Measuring Dependence of Bin-wise Separated Signals for Permutation Alignment in Frequency-domain BSS

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

Abstract— **This paper presents a new method for grouping bin-wise separated signals for individual sources, i.e., solving the permutation problem, in the process of frequency-domain blind source separation. Conventionally, the correlation coefficient of separated signal envelopes is calculated to judge whether or not the separated signals originate from the same source. In this paper, we propose a new measure that represents the dominance of the separated signal in the mixtures, and use it for calculating the correlation coefficient, instead of a signal envelope. Such dominance measures exhibit dependence/independence more clearly than traditionally used signal envelopes. Consequently, a simple clustering algorithm with centroids works well for grouping separated signals. Experimental results were very appealing, as three sources including two coming from the same direction were separated properly with the new method.**

I. Introduction

For acoustic applications of blind source separation (BSS) or independent component analysis (ICA) [1]–[3], such as solving a cocktail party problem, signals are generally mixed in a convolutive manner with reverberations. 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

$$\mathrm{x}_j(t)=\sum_{k=1}^{N}\sum_{l}\mathrm{h}_{jk}(l)\,\mathrm{s}_k(t-l),\ \ j=1,\ldots,M, \tag{1}$$

where t represents time and $\mathrm{h}_{jk}(l)$ represents the impulse response from source k to sensor j. In a practical room situation, impulse responses $\mathrm{h}_{jk}(l)$ can have thousands of taps even with an 8 kHz sampling rate. This makes the convolutive BSS problem harder to solve compared with that of simple instantaneous mixtures.

Many methods have been proposed [4]–[14] for solving 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 $\mathrm{x}_j(t)$. In the frequency domain, the convolutive mixture (1) can be approximated as an instantaneous mixture at each frequency:

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

where n represents the time frame index, f represents frequency, 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. A filtering operation for generating separated signals is performed in each frequency bin f, and then an inverse STFT is applied to the bin-wise separated signals to construct time-domain separated signals.

In order to construct proper separated signals in the time domain, frequency-domain separated signals originating from the same source should be grouped together. This problem is known as the permutation problem of frequency-domain BSS, and various methods have been proposed for its solution. Early work [6], [7] considered the smoothness of the frequency responses of the separation filters. Spatial information, such as the direction-of-arrivals of sources, can also be estimated and used [10]–[12]. For non-stationary sources such as speech, many methods [8]–[10], [13], [14] exploit the dependence of separated signals across frequencies. Of these, this paper considers the methods that employ two-stage processing [9], [10] where separation is first completed in each frequency bin by any BSS or ICA algorithm and then bin-wise separated signals are grouped together by measuring their dependence. This contrasts with the other methods [8], [13], [14] where the separation matrix in each frequency bin is adaptively updated according to the dependence of bin-wise signals across frequencies.

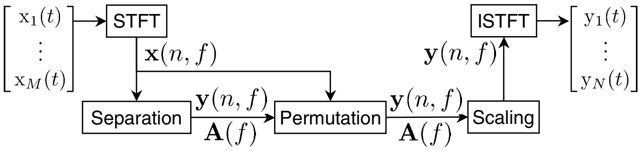


Fig. 1. System structure for frequency-domain BSS

This paper discusses how to evaluate the dependence of frequency-domain separated signals. Conventionally, the correlation coefficient of separated signal envelopes has been utilized [8]–[10]. Section III reviews the conventional method, and points out the drawback of using signal envelopes. Then, we propose a new measure in Sec. IV, which has good characteristics that help to represent dependence/independence more clearly. Section V presents two optimization methods for grouping bin-wise separated signals based on the above measure. The experimental results shown in Sec. VI are very encouraging. Section VII concludes this paper.

II. Frequency-domain BSS

This section presents an overview of frequency-domain BSS, which we consider in this paper. Figure 1 shows the system structure. First, sensor observations (1) sampled at frequency f_s 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:

$$x_j(n,f)\leftarrow\textstyle\sum_t \mathrm{x}_j(t)\,\mathrm{win}(t-n\frac{S}{f_s})\,e^{-\imath 2\pi f t}, \tag{3}$$

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 window $\mathrm{win}(t)$ is defined as non-zero only in the L-sample interval $[-\frac{L}{2}\frac{1}{f_s},\,(\frac{L}{2}-1)\frac{1}{f_s}]$ and preferably tapers smoothly to zero at each end of the interval.

Next, separation is performed in each frequency bin $f\in\mathcal{F}$:

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

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 $N\times M$

1-4244-0921-7/07 $25.00 © 2007 IEEE.

separation matrix. The set of frequencies $\mathcal{F}$ where the separation operation is performed can be limited to

$$\mathcal{F} = \{0, \dots, \frac{1}{2} f_s\} \tag{5}$$

due to the relationship of the complex conjugate:

$$x_j(n, \tfrac{m}{L} f_s) = x_j^*(n, \tfrac{L-m}{L} f_s)\,,\ \ m = 1, \dots, \tfrac{L}{2} - 1\,. \tag{6}$$

We can apply any instantaneous BSS/ICA algorithm [1]–[3] for the calculation of $\mathbf{W}$. Then, we calculate a matrix $\mathbf{A}$ whose columns are basis vectors $\mathbf{a}_i$,

$$\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_N],\ \mathbf{a}_i = [a_{1i}, \dots, a_{Mi}]^T, \tag{7}$$

in order to represent the vector $\mathbf{x}$ by a linear combination of the basis vectors:

$$\mathbf{x}(n, f) = \mathbf{A}(f)\,\mathbf{y}(n, f) = \sum_{i=1}^{N} \mathbf{a}_i(f) y_i(n, f)\,. \tag{8}$$

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

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

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

If the separation works well, we can expect the bin-wise separated signals $y_1(n, f), \dots, y_N(n, f)$ to be close to the original source signals $s_1(n, f), \dots, s_N(n, f)$ up to permutation and scaling ambiguity. The use of different subscripts in (2) and (8), i and k, indicates the permutation ambiguity. They should be related by a permutation $\Pi_f : \{1, \dots, N\} \to \{1, \dots, N\}$ for each frequency bin f as

$$i = \Pi_f(k) \tag{9}$$

so that the separated components y_i originating from the same source s_k are grouped together. The following sections will describe a procedure for deciding permutations Π_f. After Π_f are decided, the separated signals and the basis vectors are updated by

$$y_k(n, f) \leftarrow y_{\Pi_f(k)}(n, f),\ \ \mathbf{a}_k(f) \leftarrow \mathbf{a}_{\Pi_f(k)}(f),\ \ \forall\, k, n, f. \tag{10}$$

Next, the scaling ambiguity is aligned by adjusting $y_k(n, f)$ to the observation $x_J(n, f)$ of a selected reference sensor $J \in \{1, \dots, M\}$:

$$y_k(n, f) \leftarrow a_{Jk}(f) y_k(n, f),\ \ \forall\, k, n, f. \tag{11}$$

We see in (8) that $a_{Jk}(f) y_k(n, f)$ is a part of $x_J(n, f)$ that originates from source s_k.

Finally, time-domain output signals $\mathrm{y}_k(t)$ are calculated with an inverse STFT (ISTFT) of the separated signals $y_k(n, f)$.

III. Conventional measure for dependence

The correlation coefficient of bin-wise separated signal envelopes has been used [8]–[10] for measuring their dependence, and solving the permutation problem. The envelope of a bin-wise separated signal y_i is calculated by

$$v_i^f(n) = |y_i(n, f)|\,. \tag{12}$$

It is real-valued and represents the signal activity. The correlation coefficient ρ between two real-valued sequences $v_i(n)$ and $v_j(n)$ is defined as

$$\rho(v_i, v_j) = \frac{r_{ij} - \mu_i \mu_j}{\sigma_i \sigma_j} \tag{13}$$

where

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

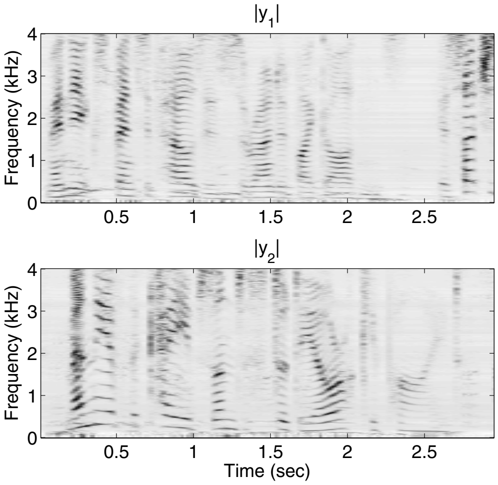


Fig. 2. Separated signal envelopes, normalized to zero mean and unit variance.

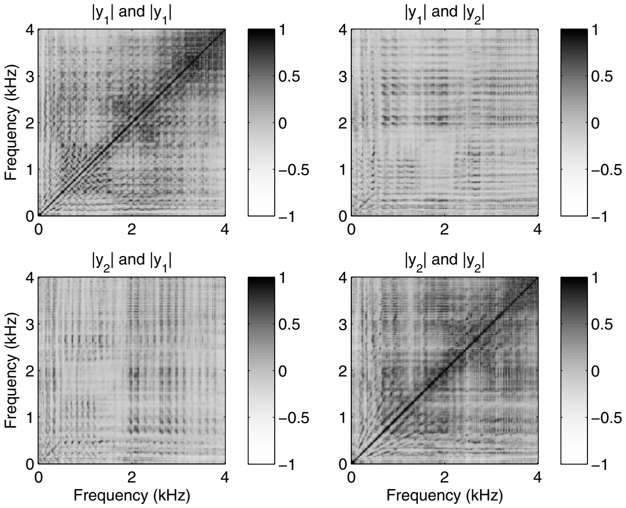


Fig. 3. Correlation coefficients between the separated signal envelopes shown in Fig. 2.

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.

We expect the correlation coefficient $\rho(v_i^f, v_k^g)$ of bin-wise separated signal envelopes to be high if they come from the same source. However, such a tendency is not always the case, and mostly can be seen only when the frequencies f and g are close together, or in a harmonic relationship such as $f \approx 2g$. Here is an example. Figure 2 shows the envelopes of successfully separated speech signals. Figure 3 shows the correlation coefficients between these envelopes. We observe that there are many frequency pairs f and g such that the envelopes are almost uncorrelated even if they belong to the same source: $\rho(v_1^f, v_1^g) \approx \rho(v_2^f, v_2^g) \approx 0$. This is due to the wide dynamic range of speech signals even if they are normalized to zero mean and unit variance, and active signals are represented with various values. This indicates the drawback of using signal envelopes.

IV. Proposed new measure

Instead of using envelopes (12) for calculating correlation coefficients (13), we propose using another type of measure that represents the dominance of the i-th separated signal in the observations. An example of such a measure is the power ratio between the i-th separated signal and the total power sum of the all separated signals:

$$\mathit{powRatio}_i(n,f) = \frac{||\mathbf{a}_i(f)\, y_i(n,f)||^2}{\sum_{k=1}^{N} ||\mathbf{a}_k(f)\, y_k(n,f)||^2}\,. \tag{14}$$

It is in the range $0 \leq \mathit{powRatio}_i \leq 1$ 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 (8) 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 signals, there are many cases where one signal is dominant due to their sparseness property.

Figure 4 shows the $\mathit{powRatio}_i$ values for the same separated signals as those shown in Fig. 2. Two characteristics should be stressed here. First, the dynamic range of the value is bounded, and active signals are uniformly represented with values close to 1. Second, the values of different sources are exclusive to each other, i.e., if $\mathit{powRatio}_1(n,f)$ is close to 1, then $\mathit{powRatio}_2(n,f)$ is close to 0. These two characteristics help us to measure the dependence clearly, as demonstrated below.

Let us here employ

$$v_i^f(n) = \mathit{powRatio}_i(n,f) \tag{15}$$

instead of envelopes (12) for calculating correlation coefficients (13). Figure 5 shows the correlation coefficients between such values. We see that the correlation coefficients are high for the same source more often than those in Fig. 3. Moreover, the correlation coefficients tend to be negative for different sources, due to the exclusive property mentioned above.

In summary, we can measure the dependence of bin-wise separated signals y_i more clearly by calculating correlation coefficients ρ with $\mathit{powRatio}_i$ values (14) rather than with envelopes $|y_i|$ (12).

V. Optimization

This section presents two optimization techniques for permutation alignment. The **first** technique is rough global optimization, which maximizes the cost function

$$\mathcal{J}(\{c_k\}, \{\Pi_f\}) = \sum_{f \in \mathcal{F}} \sum_{k=1}^{N} \rho(v_i^f, c_k) \,\big|_{i=\Pi_f(k)} \;. \tag{16}$$

The centroid c_k is calculated for each source as the average value of the measure (15) with the current permutation Π_f:

$$c_k(n) \leftarrow \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} v_i^f(n) \,\big|_{i=\Pi_f(k)}, \;\; \forall\, k, n\,, \tag{17}$$

where $|\mathcal{F}|$ is the number of elements in the set $\mathcal{F}$. The permutations Π_f are optimized to maximize the correlation coefficients ρ between the measures (14) and the current centroid:

$$\Pi_f \leftarrow \mathrm{argmax}_{\Pi} \sum_{k=1}^{N} \rho(v_i^f, c_k) \,\big|_{i=\Pi(k)}, \;\; \forall\, f \in \mathcal{F}\,. \tag{18}$$

These two operations (17) and (18) are iterated until convergence.

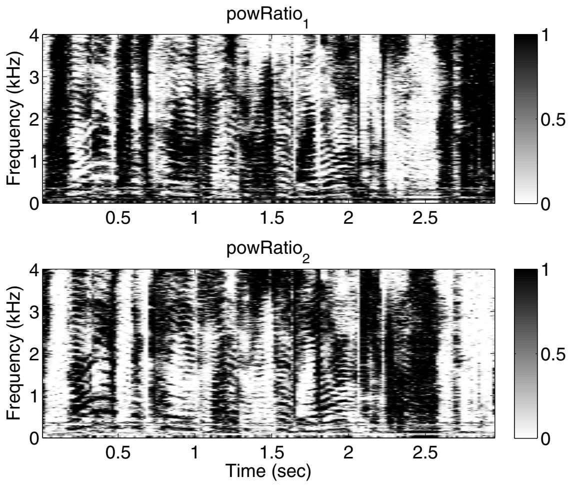


Fig. 4. The $\mathit{powRatio}_i$ values for the same separated signals whose envelopes are shown in Fig. 2.

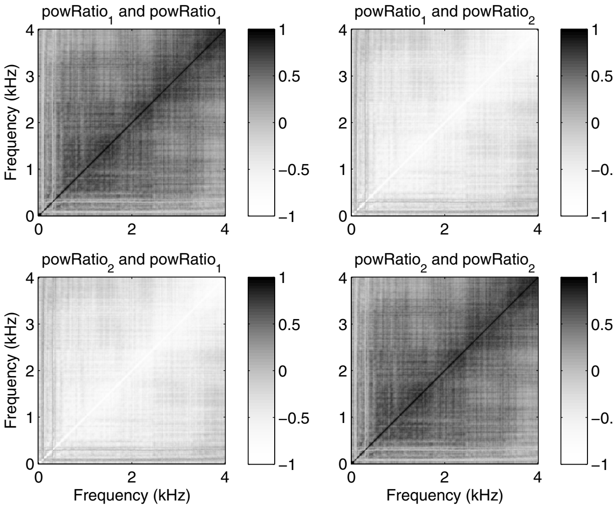


Fig. 5. Correlation coefficients between the $\mathit{powRatio}_i$ values shown in Fig. 4.

The **second** technique is for fine local optimization. It maximizes the sum of the correlation coefficients over a set of selected frequencies $\mathcal{R}(f)$ for a frequency f:

$$\Pi_f \leftarrow \mathrm{argmax}_{\Pi} \sum_{g \in \mathcal{R}(f)} \sum_{k=1}^{N} \rho(v_i^f, v_{i'}^g) \,\big|_{i=\Pi(k), i'=\Pi_g(k)} \;. \tag{19}$$

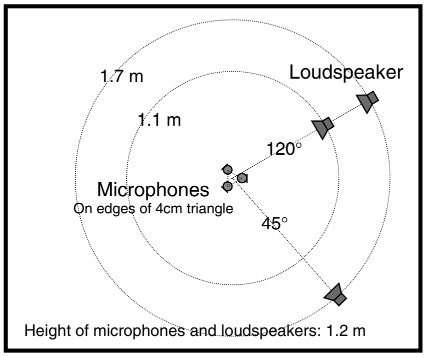
The set $\mathcal{R}(f)$ preferably consists of frequencies g where a high correlation coefficient $\rho(v_i^f, v_k^g)$ would be attained for v_i^f and $v_{i'}^g$ corresponding to the same source. We typically select adjacent frequencies $\mathcal{A}(f)$ and harmonic frequencies $\mathcal{H}(f)$ so that $\mathcal{R}(f) = \mathcal{A}(f) \cup \mathcal{H}(f)$. For example, $\mathcal{A}$ is given by

$$\mathcal{A}(f) = \{f-3\Delta f, f-2\Delta f, f-\Delta f, f+\Delta f, f+2\Delta f, f+3\Delta f\}$$

where $\Delta f = \frac{1}{L} f_s$, and $\mathcal{H}$ is given by

$$\begin{aligned}\mathcal{H}(f) = \quad & \{\mathit{round}(f/2)-\Delta f, \mathit{round}(f/2), \mathit{round}(f/2)+\Delta f, \\ & 2f-\Delta f, 2f, 2f+\Delta f\}\end{aligned}$$

where $\mathit{round}(\cdot)$ selects the nearest frequency to $\cdot$ from the set $\mathcal{F}$.



Source signals:
speeches of 3 s
Sampling rate:
8 kHz
Frame size of STFT:
1024 points (128 ms)
Frame shift of STFT:
256 points (32 ms)
Room size:
$4.45 \times 3.55 \times 2.5$ m
Reverberation time:
$\mathrm{RT}_{60} = 200$ ms

Fig. 6. Experimental conditions

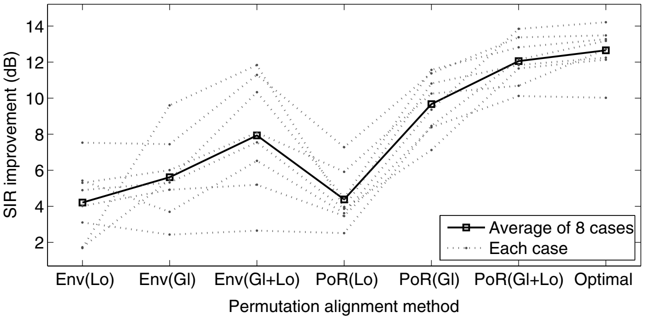


Fig. 7. SIR improvements with several permutation alignment methods.

The fine local optimization (19) is performed for a selected frequency f at a time, and repeated until no improvement is found for any frequency f.

VI. Experiments

We performed experiments to separate three speech sources with three microphones. We measured impulse responses $\mathrm{h}_{jk}(l)$ under the conditions shown in Fig. 6. Under this condition, two sources came from the same direction, and the sensor spacings were small. Thus, it was hard to exploit the spatial information of sources for permutation alignment as considered in [10]–[12]. Mixtures at the microphones were made by convolving the impulse responses and 3-second English speeches. The computational time was around 3 seconds for 3-second speech mixtures. The program was coded in Matlab and run on Athlon 64 FX-53. 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

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

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

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. $\mathrm{y}_i(t) = \sum_{k=1}^{N} \mathrm{y}_{ik}(t)$.

Experiments were conducted with 8 combinations of 3 speeches. Figure 7 shows the average SIR improvements obtained with several permutation alignment methods. The abbreviations "Env" and "PoR" indicate the methods using conventional [8]–[10] envelopes $|y_i|$ in (12) and the proposed dominance measure $\mathit{powRatio}_i$ in (14), respectively. The abbreviations "Gl" and "Lo" correspond to the rough global optimization and the fine local optimization presented in Sec. V, respectively. The entry "Optimal" represents the results with the optimal permutations calculated with a knowledge of the original source signals.

We observe the following. The rough global optimization "Gl" works well with the new measure "PoR", but not with the conventional measure of envelopes "Env". The fine local optimization "Lo" alone does not provide good results for either "Env" and "PoR". Its effectiveness can be seen in improving the moderately good solutions obtained by the global optimization "Gl". The proposed method "PoR(Gl+Lo)" achieved almost optimal results.

VII. Conclusion

This paper presented a novel method for solving the permutation problem of frequency-domain BSS. The newly proposed dominance measure (14) represents the activity of a separated signal y_i with the favorable characteristics described in Sec. IV. Thanks to these characteristics, the dependency of separated signals is clearly measured by using correlation coefficients (13). This allows the effective use of the global optimization technique proposed in Sec. V. The proposed method worked surprisingly well in the experiments, as two sources coming from the same direction were separated in a real room.

References

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

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

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

[4] S. C. Douglas, H. Sawada, and S. Makino, "A spatio-temporal FastICA algorithm for separating convolutive mixtures," in *Proc. ICASSP 2005*, vol. V, Mar. 2005, pp. 165–168.

[5] W. Kellermann, H. Buchner, and R. Aichner, "Separating convolutive mixtures with TRINICON," in *Proc. ICASSP 2006*, vol. V, May 2006, pp. 961–964.

[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] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency domain blind source separation using small and large spacing sensor pairs," in *Proc. ISCAS 2004*, vol. V, May 2004, pp. 1–4.

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

[14] 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.