Fig. 1. Flow of frequency-domain BSS

$= [Y_1(f,t), \ldots, Y_N(f,t)]^T$ and $\mathbf{W}(f)$ is an $N \times M$ separation matrix whose elements are $W_{ij}(f)$. Any complex-valued ICA algorithm can be used in this scheme.

The ICA solution in each frequency bin has permutation and scaling ambiguity: even if we permute the rows of $\mathbf{W}(f)$ or multiply a row by a constant, it is still an ICA solution. The permutation ambiguity should be fixed so that $Y_i(f,t)$ at all frequencies correspond to the same source $s_i(t)$. Thus, the rows of $\mathbf{W}(f)$ are permuted $\mathbf{W}(f) \leftarrow \mathbf{P}(f)\mathbf{W}(f)$ by a permutation matrix $\mathbf{P}(f)$ obtained by a method, such as those discussed in Sec. 3. The scaling ambiguity is solved by the frequency-domain version of the minimal distortion principle, $\mathbf{W}(f) \leftarrow \mathrm{diag}[\mathbf{W}^{-1}(f)]\mathbf{W}(f)$, to make $Y_i(f,t)$ as close to $X_i(f,t)$ as possible [4, 8]. Then, we solve the circularity problem by the spectral smoothing described in Sec. 4. Finally, time-domain separation filters $w_{ij}(l)$ are obtained by applying inverse DFT to $W_{ij}(f)$.

3. THE PERMUTATION PROBLEM

Various methods have been proposed for solving the permutation problem. Let us begin with the direction of arrival (DOA) approach, where the DOAs of source signals are estimated to align permutations. The methods described in [9, 10] plot the directivity patterns formed by a separation matrix, and estimate the direction of a source as the minimum of a directivity pattern. In practice, the methods only work for two sources since the directivity patterns become too complicated to analyze for more than two sources.

We have proposed another way of estimating directions that works for any number of sources [11]. It first calculates the inverse $\mathbf{W}^{-1}(f)$, or the Moore-Penrose pseudoinverse $\mathbf{W}^{+}(f)$ if $N < M$, of the separation matrix $\mathbf{W}(f)$ obtained by ICA. Then, the direction θ_i of a source corresponding to the i-th row of $\mathbf{W}(f)$ is calculated by

$$\theta_i = \arccos \frac{\arg(\,[\mathbf{W}^{-1}]_{ji} \,/\, [\mathbf{W}^{-1}]_{j'i})}{2\pi f c^{-1}(d_j - d_{j'})}, \qquad (1)$$

where c is the propagation velocity and d_j is the position of sensor j. The scaling ambiguity of the ICA solution is eliminated by taking the ratio $[\mathbf{W}^{-1}]_{ji} \,/\, [\mathbf{W}^{-1}]_{j'i}$ of two elements from the same column. Figure 2 shows DOA estimations for mixtures of four sources obtained with (1). We see that directions are well estimated and a permutation matrix $\mathbf{P}(f)$ can be obtained by sorting the estimated directions at each frequency. However, at some frequencies (especially low frequencies), estimations are not obtained or are inaccurate. Therefore, the DOA approach alone does not provide a highly precise solution as shown at "D" in Fig. 6.

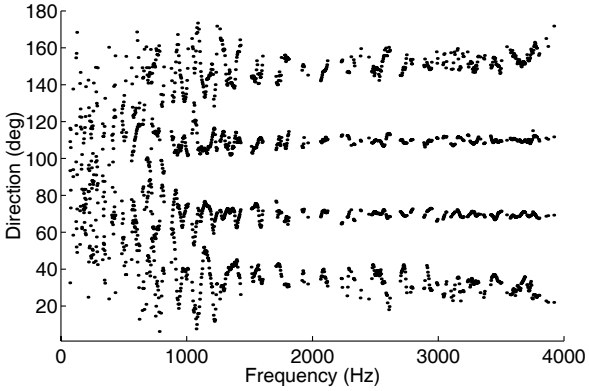

Fig. 2. DOA estimations for four sources using ICA

We also employ the correlation approach [7, 8] to align permutations more precisely. We use the envelope $v_i^f(t) = |Y_i(f,t)|$ of a separated signal $Y_i(f,t)$ to measure correlation. The correlation between two signals $x(t)$ and $y(t)$ is defined as $\mathrm{cor}(x,y) = (\mu_{x\cdot y} - \mu_x \cdot \mu_y)/(\sigma_x \cdot \sigma_y)$, where μ_x is the mean and σ_x is the standard deviation of x. Envelopes have high correlation at neighboring frequencies if separated signals correspond to the same signal. Let Π_f be a permutation corresponding to the inverse $\mathbf{P}^{-1}(f)$ of a permutation matrix $\mathbf{P}(f)$. A simple criterion for deciding Π_f is to maximize the sum of the correlations between neighboring frequencies within distance δ:

$$\Pi_f = \mathrm{argmax}_\Pi \sum_{|g-f|\le\delta} \sum_{i=1}^{N} \mathrm{cor}(v^f_{\Pi(i)}, v^g_{\Pi_g(i)}), \qquad (2)$$

where Π_g is the permutation at frequency g. This criterion is based on local information and has a drawback in that mistakes in a narrow range of frequencies may lead to the complete misalignment of the frequencies beyond the range. As shown at "C" in Fig. 6, the correlation approach alone does not provide a robust solution.

Our method effectively integrates these two approaches to solve the permutation problem robustly and precisely [11]. First, we decide permutations for frequency bins where the confidence of the DOA estimation is sufficiently high. Let $\mathcal{F}$ be the set of frequency bins where the permutation is already decided. Then, we apply (2) to frequency bins that are close neighbors with $f \in \mathcal{F}$. This procedure can avoid a consecutive misalignment. However, the permutations at low frequencies are not usually decided at this stage because the DOA estimations are unreliable as shown in Fig. 2. To decide permutations for these frequencies, we utilize the harmonic structure of a signal. If the signals are speech, there is a strong correlation between the envelopes of a frequency f and its harmonics $2f, 3f$ and so forth. Thus, we decide the permutation at frequency f with high confidence, if the sum shown below can be clearly maximized:

$$\Pi_f = \mathrm{argmax}_\Pi \sum_{g=2f,3f,\ldots} \sum_{i=1}^{N} \mathrm{cor}(v^f_{\Pi(i)}, v^g_{\Pi_g(i)}).$$

Finally, we apply (2) again for frequencies where the permutation is not yet decided.

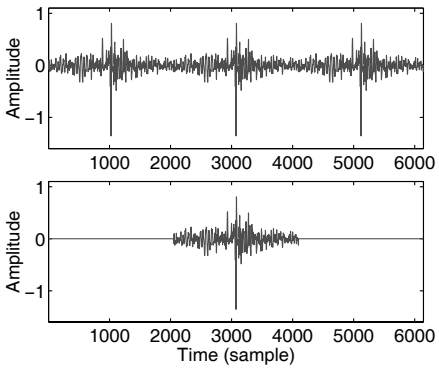

Fig. 3. Periodical time-domain filter represented by frequency responses sampled at $L = 2048$ points (above) and its one-period realization (below).

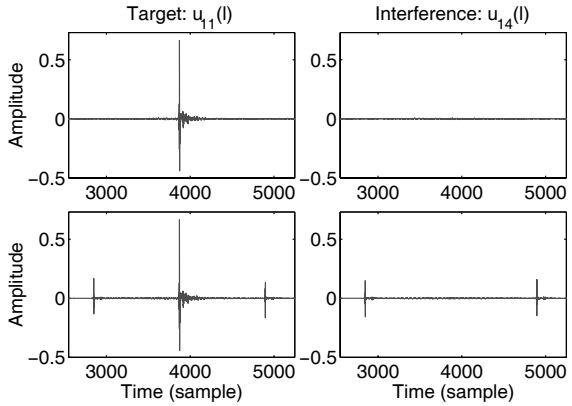

Fig. 4. Impulse responses $u_{ik}(l)$ obtained with the periodical filter (above) and with its one-period realization (below).

4. THE CIRCULARITY PROBLEM

The frequency-domain BSS described in Sec. 2 is influenced by the circularity of discrete frequency representation. The circularity refers to the fact that frequency responses sampled at L points with an interval f_s/L (f_s: sampling frequency) represent a periodical time-domain signal whose period is L/f_s. Figure 3 shows two time-domain filters. The upper one is a periodical infinite-length filter represented by frequency responses $W_{ij}(f)$ calculated by ICA at L points. Since this filter is unrealistic, we usually use its one-period realization shown in the lower part.

However, such one-period filters may cause a problem. Figure 4 shows impulse responses from a source $s_k(t)$ to an output $y_i(t)$: $u_{ik}(l) = \sum_{j=1}^{M} \sum_{\tau=0}^{L-1} w_{ij}(\tau) h_{jk}(l-\tau)$. Those on the left $u_{11}(l)$ correspond to the extraction of a target signal, and those on the right $u_{14}(l)$ correspond to the suppression of an interference signal. The upper responses are obtained with the infinite-length filters, and the lower ones with the one-period filters. We see that the one-period filters create spikes, which distort the target signal and degrade the separation performance.

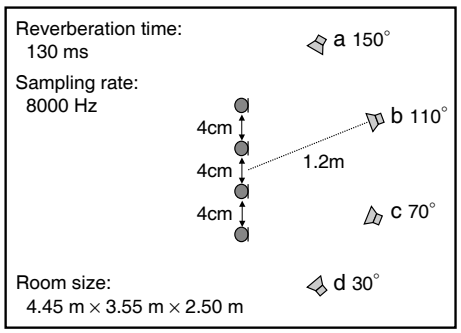

Fig. 5. Experimental conditions

Here, we consider two reasons for these spikes. One is that the frequency responses are under-sampled and the corresponding time-domain filter has an overlap with another period. ICA solutions separately obtained in frequency bins generally require the time-domain filters to be longer than L. The other reason is that adjacent periods work together to perform some filtering even if the first problem is solved. The effect of the second problem can be mitigated if the amplitude of the filter coefficients around both ends is small. It might be thought that a sufficiently large L would solve these problems. However, an excessively long STFT frame results in fewer samples at each frequency and worse ICA solutions [12].

Our approach to this problem involves controlling the frequency responses $W_{ij}(f)$ so that the corresponding time-domain filter $w_{ij}(l)$ fits length L and has small amplitude around the ends. This is carried out by windowing $w_{ij}(l) \cdot g(l)$ with a window $g(l)$ that tapers smoothly to zero at each end, such as a Hanning window. With this operation, frequency responses $\mathbf{W}(f)$ obtained by ICA are smoothed as $\mathbf{W}(f) \leftarrow \sum_{\phi=0}^{f_s-\Delta f} G(\phi)\mathbf{W}(f-\phi)$, where $G(f)$ is the frequency response of $g(l)$ and $\Delta f = f_s/L$. If a Hanning window is used, the frequency responses are smoothed as $\mathbf{W}(f) \leftarrow [\mathbf{W}(f-\Delta f) + 2\mathbf{W}(f) + \mathbf{W}(f+\Delta f)]/4$. The windowing successfully eliminates the spikes. However, it changes the frequency response obtained by ICA and causes an error. Thus, we minimize the error by adjusting the scaling of the ICA solution before windowing. See [13] for the details of the error and how to minimize it.

5. EXPERIMENTAL RESULTS

We performed experiments to separate speech signals in an environment whose conditions are summarized in Fig. 5. We tested cases of two, three and four sources whose positions are indicated in Table 1. The sensors were arranged linearly, and the number of sensors used was the same as the number of sources. We used filters of length $L = 2048$ because this length performed the best under the conditions. The ICA algorithm used was FastICA [3] followed by InfoMax combined with the natural gradient [2]. The results shown in Table 1 are the average of eight combinations of 7-second speeches. The signal-to-interference ratio (SIR) at

Table 1. Overall results for the batch processing

#sources / position	2 / a b		3 / a b c		4 / a b c d	
spectral smoothing	no	yes	no	yes	no	yes
SIR (dB)	18.0	18.7	13.0	14.4	9.5	12.0
execution time (s)	9.9	9.9	18.7	18.8	27.8	27.9

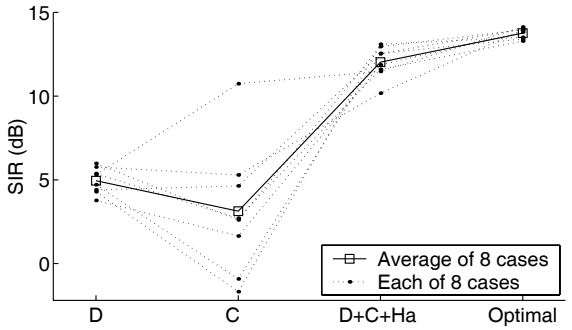

Fig. 6. Separation performance for four sources with different methods for the permutation problem

output i is calculated as the ratio of the power of a target component $\sum_l u_{ii}(l)s_i(t-l)$ and interference components $\sum_{k\neq i}\sum_l u_{ik}(l)s_k(t-l)$. We see that the spectral smoothing discussed in Sec. 4 improves the average SIR in every setup, especially with three and four sources.

Figure 6 shows separation performance for four sources with different methods for the permutation problem discussed in Sec. 3: "D" is the DOA approach alone, "C" is the correlation approach alone, "D+C+Ha" is the proposed method, and "Optimal" is the optimal solution obtained by utilizing the information of $s_k(t)$ and $h_{jk}(l)$. We see that the performance of "D" was stable but insufficient, the performance of "C" was unstable, and "D+C+Ha" performed very well and close to "Optimal".

Fast processing and convergence is one of the advantages of frequency-domain BSS. By using shorter filters $L = 1024$ and decreasing the number of iterations in the natural gradient, the BSS system performs in real-time for three sources. We used the same system structure as the one for two sources described previously [14]. Figure 7 shows the SIR for each source, where the source at position "d" started to move to position "c" at a time of 15 seconds. Since the filter coefficients were updated every 3 seconds, the system tracked the moving and recovered the SIRs.

6. CONCLUSIONS

This paper presented effective methods for overcoming the two major problems of frequency domain BSS. We succeeded in separating many sources mixed in a real environment with a practical execution time. The results shown here were for up to four sources with linearly arranged sensors. We have also separated six sources with a planar array of eight sensors based on similar techniques [15].

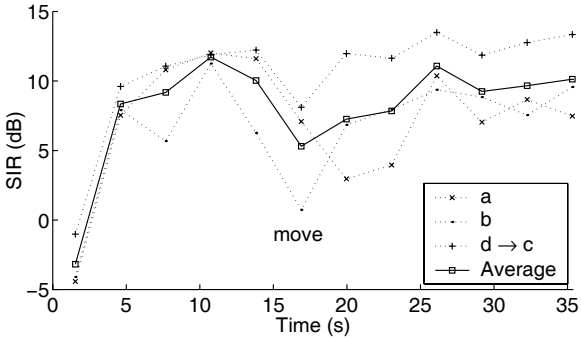

Fig. 7. Real-time processing for moving sources

7. REFERENCES

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

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

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

[4] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. ICA 2001*, Dec. 2001, pp. 722–727.

[5] S. C. Douglas and X. Sun, "Convolutive blind separation of speech mixtures using the natural gradient," *Speech Communication*, vol. 39, pp. 65–78, 2003.

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

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

[8] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, Oct. 2001.

[9] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. ICASSP 2000*, June 2000, pp. 3140–3143.

[10] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. ICASSP 2002*, May 2002, pp. 881–884.

[11] 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," in *Proc. ICA2003*, Apr. 2003, pp. 505–510.

[12] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.

[13] H. Sawada, R. Mukai, S. de la Kethulle, S. Araki, and S. Makino, "Spectral smoothing for frequency-domain blind source separation," in *Proc. IWAENC 2003*, Sept. 2003, pp. 311–314.

[14] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Robust real-time blind source separation for moving speakers in a room," in *Proc. ICASSP 2003*, Apr. 2003, pp. 469–472.

[15] R. Mukai, H. Sawada, S. de la Kethulle, S. Araki, and S. Makino, "Array geometry arrangement for frequency domain blind source separation," in *Proc. IWAENC 2003*, Sept. 2003, pp. 219–222.