# Blind separation and localization of speeches in a meeting situation

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

*Abstract*— **The technique of blind source separation (BSS) has been well studied. In this paper, we apply the BSS technique, particularly based on independent component analysis (ICA), to a meeting situation. The goal is to enhance the spoken utterances and to estimate the location of each speaker by means of multiple microphones. The technique may help us to take the minutes of a meeting.**

## I. Introduction

The theory and algorithms of blind source separation (BSS) and independent component analysis (ICA) became popular in the signal processing community, as many textbooks [1]–[3] have been published. One of the well recognized applications of ICA/BSS is the separation of speeches mixed in a real-room reverberant environment (i.e. solving a cocktail party problem). The difficulty of this problem lies in the fact that the mixing system is not simply instantaneous but convolutive. Thus, additional effort has been devoted to the separation of convolutive speech mixtures by many researchers in recent years, as reported in many papers [4]–[12]. Some of these papers describe effective methods that can attain successful separation, namely an improvement in the signal-to-interference ratio (SIR) of 10∼15 dB or more, for real recorded speech mixtures. However, the situations are commonly well set up so that these conditions are satisfied:

1) The mixing system is roughly time-invariant.
2) All speakers are active most of the time.
3) The number of speakers is known, and less than or equal to the number of microphones.

This paper reports a trial where BSS techniques are applied to a meeting situation. The goal is to obtain the enhanced spoken utterances of each speaker and to identify the location of each speaker from the signals observed at multiple microphones. It should be noted that the above conditions are not necessarily met in a meeting situation. First, it is hard to assume time-invariance of the mixing system for hundreds or thousands of seconds of observed signals. Our approach here is to apply block processing to signals of several seconds where time-invariance can be assumed. Second, speaker activity is non-stationary in a meeting, and the second condition in the above list is basically not met. If a person speaks continuously in one block and then falls silent throughout the next block, the separation filters of the BSS system change drastically. If such situations happen frequently, a speaker in the specific BSS output may change across blocks. We call this problem a block-wise permutation problem, and propose a solution based on speaker localization in this paper.

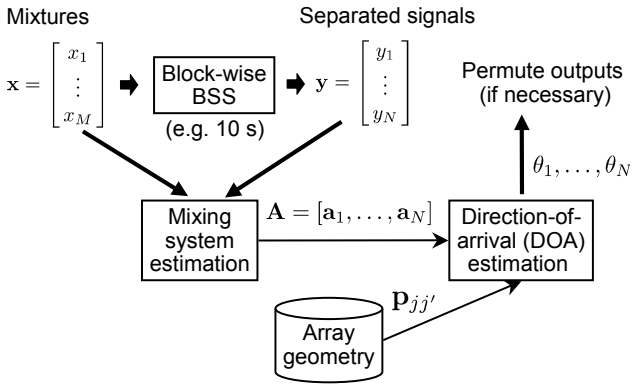


Fig. 1. Global flow of the method

As related work, there are many meeting room projects (e.g. [13]–[15]). The goal of such projects is to develop techniques for the transcription, summarization and understanding of a meeting. Of the many research topics in this field, speaker diarization [16], [17], which identifies when each participant speaks, is one that is strongly related to the goal of this paper. However, they generally assumed that at most one speaker is active at a time, and BSS techniques have not yet been substantially considered in these projects.

This paper is organized as follows. Section II presents an overview of the proposed method, which basically consists of block-wise BSS and speaker localization. An implementation of the block-wise BSS is presented in Sec. III. Since we employ a small size microphone array in this trial, where the microphone spacing is around 4 cm, we focus only on the direction of each speaker in the processing of speaker localization. Therefore, a method for estimating the direction-of-arrival (DOA) of each speaker is presented in Sec. IV. Also, the clustering of such estimated DOAs for aligning block-wise permutations is explained in the same section. Experimental results are shown in Sec. V, and Sec. VI concludes this paper.

## II. Global flow

Figure 1 shows the global flow of the proposed method. We apply block-wise BSS to several seconds (e.g. 10 seconds) of observed mixtures, to obtain separated signals for each

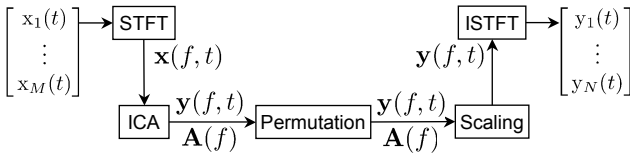


Fig. 2. Structure of frequency-domain BSS

block. As discussed in the introduction, there is a possibility that block-wise permutation occurs. The problem is solved by estimating the DOA for each speaker in each block. To accomplish this, we first estimate the mixing system from the BSS results. And then the DOA of each speech is estimated together with the array geometry. Finally, we permute the outputs of block-wise BSS if necessary so that the same speaker appears at the same BSS output.

## III. Block-wise BSS

We can apply any method/approach [4]–[12] to the block-wise BSS in Fig. 1. We here adopt a frequency-domain approach similar to [6]–[10] for convolutive BSS. This section explains the procedure in detail.

Figure 2 shows the system structure. First, sensor observations $\mathrm{x}_1(t), \ldots, \mathrm{x}_M(t)$ in the time domain sampled at frequency $f_s$ are converted into frequency-domain time-series signals $x_1(f,t), \ldots, x_M(f,t)$ by a short-time Fourier transform (STFT) with frame size $L$:

$$x_j(f,t) \leftarrow \sum_{q=-L/2}^{L/2-1} \mathrm{x}_j(t+q)\,\mathrm{win}(q)\,e^{-\imath 2\pi f q}, \qquad (1)$$

for all discrete frequencies $^\forall f \in \mathcal{F} = \{0, \frac{1}{L}f_s, \ldots, \frac{L-1}{L}f_s\}$, and for time $t$ which is now down-sampled with a distance equal to the frame shift. We typically use a window $\mathrm{win}(q)$ that tapers smoothly to zero at each end, such as a Hanning window $\mathrm{win}(q) = \frac{1}{2}(1 + \cos\frac{2\pi q}{L})$.

Next, ICA is employed in each frequency bin:

$$\mathbf{y}(f,t) = \mathbf{W}(f)\,\mathbf{x}(f,t), \quad ^\forall f \in \mathcal{F}, \qquad (2)$$

where $\mathbf{x} = [x_1, \ldots, x_M]^T$ is an observation vector, $\mathbf{y} = [y_1, \ldots, y_N]^T$ is a separated signal vector, and $\mathbf{W}$ is an $N \times M$ separation matrix. We can apply any instantaneous ICA [1]–[3] for the calculation of $\mathbf{W}$. Then, we calculate a matrix $\mathbf{A}$ whose columns are basis vectors $\mathbf{a}_i$,

$$\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_N],\ \mathbf{a}_i = [a_{1i}, \ldots, a_{Mi}]^T, \qquad (3)$$

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

$$\mathbf{x}(f,t) = \mathbf{A}(f)\,\mathbf{y}(f,t) = \sum_{i=1}^{N} \mathbf{a}_i(f) y_i(f,t), \quad ^\forall f \in \mathcal{F}. \qquad (4)$$

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

$$\mathbf{A} = \mathrm{E}\{\mathbf{x}\mathbf{y}^H\}(\mathrm{E}\{\mathbf{y}\mathbf{y}^H\})^{-1}, \qquad (5)$$

which minimizes $\mathrm{E}\{||\mathbf{x} - \mathbf{A}\mathbf{y}||^2\}$. The notation $\cdot^H$ represents the conjugate transpose of $\cdot$.

The next step is to solve the internal permutation problem, which is caused by the fact that ICA (2) is employed independently in each frequency bin and ICA solutions have permutation ambiguities. Many methods have been proposed [6]–[10] for the internal permutation problem. However, we employ a recently developed approach in the experiments, which is explained in detail in a paper under review [18]. Whatever method we use, we calculate a permutation $\Pi_f : \{1, \ldots, N\} \to \{1, \ldots, N\}$ for each frequency bin $f$, and then the separated signals and the basis vectors are permuted by

$$y_i(f,t) \leftarrow y_{\Pi_f(i)}(f,t), \quad \mathbf{a}_i(f) \leftarrow \mathbf{a}_{\Pi_f(i)}(f), \quad ^\forall i, f, t. \qquad (6)$$

so that the separated components $y_i$ originating from the same source have the same index at all frequency bins.

Next, scaling ambiguities of the ICA solutions are aligned by adjusting $y_i(f,t)$ to the observation $x_J(f,t)$ of a selected reference sensor $J \in \{1, \ldots, M\}$:

$$y_i(f,t) \leftarrow a_{Ji}(f) y_i(f,t), \quad ^\forall i, f, t. \qquad (7)$$

We see in (4) that $a_{Ji}(f) y_i(f,t)$ is a part of $x_J(f,t)$.

Finally, time-domain output signals $\mathrm{y}_i(t)$ are calculated by applying an inverse STFT (ISTFT) to the separated signals $y_i(f,t)$.

## IV. DOA estimation and clustering

This section presents a method for estimating the DOAs of sources from the result of block-wise BSS. The flow is depicted in the lower half of Fig. 1.

### A. Mixing system estimation

First, the mixing system is estimated from the BSS results. If we take the frequency-domain approach, the frequency responses of the mixing system have already been estimated as $\mathbf{A}(f)$ by (5). If we take another approach (e.g. time-domain BSS) for the block-wise BSS and thus obtain just the set of time-domain mixtures $\mathrm{x}_1(t), \ldots, \mathrm{x}_M(t)$ and separated signals $\mathrm{y}_1(t), \ldots, \mathrm{y}_N(t)$, the mixing system could be estimated as a matrix of filters $\mathbf{A}(l)$ by performing a least-mean-square estimation in the time domain. However, this is computationally demanding if the filters are long. An efficient alternative way is to convert the time-domain signals into frequency-domain time-series signals by STFT (1), and apply the same equation (5) to the frequency-domain vectors $\mathbf{x}(f,t)$ and $\mathbf{y}(f,t)$ to estimate $\mathbf{A}(f)$ for each frequency $f$.

As shown in (3), the columns of matrix $\mathbf{A}$ are called basis vectors. The $i$-th basis vector $\mathbf{a}_i = [a_{1i}, \ldots, a_{Mi}]^T$ is an estimation of the way in which the source $i$ was observed at all the sensors. This means that we can estimate the source directions by analyzing the basis vectors.

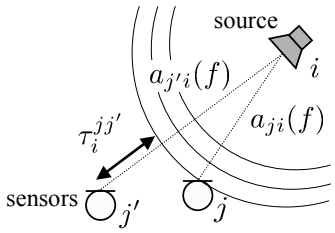


Fig. 3. Time-difference-of-arrival (TDOA) estimation

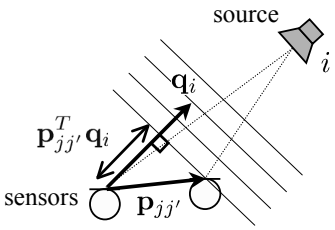


Fig. 4. DOA estimation with two sensors

### B. TDOA estimation

Before estimating the DOA of a source, we estimate the time-difference-of-arrival (TDOA) $\tau_i^{jj'}$ of source $i$ between sensors $j$ and $j'$ (Fig. 3). We employ the well-known GCC-PHAT (Generalized Cross Correlation PHAse Transform) function [19], [20] for TDOA estimations. In the original formulation, frequency-domain sensor observations $x_j(f,t)$ and $x_{j'}(f,t)$ are used in the GCC function. In our scenario, however, sensor observations may contain multiple source components, which disturb the estimation. On the other hand, the basis vectors $\mathbf{a}_i(f)$, $f \in \mathcal{F}$, of source $i$ represent information specific only to source $i$. Thus, we use basis vector elements $a_{ji}(f)$ and $a_{j'i}(f)$ corresponding to sensors $j$ and $j'$ instead. The formula

$$\tau_i^{jj'} = \text{argmax}_\tau \sum_f \frac{a_{ji}(f)a_{j'i}^*(f)}{|a_{ji}(f)a_{j'i}^*(f)|} e^{\imath 2\pi f \tau}, \tag{8}$$

where $\cdot^*$ denotes the complex conjugate of $\cdot$, estimates the TDOA of source $i$ between sensors $j$ and $j'$.

### C. DOA estimation with array geometry

Then, by incorporating the array geometry information, the DOA of source $i$ can be estimated. Let us define a DOA vector

$$\mathbf{q}_i = \begin{bmatrix} \cos\theta_i \cos\phi_i \\ \sin\theta_i \cos\phi_i \\ \sin\phi_i \end{bmatrix}$$

that has unit norm $||\mathbf{q}_i|| = 1$. The azimuth and elevation of the corresponding source $i$ are represented by $\theta_i$ and $\phi_i$, respectively.

If we start by considering a simple two-sensor case, such as that shown in Fig. 4, the DOA vector of source $i$ should satisfy

$$\mathbf{p}_{jj'}^T \mathbf{q}_i = \tau_i^{jj'} \cdot v \tag{9}$$

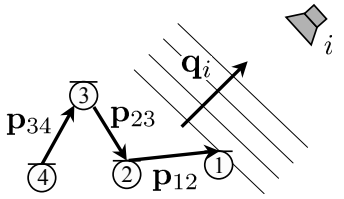


Fig. 5. DOA estimation with multiple sensor pairs

where $\mathbf{p}_{jj'}$ is a 3-dimensional vector representing the location of sensor $j$ relative to that of sensor $j'$, $\tau_i^{jj'}$ is the TDOA estimated by (8), and $v$ represents the signal velocity. The possible solutions for the DOA vector $\mathbf{q}_i$ in (9) represent a cone, not a direction [10].

If we have more sensor pairs that provide additional equations, the direction can be specified as the intersection of the cones given by the multiple equations. Figure 5 shows a situation where we have four sensors and take three sensor pairs into consideration. In such a case, we have the following simultaneous linear equations for the DOA vector $\mathbf{q}_i$:

$$\begin{bmatrix} \mathbf{p}_{12}^T \\ \mathbf{p}_{23}^T \\ \mathbf{p}_{34}^T \end{bmatrix} \mathbf{q}_i = \begin{bmatrix} \tau_i^{12} \\ \tau_i^{23} \\ \tau_i^{34} \end{bmatrix} v\,.$$

In most cases, such simultaneous linear equations do not have an exact solution for $\mathbf{q}_i$ because it is hard to obtain precise array geometry information, and hard to estimate TDOAs correctly in a real reverberant situation. Hence we practically compromise with an approximated solution. An efficient way [10] is to multiply the Moore-Penrose pseudo-inverse of the array geometry matrix

$$\mathbf{D} = \begin{bmatrix} \mathbf{p}_{12}^T \\ \mathbf{p}_{23}^T \\ \mathbf{p}_{34}^T \end{bmatrix}$$

to obtain

$$\mathbf{q}_i = \mathbf{D}^+ \begin{bmatrix} \tau_i^{12} \\ \tau_i^{23} \\ \tau_i^{34} \end{bmatrix} v\,,$$

and then normalize it to unit-norm

$$\mathbf{q}_i = \frac{\mathbf{q}_i}{||\mathbf{q}_i||}\,.$$

### D. DOA clustering and block-wise permutation alignment

So far, we have calculated DOA vectors $\mathbf{q}_i$ for each block and for each output of the block-wise BSS. Figure 6 shows an example of calculated DOAs. These data were obtained for the situation shown in Fig. 9, where we have four speakers and four microphones. The DOA estimations are simply represented by the azimuth values $\theta_i$ in degrees, and the elevations $\phi_i$ are all zero, since all the microphones were placed in a 2-dimensional space in this case. There are four different plot symbols, each of which corresponds to each of the BSS outputs. As seen in Fig. 6, the BSS outputs of the speakers interchange at some block transition points. This shows an example of the block-wise permutation problem.

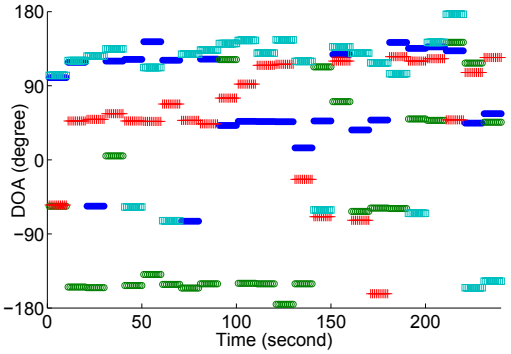


Fig. 6. Unsorted DOA estimations from the results of block-wise BSS.

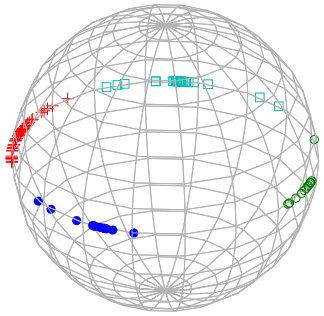

Fig. 7. Clustering DOA estimations into four classes

In order to align these block-wise permutations, we apply a clustering algorithm to the DOA vectors to identify the DOA cluster for each speaker. There are many clustering algorithms that can be used for this purpose. In this paper we employ a simple k-means algorithm [21], assuming that the number of speakers is known a priori. Figure 7 shows the clustering result applied to the DOA estimations shown in Fig. 6. In this figure, DOA vectors are plotted on a unit circle since the microphone array was 2-dimensional.

Based on the clustering result, we can permute the block-wise BSS outputs so that the same speaker appears at the same BSS output. Figure 8 shows the sorted DOA estimations. We observe that the misalignments found in Fig. 6 are reduced.

Sometimes, someone is silent for the whole time in a block, and consequently multiple DOA estimations belong to one cluster in the block. In such a case, we preserve only the DOA estimation whose corresponding separated signal is the loudest, and reassign the other DOA estimations to other clusters.

## V. Experiments

Figure 9 shows the meeting situation to which we applied the BSS system. We have four speakers and four microphones together with some noise sources such as desktop computers (PC) and a projector. Each of the four human speakers was sitting on a chair. The four microphones were arranged in the horizontal 2-dimensional plane about 4 cm apart. The reverberation time of the room was around $\mathrm{RT}_{60} = 350$ ms.

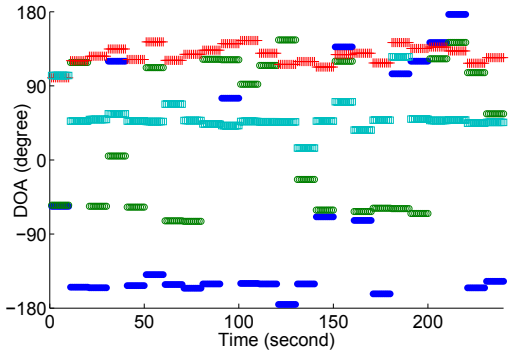


Fig. 8. Sorted DOA estimations from the results of block-wise BSS.

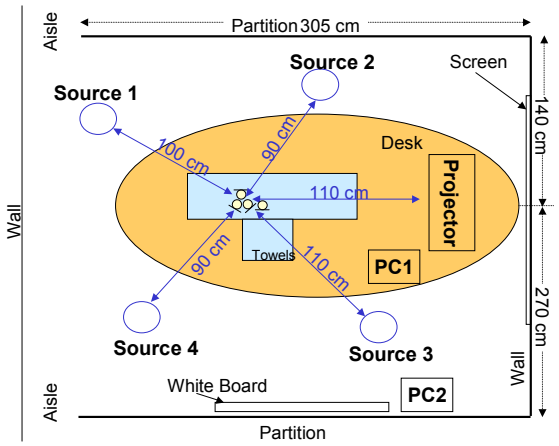


Fig. 9. Meeting situation

We recorded mixed sounds consisting of human speech and PC/projector noise as the input signals for the BSS system. We also measured the impulse response $\mathrm{h}_{ji}(l)$ from a source $i$ to a microphone $j$ for a quantitative evaluation of the separation performance in terms of signal-to-interference ratio (SIR) improvement. The improvement was calculated by $\mathsf{OutputSIR}_i - \mathsf{InputSIR}_i$ for each output $i$. 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 $\mathrm{s}_i$ is the $i$-th source, $J \in \{1,\ldots,M\}$ is the index of a selected reference sensor, and $\mathrm{y}_{ik}(t)$ is the component of $\mathrm{s}_k$ that appears at output $\mathrm{y}_i(t)$, i.e. $\mathrm{y}_i(t) = \sum_{k=1}^{N}\mathrm{y}_{ik}(t)$.

Examples of recorded mixtures and separated sounds are shown in Figs. 10 and 11, respectively. The utterances of the speakers were well separated, and also the PC/projector noise was reduced, as a result of BSS. In an informal listening test, however, we observed that some leaks could still be heard. Estimated DOAs for recorded mixtures have already

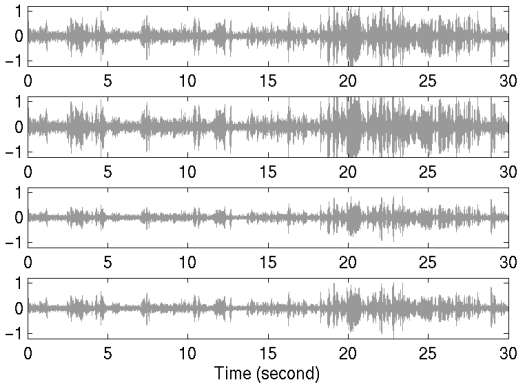


Fig. 10. Four channels of observed mixed signals at the four microphones in the meeting situation. 30 second excerpts.

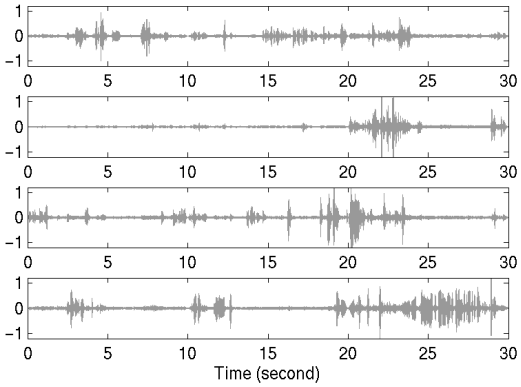


Fig. 11. Separated signals generated from the mixed signals shown in Fig. 10 with the BSS system.

been shown in Fig. 8, whose result was obtained through the intermediate results shown in Figs. 6 and 7.

An evaluation in terms of the SIR was conducted by using the measured impulse responses and speech signals drawn from a speech database. Table I shows the average SIR improvements for 8 different speech combinations. We performed tests using three different lengths (30, 60, and 90 seconds) of total data, while the block length for BSS was 10 seconds in all cases. To observe the effect of block-wise permutation, we compare the results that we obtained when block-wise permutations were aligned and untouched. As the total data length increases, the SIR improvement decreases in the untouched case. This is why the chance for block-wise permutation increases as the total data length increases.

TABLE I

SIR IMPROVEMENTS. AVERAGE OF 8 SPEECH COMBINATIONS.

| Block-wise permutation | 30 seconds | 60 seconds | 90 seconds |
|---|---|---|---|
| Aligned | 14.99 dB | 15.54 dB | 14.29 dB |
| Untouched | 9.87 dB | 7.98 dB | 7.62 dB |

## VI. CONCLUSION

This paper discussed the application of BSS techniques to a meeting situation. The proposed approach consists of block-wise BSS and DOA clustering. The separation performance is good in terms of SIR improvement (Table I) and with the waveforms (Fig. 11). However, we can still hear another speaker's utterances and their reverberations in the separated signals. Post-processing is a way to reduce such interference. Other future work will include an online implementation and tracking, a combination with a speech recognition system, and the application to underdetermined cases where the speakers outnumber the microphones.

## 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] 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] 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.

[10] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation of many speech signals using near-field and far-field models," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 83 683, 13 pages, 2006.

[11] 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.

[12] 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*, (accepted for future publication).

[13] [Online]. Available: http://nist.gov/speech/tests/rt/rt2002/

[14] [Online]. Available: http://www.icsi.berkeley.edu/Speech/mr/

[15] [Online]. Available: http://chil.server.de/servlet/is/101/

[16] D. Ellis and J. Liu, "Speaker turn segmentation based on between-channel differences," in *Proc. ICASSP 2004 NIST Meeting Recognition Workshop*, 2004, pp. 112–117.

[17] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. ICASSP 2005*, vol. V, Mar. 2005, pp. 953–956.

[18] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. 2007 IEEE International Symposium on Circuits and Systems (ISCAS 2007)*, 2007, (submitted).

[19] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustic, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[20] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 3, pp. 288–292, May 1997.

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