# A Region-Growing Permutation Alignment Approach in Frequency-Domain Blind Source Separation of Speech Mixtures

Lin Wang, Heping Ding, *Senior Member, IEEE*, and Fuliang Yin

***Abstract*—The convolutive blind source separation (BSS) problem can be solved efficiently in the frequency domain, where instantaneous BSS is performed separately in each frequency bin. However, the permutation ambiguity in each frequency bin should be resolved so that the separated frequency components from the same source are grouped together. To solve the permutation problem, this paper presents a new alignment method based on an inter-frequency dependence measure: the powers of separated signals. Bin-wise permutation alignment is applied first across all frequency bins, using the correlation of separated signal powers; then the full frequency band is partitioned into small regions based on the bin-wise permutation alignment result. Finally, region-wise permutation alignment is performed in a region-growing manner. The region-wise permutation correction scheme minimizes the spreading of the misalignment at isolated frequency bins to others, hence to improve permutation alignment. Experiment results in simulated and real environments verify the effectiveness of the proposed method. Analysis demonstrates that the proposed frequency-domain BSS method is computationally efficient.**

***Index Terms*—Blind source separation (BSS), convolutive mixture, frequency domain, permutation problem, power ratio, region growing.**

## I. Introduction

BLIND source separation (BSS) is a technique for recovering the original source signals from observed signals with the mixing process unknown [1]. BSS has a lot of potential applications including noise robust speech recognition, crosstalk separation in telecommunications, biomedical signal analysis, and so on. It has attracted considerable attention in research communities.

Manuscript received June 29, 2009; revised November 13, 2009, January 28, 2010, and May 07, 2010; accepted May 21, 2010. Date of publication July 12, 2010; date of current version December 03, 2010. This work was supported in part by the National Natural Science Foundation of China under Grants 60772161 and 60372082, in part by the Specialized Research Fund for the Doctoral Program of Higher Education of China under 200801410015, and in part by NRC-MOE Research and a Post-Doctoral Fellowship Program from Ministry of Education of China and National Research Council of Canada. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jingdong Chen.

L. Wang is with the School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China, and also with the Acoustics and Signal Processing Group, Institute for Microstructural Sciences, National Research Council Canada, Ottawa, ON K1A 0R6, Canada (e-mail: wanglin_2k@sina.com).

H. Ding is with the Acoustics and Signal Processing Group, Institute for Microstructural Sciences, National Research Council Canada, Ottawa, ON K1A 0R6, Canada (e-mail: heping.ding@nrc-cnrc.gc.ca).

F. Yin is with the School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: flyin@dlut.edu.cn).

Digital Object Identifier 10.1109/TASL.2010.2052244

One well recognized BSS application is the separation of audio sources that have been mixed and recorded by multiple microphones in a real room, known as a cocktail party environment. The challenge of the problem is that the mixing process is convolutive, where the observations are combinations of the unknown filtered versions of the sources. Approaches to solve the convolutive blind source separation problem [2] can be classified into two categories: time-domain and frequency-domain. In time-domain BSS, the separation network is derived by optimizing a time-domain cost function [3]–[5]. These approaches may not be effective due to slow convergence and large computation load. In frequency-domain BSS, the observed time-domain signals are converted into the time–frequency domain by short-time Fourier transform (STFT), then an instantaneous BSS algorithm is applied to each frequency bin, after which the separated signals of all frequency bins are combined and inverse-transformed to the time domain [6]–[8]. Although satisfactory instantaneous separation may be achieved in all frequency bins, combining them to recover the original sources is a challenge because there are unknown permutations associated with individual frequency bins. This permutation ambiguity should be looked after properly so that the separated frequency components from the same source are grouped together. Besides the conventional frequency-domain BSS, methods based on sparsity representation have emerged in recent years and are also promising for convolutive blind source separation [9].

There are three common strategies used to solve the permutation problem. The first is to make the separation filters smooth in the frequency domain [4], [10], [11]. This may be achieved by limiting their lengths. It has been proved in [11] that the permutation ambiguity can be avoided if the separation filters are short enough relative to the FFT block size. Besides, the spectral continuity of separation filters can also be exploited to align the permutation [12]. The second strategy is to exploit the dependence of separated signals across frequencies. The inter-frequency dependence of signal envelope was exploited in [13] to align the permutation. The inter-frequency dependence of signal power was exploited in [14], which shows a clearer dependence than the envelope measure does. The permutation alignment using inter-frequency dependence may be precise, but a misalignment at a frequency bin may lead to a big misalignment beyond that frequency. This is referred to as misalignment spread. References [15] and [16] incorporated the inter-frequency dependences into instantaneous BSS, so that bin-wise separation and permutation alignment can be obtained simultaneously. The third strategy is to exploit the position information of sources

such as direction-of-arrival (DOA) [17], [18]. It is believed that contributions from the same source are likely to come from the same direction. By estimating the arriving delay of sources or analyzing the directivity pattern formed by a separation matrix, source direction can be estimated and permutations aligned. Compared to the second strategy, the third one is more robust because the source direction estimation at one frequency bin does not depend on other frequencies. However, the source direction estimation is often not precise, especially in reverberant environments. Moreover, the geometry of the microphone array must be known in order for the source direction to be estimated. To distinguish it from a fully blind method which only needs observed signals, we regard the third strategy as semi-blind.

Generally, source direction information and signal inter-frequency dependence can be combined to get a better permutation alignment [17]. Since the direction-of-arrival information is not available in many cases, we attempt to solve the problem using only one piece of inter-frequency dependence information: the powers of separated signals. This measure exhibits a clear dependence across frequencies for separated signals, and the permutation alignment based on it is correct in most frequency bins. However, errors still occur in some frequency bins, resulting in permutation misalignment spread which may destroy signal reconstruction. To minimize the spread, this paper proposes a novel scheme called "region-growing permutation alignment approach." Bin-wise permutation alignment is applied first across all frequency bins, using the correlation of separated signal powers; then the full frequency band is partitioned into small regions based on the bin-wise permutation alignment result; finally, region-wise permutation alignment is performed in a region-growing manner. The region-wise permutation correction scheme minimizes the spreading of the misalignment at isolated frequency bins to others, hence to improve permutation alignment. Experimental results have verified the effectiveness of the proposed method.

The rest of the paper is organized as follows. The principle of frequency-domain blind source separation is discussed in Section II. The proposed permutation alignment scheme based on power measure is described in detail in Section III. Experiment results are presented in Section IV. Computation cost analysis of the proposed algorithm is given in Section V. Finally, Section VI concludes the paper.

## II. Frequency-Domain Blind Source Separation

Supposing $N$ sources and $M(M \geq N)$ sensors in a real-world acoustic scenario, the source vector $s(n) = [s_1(n), \cdots, s_N(n)]^T$, and the observed vector $x(n) = [x_1(n), \cdots, x_M(n)]^T$, the mixing channels can be modeled by finite impulse response (FIR) filters of length $P$, the convolutive mixing process is formulated as

$$x(n) = H(n) * s(n) = \sum_{p=0}^{P-1} H(p)s(n-p) \tag{1}$$

where $H(n)$ is a sequence of $M \times N$ matrices containing the impulse responses of mixing channels. For separation, we use FIR unmixing filters of length $L$ and obtain estimated source signal vector $y(n) = [y_1(n), \cdots, y_N(n)]^T$ by

$$y(n) = W(n) * x(n) = \sum_{l=0}^{L-1} W(l)x(n-l) \tag{2}$$

where $W(n)$ is a sequence of $N \times M$ matrices containing the unmixing filters.

The unmixing network $W(n)$ can be obtained by a frequency-domain BSS approach. After transforming the signals to the time–frequency domain using blockwise $L$-point short-time Fourier transform (STFT), the convolution becomes a multiplication

$$X(m,f) = H(f)S(m,f) \tag{3}$$
$$Y(m,f) = W(f)X(m,f) \tag{4}$$

where $m$ is a decimated version of the time index $n$, $X(m,f)$ and $Y(m,f)$ are the STFTs of $x(n)$ and $y(n)$, respectively, $H(f)$ and $W(f)$ are Fourier transforms of $H(n)$ and $W(n)$, respectively, and $f \in [f_0, \cdots, f_{L/2}]$ is the frequency.[1]

In the frequency domain, it is possible to separate each frequency bin independently using complex-valued instantaneous BSS algorithms such as FastICA [1], [19] and Infomax [20], [21], which are considered quite mature. However, there are scaling and permutation ambiguities in each bin. This is expressed as

$$Y(m,f) = W(f)X(m,f) = \Lambda(f)D(f)S(m,f) \tag{5}$$

where $D(f)$ is a permutation matrix and $\Lambda(f)$ a scaling matrix, all at frequency $f$. It is necessary to correct the scaling and permutation ambiguities before transforming the signals back to the time domain.

The permutation at each bin should be aligned so that the separated components originating from the same source are grouped together. The permutation correction is a challenging problem, and will be addressed in Section III.

The scaling ambiguity can be resolved relatively easily, by using the Minimal Distortion Principle [22]

$$W_s(f) = diag\left(W_p^{-1}(f)\right) \cdot W_p(f) \tag{6}$$

where $W_p(f)$ is $W(f)$ after permutation correction and $W_s(f)$ is the one after scaling correction, $(\cdot)^{-1}$ denotes inversion of a square matrix or pseudo inversion of a rectangular matrix; $\mathrm{diag}(\cdot)$ retains only the main diagonal components of the matrix.

Finally, the unmixing network $W(n)$ is obtained by inverse Fourier transforming $W_s(f)$, and the estimated source $y(n)$ is obtained by filtering $x(n)$ through $W(n)$. The workflow of the frequency-domain BSS is shown in Fig. 1.

[1]Only positive frequencies are addressed; the negative frequency range is looked after by exploiting the conjugate symmetry property of FFT on real signals.

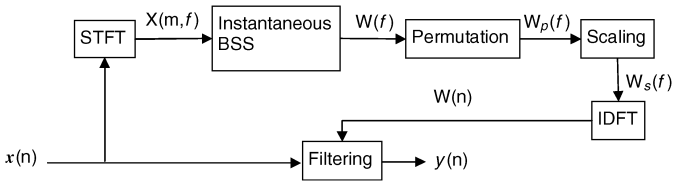



Fig. 1. Workflow of frequency-domain blind source separation.

## III. Permutation Alignment

### A. Inter-Frequency Dependence Measure

The inter-frequency dependence of speech sources can be exploited to align the permutations across all frequency bins. The correlation between separated signal envelopes is commonly used as a measure of inter-frequency dependence. However, this dependence is only clearly exhibited among a small set of frequencies. Another inter-frequency dependence measure proposed in [14] is the correlation between power ratios of separated signals. It exhibits a clearer inter-frequency dependence among all frequencies.

The $M \times N$ mixing network at frequency $f$ can be estimated from the separation network by

$$A(f) = W^{-1}(f) = [a_1(f), \cdots, a_N(f)] \tag{7}$$

where $a_i(f)$ is the $i$th column vector of $A(f)$. The observed signal can be decomposed by

$$X(m,f) = \sum_{i=1}^{N} a_i(f) Y_i(m,f) \tag{8}$$

where $Y_i(m,f)$ is the $i$th component of $Y(m,f)$, i.e., $Y(m,f) = [Y_1(m,f), \cdots, Y_N(m,f)]^T$.

A power ratio measure is calculated to represent the dominance of the $i$th separated signal in the observations at frequency $f$. It is defined as

$$v_i^f(m) = \frac{\|a_i(f)Y_i(m,f)\|^2}{\sum_{k=1}^{N} \|a_k(f)Y_k(m,f)\|^2} \tag{9}$$

where the denominator is the total power of the observed signals $X(m,f)$, and the numerator is the power of the $i$th separated signal. Being in the range [0, 1], (9) is close to 1 when the $i$th separated signal is dominant, and close to 0 when others are dominant. The power ratio measure can clearly exhibit the signal activity due to the sparseness of speech signals.

The correlation coefficient of signal power ratios can be used for measuring inter-frequency dependence and solving the permutation problem. The normalized bin-wise correlation coefficient between two power ratio sequences $v_i^{f_1}(m)$ and $v_j^{f_2}(m)$ is defined as

$$\rho\left(v_i^{f_1}, v_j^{f_2}\right) = \frac{r_{ij}(f_1,f_2) - \mu_i(f_1)\mu_j(f_2)}{\sigma_i(f_1)\sigma_j(f_2)} \tag{10}$$

where $i$ and $j$ are indices of two separated channels, $f_1$ and $f_2$ are two frequencies, $r_{ij}(f_1,f_2) = E\{v_i^{f_1} v_j^{f_2}\}$, $\mu_i(f) = E\{v_i^f\}$, $\sigma_i(f) = \sqrt{E\{(v_i^f)^2\} - \mu_i^2(f)}$ are, respectively, the correlation, mean, and standard deviation at time $m$ (The time index $m$ is omitted for clarity). Note that $E\{\cdot\}$ denotes expectation. Being in the range of $[-1,1]$, (10) equals 1 when the two sequences are identical. In general, (10) tends to be high if output channels $i$ and $j$ originate from the same source and low if they represent different sources. This property will be used for aligning the permutation.

Reference [14] provided a two-step optimization scheme for permutation alignment with the power ratio measure. First, the power ratios are calculated for all frequency bins and all $N$ separated signals, and then clustered into $N$ subgroups—corresponding to $N$ separated signals, respectively. This is a global permutation alignment process. Second, the permutation is aligned bin by bin. This is a local permutation alignment process. The overall alignment performance depends crucially on the accuracy of the clustering result of the first step. However, the clustering process is unsupervised and it is difficult to control it. In some cases, large misalignment occurs during the process, resulting in a failure in the overall permutation alignment. Next, we will propose a new permutation alignment scheme without a clustering process.

### B. Proposed Permutation Alignment Scheme

As discussed above, (10) tends to be high when $i$ and $j$ belong to the same source, but such a dependence is not always evident. Generally, the misalignment at an isolated frequency bin may spread to other frequencies easily, causing a big misalignment beyond it. A region-growing permutation scheme is proposed which can avoid this problem. The scheme is described in 6 steps as follows.

Step 1) Calculate the power ratio $v_i^f(m)$ for all frequency bins and all separated signals by (9), where $m$ is time index, $f \in [f_0, \cdots, f_{L/2}]$ is frequency bin index, $L$ is FFT size, and $i \in [1, \cdots, N]$ is signal channel index.

Step 2) Correct permutation bin by bin from $f_2$ to $f_{L/2}$, so that the power ratio time sequences at each frequency bin has the highest correlation coefficient with the previous bin. For current bin $f$ and previous bin $f-1$ (supposing the correct permutation $\Pi_{f-1}$ is known for bin $f-1$), select a permutation $\Pi_f : \{1, \cdots, N\} \to \{1, \cdots, N\}$ which maximizes the average value of correlation coefficients. This is expressed as

$$\Pi_f \leftarrow \arg\max_{\Pi} \sum_{k=1}^{N} \rho\left(v_i^f, v_{i'}^{f-1}\right)\Big|_{i=\Pi_f(k), i'=\Pi_{f-1}(k)}. \tag{11}$$

Now the average correlation coefficient $\rho_f$ for permutation $\Pi_f$ is

$$\rho_f = \frac{1}{N} \sum_{k=1}^{N} \rho\left(v_i^f, v_{i'}^{f-1}\right)\Big|_{i=\Pi_f(k), i'=\Pi_{f-1}(k)}. \tag{12}$$

Step 3) Divide the full frequency band into low and high bands, and process them separately in step 4–5.

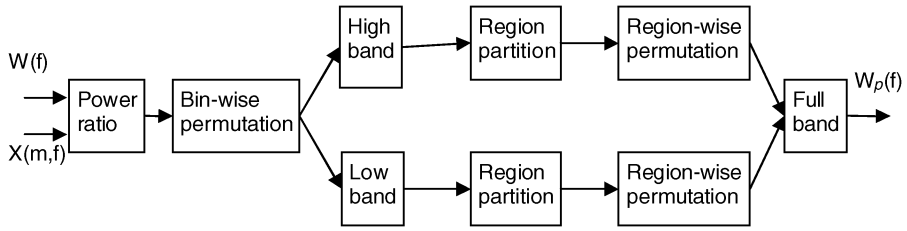



Fig. 2. Workflow of the proposed permutation alignment method.

Step 4) Partition the high (low) frequency band into $K$ regions, with the highly related frequency bins assigned to the same region. The criteria for obtaining region $R_l$ are

$$\begin{cases} 1) \text{ The index } f \text{ should be consecutive in region } R_l \\ 2) \ \forall f \in R_l, \rho_f > U_{th} \end{cases}$$

where $U_{th}$ is a threshold. When the bin-wise correlation coefficient is higher than the threshold $U_{th}$, the permutation alignment is regarded as correct. The selection of $U_{th}$ will be discussed later.

Step 5) Select a region with the largest number of elements as a seed; merge with its neighboring regions on both sides in a region-growing style until the new region covers the full high (low) frequency band. The merging procedure for the seed region $R_l$ and a neighboring region $R_{l+1}$ is as follows.

5.1) Calculate the $N$ centroids $c_k^l$ and $c_k^{l+1}$ for regions $R_l$ and $R_{l+1}$, respectively, by

$$c_k^l(m) = \frac{1}{N_l} \sum_{f \in R_l} v_i^f(m)|_{i=\Pi_f(k)}, \quad k = 1, \cdots, N \tag{13}$$

where $N_l$ is the number of elements in region $R_l$.

5.2) Find a permutation $\Pi_{l+1}$ (for region $R_{l+1}$) that maximizes the average value of correlation coefficient between $c^l$ and $c^{l+1}$, which is expressed by

$$\Pi_{l+1} \leftarrow \arg\max_{\Pi} \sum_{k=1}^{N} \rho\left(c_i^l, c_{i'}^{l+1}\right)\Big|_{i=\Pi_l(k), i'=\Pi_{l+1}(k)} \tag{14}$$

5.3) After permutation correction for frequency bins in region $R_{l+1}$ with $\Pi_{l+1}$, merge $R_l$ and $R_{l+1}$ into a new seed region.

Step 6) After permutation corrections in high and low frequency bands, calculate the centroids for both bands and merge them together with a manner similar to step 5.

Based on the description above, the workflow of the proposed scheme is shown in Fig. 2. It aligns permutations based on the power ratio measure, which represents the dominance of one source signal in the total observed signals at every frequency bin. The region-growing idea behind the proposed method is somewhat similar to the one proposed by Murata [13]. The Murata method sorts frequency bins in increasing order of similarity among independent components, and then aligns the permutation of the sorted frequency bins one by one by maximizing the correlation between the current bin and a set of bins in which the permutation has already been decided. This method does not limit the frequency range in which correlations are calculated, because it assumes high correlations of envelopes even between frequencies that are not close neighbors. This assumption is not satisfied for all pairs of frequencies, especially when many sources are mixed. As a result, permutations may be misaligned at many frequencies. Alternatively, the proposed method in this paper aligns permutation in neighboring regions before growing the permutation aligned regions to the whole frequency band. It exploits the dependence between neighboring frequency bins more efficiently.

It is observed that speech signals generally have nonuniform energy distribution across the frequency, with more energy in low frequency regions. Consequently, inter-frequency dependence in the low frequency band may not be as evident as in the high band. If we process signal in one full band, misalignment spread may still be observed at low frequencies in some cases. On the other hand, the fact that we process in low and high frequency bands separately before merging them helps reduce the misalignment spread and leads to better permutation alignment. For speech signals, most energy is usually in frequencies below 500–700 Hz; thus, we choose the split frequency at 600 Hz. This value works well in most speech cases. In a similar manner, the full frequency band could also be divided into three or more subbands before processing, but this multiple-band scheme is not expected to improve the separation result further for speech signals, which generally have energy concentrated in low frequencies. For signals other than speech, a multiple-band scheme may be considered. This can be a topic for further research. In Step 4, the threshold $U_{th}$ is an important parameter since it directly affects the partitioning of the regions. $U_{th}$ is determined empirically by

$$U_{th} = \min(U_{th1}, U_{th2}) \tag{15}$$

where function $\min(\cdot)$ returns the minimal argument, $U_{th1} = 0.7$, and $U_{th2}$ is the 80%th smallest value[2] in the correlation coefficients set calculated by (12) after bin-wise permutation alignment in the high (low) band.[3] In Step 5, the proposed method attempts to start from a region with strong inter-frequency dependence. Generally, a region with largest number of seeds may have a high average correlation value; thus, we start growing from it. In summary, since the region-wise permutation alignment employs a region-growing style, the misalignment at some frequency does not affect the

[2] $p\%$th smallest value is the $p'$th ranked value from smallest to highest, where $p'$ equals $p\%$ of the size of the data set.

[3] Some experiments will be given in Section IV-D on how to choose $U_{th}$.

permutation result and therefore will not spread to other frequencies. The validity of the proposed method will be verified in the experiments.

## IV. Experiment Results and Analysis

We evaluate the performance of the proposed method using both simulated and real data. The experiments are composed of four parts. The first part exhibits the permutation alignment result of the proposed method with simulated data. The second verifies the separation performance in different conditions with simulated data and compares it with other well-known frequency-domain BSS algorithms. The third part investigates the performance of the proposed algorithm in real environments. The last one examines how threshold (15) affects the performance.

The separation performance is measured by signal-to-interference ratio (SIR), in dB. The input and output SIRs for the $J$th channel are defined respectively as

$$\mathrm{SIRIN}_J = 10\log_{10}\frac{\sum_n \left|\sum_l h_{JJ}(l)s_J(n-l)\right|^2}{\sum_{k\neq J}\sum_n \left|\sum_l h_{Jk}(l)s_k(n-l)\right|^2} \quad (16)$$

$$\mathrm{SIROUT}_J = 10\log_{10}\frac{\sum_n \left|\sum_l g_{Jp(J)}(l)s_{p(J)}(n-l)\right|^2}{\sum_{k\neq p(J)}\sum_n \left|\sum_l g_{Jk}(l)s_k(n-l)\right|^2} \quad (17)$$

where $n$ is time index, $J = 1, \cdots, N$, and $p(J)$ is the index of the output channel where the $J$th source appears, $h_{Jk}(n)$ is an element of $H(n)$ [see (1)], and $g_{Jk}(n)$ is an element of the overall impulse response matrix $G(n) = W(n) * H(n)$.

The implementation detail of the algorithm is as follows. The Tukey window is used in short-time Fourier transform, with a shift size of 1/4 window length. The instantaneous BSS is implemented by means of the Scaled Infomax [23], which can converge to the optimal solution within 100 iterations. In this paper, we set the iteration number as 100. The scaling ambiguity is solved by using the Minimum Distortion Principle (6). The smoothing method proposed in [24] is applied in order to reduce spikes due to the circularity effect of the FFT. The processing bandwidth is between 100 and 3750 Hz (sampling rate being 8 kHz).

### A. Permutation Alignment Experiment

In this experiment the proposed algorithm is applied to the problem with three microphones and three sources in a simulated environment. The environment is shown in Fig. 3, where only microphones A, B, C and sources 1, 2, 5 are used. The source signals are two male speeches and one female speech of 8 seconds long each. All sources and microphones are 1.5 m high. The simulated room reverberation time is $RT_{60} = 130$ ms, time required for the sound level to decrease by 60 dB. One typical room impulse response is shown in Fig. 4. It has been obtained using the image source method [26], and the reverberation time was controlled by varying the absorption coefficient of the wall. The STFT frame size is 1024 samples with a shift size of 256.

To demonstrate the permutation alignment performance, we show in Fig. 5 the results at three stages: (a) before permutation alignment, (c) after bin-wise permutation alignment, and (d) after region-wise permutation alignment. Fig. 5(b) gives the correlation coefficients (12) across frequencies after bin-wise permutation alignment. The permutation result is evaluated using the method proposed in [25], supposing the mixing filters are known. The detailed process is as follows.

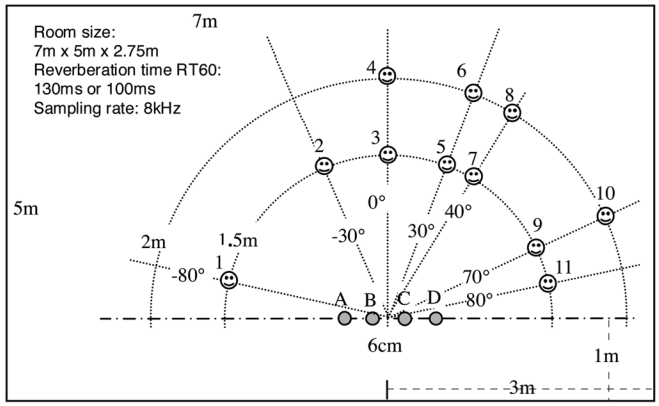



Fig. 3. Simulated room environment.

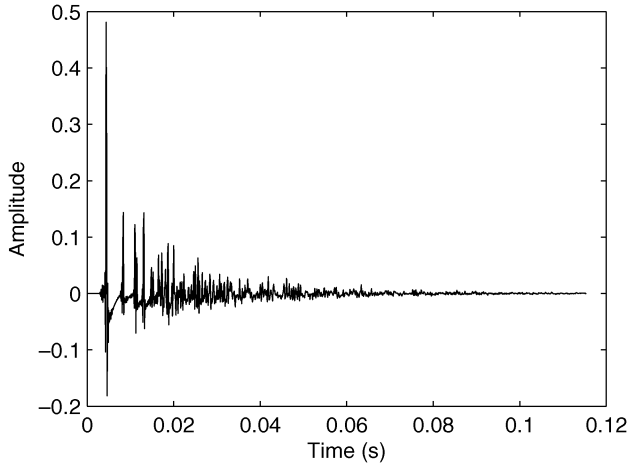



Fig. 4. Simulated room impulse response, $RT_{60} = 130$ ms.

Given the mixing matrix $H(f)$ and the unmixing matrix $W(f)$ at each frequency bin, we consider $G(f) = W(f)H(f)$. The correct permutation at channel $i$ corresponds to the maximal value in the $i$th row, which is expressed as

$$perm_i = \arg\max_j |G_{ij}(f)|. \quad (18)$$

It can be seen from Fig. 5 that the permutation ambiguity is very severe before permutation alignment; the ambiguity is mitigated after bin-wise permutation alignment but large misalignments still occur. Comparing Fig. 5(b) to (c), it can be observed that the correlation coefficient is high in most frequency bins. At bins with a high correlation coefficient, the permutation tends to be correct, whereas the permutation may or may not be correct at bins with a low correlation. At bins with incorrect permutation, the errors may leak to nearby regions, causing misalignment spread. The proposed method can solve the problem well. As seen in Fig. 5(d), the ambiguity is almost eliminated after region-wise permutation alignment except some isolated

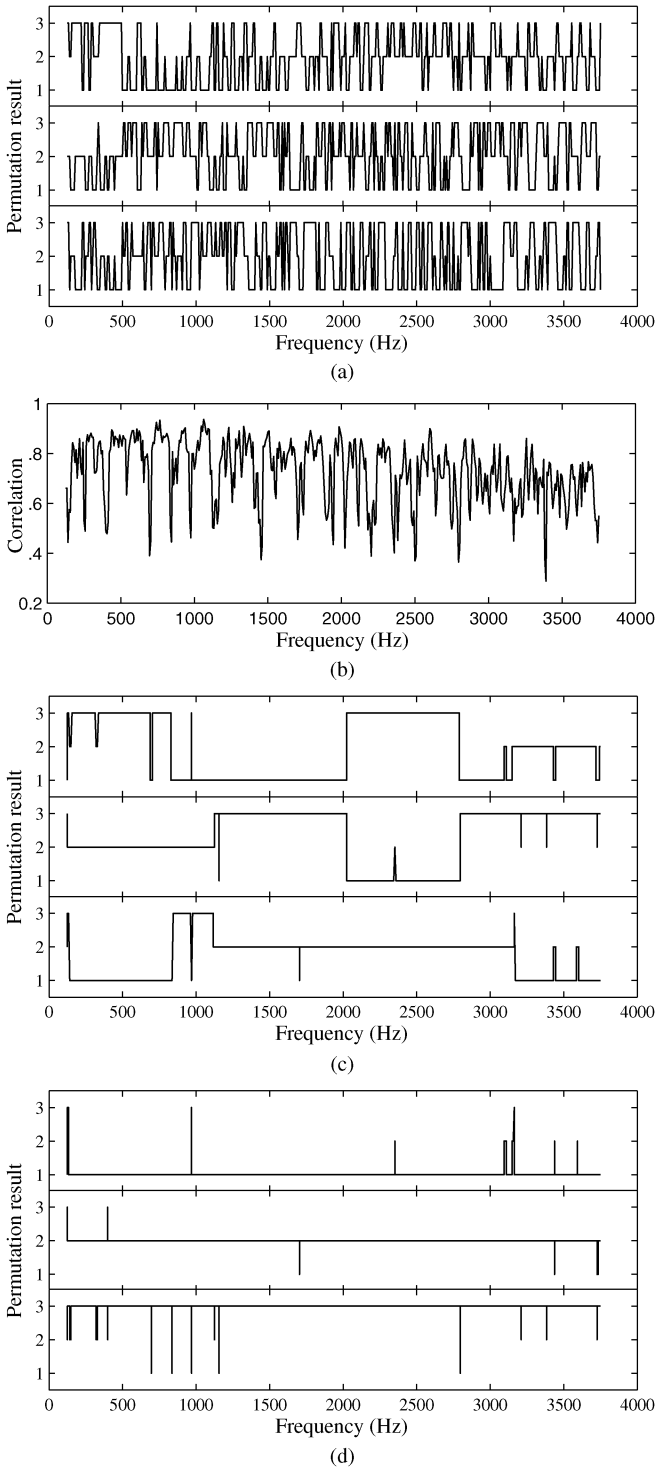



Fig. 5. Permutation result of the proposed method. (a) Before permutation alignment. (b) Average correlation coefficient. (c) After bin-wise permutation alignment. (d) After region-wise permutation alignment.

frequency bins. These misalignments do not spread to nearby frequencies. Another phenomenon observed in Fig. 5(d) is that a misalignment can occur alone—without being paired with one in another channel. For example, the one at 1703 Hz occurs only in channel 2. That is believed to be due to the bad instantaneous separation at this bin, not the permutation alignment error. Finally, the average output SIR is 13.4 dB, a 16.2-dB improvement over the average input SIR of $-2.8$ dB.

### *B. BSS Experiments in Different Simulation Conditions*

In this experiment, the proposed method is applied to separation scenarios in different simulated conditions. The environment is still shown in Fig. 3. The room reverberation time is $RT_{60} = 100$ ms. The STFT frame size is 2048 with a shift size of 512. Various $2 \times 2$ (2 sources and 2 microphones) and $4 \times 4$ (4 sources and 4 microphones) simulation cases are carried out. For the $2 \times 2$ case, microphones B, C, and seven different combinations of source locations are used. The sources are one male speech and one female speech of 8 seconds each. For the $4 \times 4$ case, all microphones, and five different combinations of source locations are used. The sources are two male speeches and two female speeches of 8 seconds each.

The proposed algorithm is compared with six frequency-domain BSS algorithms. The first one was proposed by Sawada which corrects the permutation ambiguity using both signal inter-frequency correlation and source direction information [17]. The second is the independent vector analysis (IVA) method which incorporates inter-frequency dependence evaluation into the instantaneous BSS criterion to obtain bin-wise separation and permutation alignment simultaneously [15]. The third is the Murata method which aligns permutation based on the correlation of signal envelopes [13]. We call it "Murata(Env) method." The fourth is the same as the third, except that the inter-frequency dependence measure of signal envelope is replaced by the power ratio measure—we call it "Murata(Pow) method." The fifth is the method proposed in [14] which originally resolves permutation problem based on power ratio measure. We call it "PowerRatio method." The last one is blind separation with permutation corrected by supposing the mixing filters are known, which we call "Benchmark" [25]. To evaluate the performance of the proposed permutation alignment scheme efficiently, we use the same frequency-domain BSS processing for the proposed method and methods 3–6 (Murata(Env), Murata(Pow), PowerRatio, and Benchmark): with all processing stages and parameters fixed but different permutation correction schemes.

Table I shows the average output SIRs for all cases,[4] where "n/a" means the algorithm fails in the experiment condition. Note that the average input SIR is about 0 dB for the $2 \times 2$ case and $-5$ dB for the $4 \times 4$ case. As shown in Table I, the proposed algorithm outperforms the Sawada method and the IVA method in most cases. The Sawada method does not show consistent separation performance for all cases, as do the proposed and IVA methods. The Murata(Env) method performs as well as the Murata(Pow) method for $2 \times 2$ cases, while the latter one performs better for $4 \times 4$ cases by using the power ratio measure instead of the signal envelope measure. The performance for $4\times 4$ cases is further improved by the proposed method, which employs a region-growing permutation alignment scheme. The PowerRatio method shows similar performance as the proposed method does for $2 \times 2$ cases, while performs much worse for $4 \times 4$ cases. As discussed earlier, the performance of the PowerRatio method depends greatly on the clustering result of its global permutation step. For 4-source cases, large misalignment

[4]Parts of the experiment results were obtained from [16], and used directly for comparison.

TABLE I
BSS SEPARATION RESULTS (OUTPUT SIR IN dB) FOR DIFFERENT CONDITIONS

| Source location | Sawada | IVA | Murata (Env) | Murata (Pow) | PowerRatio | Proposed | Benchmark |
|---|---|---|---|---|---|---|---|
| 1, 11 | n/a | 16.4 | 24.9 | 24.9 | 24.9 | 24.8 | 24.9 |
| 2, 7 | 16.9 | 15.8 | 24.2 | 24.9 | 24.7 | 24.9 | 24.9 |
| 5, 7 | 8.1 | 16.5 | 12.7 | 14.2 | 13.4 | 13.8 | 14.5 |
| 9, 11 | n/a | 11.7 | 10.5 | 10.1 | 9.5 | 10.9 | 11.8 |
| 3, 4 | 14.0 | 15.3 | 17.2 | 17.5 | 17.5 | 17.5 | 17.6 |
| 5, 6 | 11.5 | 14.9 | 13.3 | 15.2 | 14.8 | 15.1 | 15.2 |
| 9, 10 | 15.7 | 14.9 | 13.4 | 12.9 | 12.7 | 13.5 | 15.1 |
| 1, 2, 7, 9 | 16.5 | 11.5 | 6.7 | 8.9 | 7.9 | 17.2 | 18.3 |
| 2, 3, 7, 9 | 9.9 | 10.4 | 9.5 | 12.7 | 6.1 | 13.3 | 15.1 |
| 3, 5, 7, 9 | 10.1 | 11.2 | 5.7 | 8.2 | 7.8 | 9.1 | 12.5 |
| 5, 7, 8, 9 | 8.5 | 8.2 | 3.0 | 7.1 | 4.2 | 10.0 | 12.0 |
| 2, 3, 4, 5 | 6.6 | 7.7 | 5.2 | 8.3 | 7.7 | 9.7 | 11.2 |

occurs during the clustering process, resulting in the poor separation performance. Comparing to the benchmark, all methods show excellent performance for $2 \times 2$ cases, and degraded performance for $4 \times 4$ cases. However, the proposed method is closer to the ideal separation results than other methods especially for $4 \times 4$ cases. In a word, with both power ratio measure and region-growing permutation scheme, the proposed method shows superiority to other methods.

The separation performance depends mainly on two factors: instantaneous BSS and permutation alignment. It is easier to separate two sources at each frequency bin when they are far apart, because the transfer functions to sensors are different. On the other hand, it becomes difficult for instantaneous BSS to do the job when the two sources are close and have similar transfer functions to sensors. The degraded separation further affects the permutation alignment. Thus, as shown in Table I, huge improvement is obtained by the proposed method when the sources are spaced far apart.

### C. BSS Experiment in Real Room Conditions

To evaluate the performance of the proposed method in real environments, we use real recorded data downloaded from the internet.[5] The recording was made in a room of size $3.55$ m $\times$ $4.45$ m $\times$ $2.5$ m, the room reverberation time was $RT_{60} = 130$ ms, the source signals were two male speeches and two female speeches of 7 seconds long each, the sampling rate was 8 kHz. Various $n \times n$ experiment conditions were tested. In the dataset, individual contributions from the sources to the microphones are also available. These can be used for calculating the SIR. The data was provided by Sawada, who also gave the separation results with his own algorithm [17]. In the experiment, the STFT frame length is 2048 with a shift size of 512. As shown in Table II, the separation performance of the two methods is comparable. However, the Sawada method needs the geometrical information of microphone array while the proposed method does not rely on that.

### D. BSS Experiments With Different Threshold Parameters

The threshold $U_{th}$ in (15) is critical for region partition: the bin-wise permutation is deemed correct when the corresponding correlation coefficient exceeds it. $U_{th}$ can be either constant or adaptive. When adaptive, the threshold is determined according to the bin-wise permutation result, for example, the 80%th smallest value in the correlation coefficient set calculated from (12) after bin-wise permutation alignment. To make a more reliable decision, we combine the constant and adaptive schemes to come up with (15).

[5]See <http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/bss2to4/index.html>.

TABLE II
BSS SEPARATION RESULTS (OUTPUT SIR) FOR REAL-RECORDED DATA

| Condition | Input SIR (dB) | Sawada BSS (dB) | Proposed BSS (dB) |
|---|---|---|---|
| 2 x 2 | 0.0 | 17.3 | 17.5 |
| 3 x 3 | -2.9 | 12.5 | 11.9 |
| 4 x 4 | -4.7 | 9.3 | 8.2 |

Experiments were conducted to determine $U_{th}$. The environment is as shown in Fig. 3 with a reverberation time of 100 ms. Various $2 \times 2$ and $4 \times 4$ simulation cases were carried out. The combinations of source locations are given in Table I. Two STFT frame sizes, 1024 and 2048, were tested, respectively. The experiments were repeated three times—with different speakers each time. Finally, the average performance was calculated from the results.

We first determine the constant threshold $U_{th1}$. $U_{th} = U_{th1}$ is set at 0.1 to 0.9 with increments of 0.1. The separation results in different conditions are shown in Fig. 6. It can be seen from Fig. 6 that, in $2 \times 2$ cases, which are easier, the performance does not vary much when $U_{th} \geq 0.2$. In the more challenging cases, the $4 \times 4$ mixture, the performance varies significantly with the threshold value and peaks at $U_{th} = 0.7$. Thus, it is determined as the optimal value for $U_{th1}$.

Next, the adaptive parameter $U_{th2}$ is chosen as the $per$th smallest value in the correlation coefficient set after bin-wise permutation alignment, where $per$ is selected from 10% to 90% with 10% increments. The final threshold is calculated as $U_{th} = \min(0.7, U_{th2})$. The separation results in different conditions are depicted in Fig. 7. It can be seen that, in both $2 \times 2$ and $4 \times 4$ cases, the performance improves with $per$, and flattens when $per \geq 40\%$. In other words, $per$ can be any value between 40% to 90%. We set $per$ at 80%.

In summary, the parameters in (15) are set experimentally, where $U_{th1} = 0.7$, and $U_{th2}$ equals the 80%th smallest value in the correlation coefficient set. Although these values are determined heuristically and experimentally, they work well in most cases. It may be a future research topic to find a smarter scheme for determining the threshold.

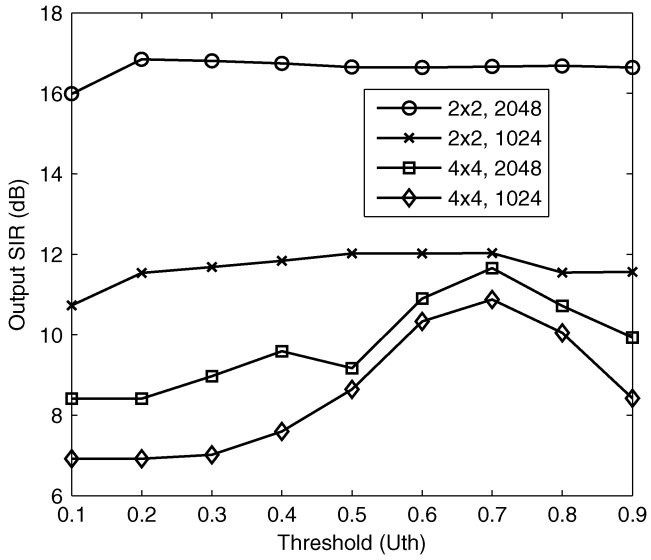



Fig. 6. BSS separation results versus constant threshold $U_{th}$.

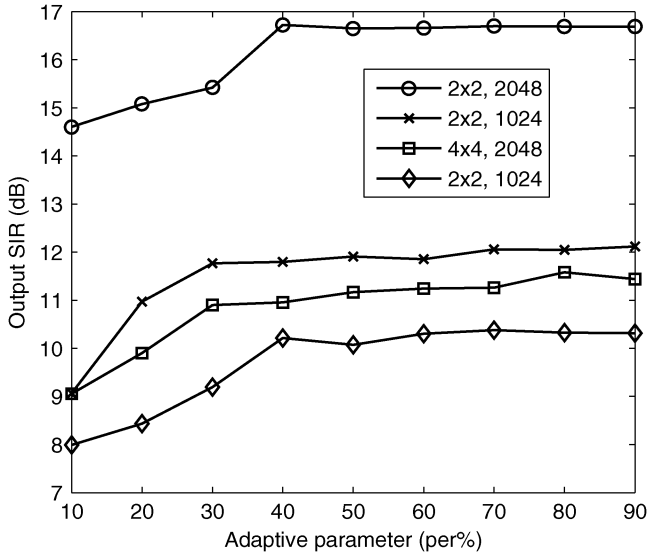



Fig. 7. BSS separation results versus adaptive threshold parameter $per$.

## V. Computation Cost Analysis

In this section, we analyze the computation cost of the proposed BSS algorithm. Suppose there are $N$ sources and $N$ microphones, the length of input signals is $T$, and the STFT frame length is $L$ with a window shift $= L/4$. After STFT, the number of data points available for each frequency bin is approximately $B = T/\text{shift}$. From Figs. 1 and 2, the computation of the frequency-domain BSS is mainly composed of three algorithm blocks: STFT, instantaneous BSS and permutation, while the computation of the proposed permutation algorithm is mainly composed of three algorithm blocks: power ratio calculation, bin-wise permutation, and region-wise permutation. Given the above, the computation cost of the proposed BSS algorithm is summarized in Table III. For convenience, only complex-valued multiplication operations are considered; besides, the computation for (2), the unmixing filtering, is not taken into account.

TABLE III
Computation Cost of the Proposed BSS Algorithm

| Algorithm block | Computations |
|---|---|
| STFT | $N \cdot B \cdot (L/2) \cdot \log L$ |
| Instantaneous BSS | $(L/2) \cdot iter \cdot cpi$ |
| Power Ratio | $(L/2) \cdot (N^2B + NB)$ |
| Bin-wise permutation | $(L/2) \cdot N^2B$ |
| Region-wise permutation | $R \cdot N^2B$ |

In Table III, the term $iter$ is the number of iterations for Scaled Infomax algorithm, and $cpi$ stands for computations per iteration, which is approximately

$$cpi = 2N^2B. \tag{19}$$

The term $R$ is the number of regions involved in region-wise permutation, and $R < L/2$.

To summarize, the total computation cost of the proposed BSS algorithm for a speech mixture of length $T$ is

$$\begin{aligned} c_{\text{total}} &= LN^2B \cdot \left( iter + \frac{1}{2} + \frac{1}{2N} + \frac{1}{2} + \frac{R}{L} + \frac{\log L}{2N} \right) \\ &< LN^2B \cdot \left( iter + 2 + \frac{\log L}{2N} \right) \\ &= 4TN^2 \cdot \left( iter + 2 + \frac{\log L}{2N} \right). \end{aligned} \tag{20}$$

Generally, $N \geq 2$ and $L \leq 65536$; thus, $c_{\text{total}} < 4TN^2 \cdot (iter + 6)$. For the total $TN$ input data points, the average computation for each data point is

$$c_{avg} < 4N \cdot (iter + 6). \tag{21}$$

Equation (21) shows that the computation cost for each input data sample is $\mathrm{O}(N)$. We think the result is quite acceptable. For example, the computation cost for the $4 \times 4$ case in the experiment with $iter = 100$ involves less than 1700 complex-valued multiplications.

It can be seen from (20) that instantaneous BSS spends most computation effort in frequency-domain BSS. To be precise, we calculate the ratio between the computation required by the permutation algorithm and by instantaneous BSS as

$$\frac{\frac{1}{2} + \frac{1}{2N} + \frac{1}{2} + \frac{R}{L}}{iter} \approx \frac{2}{iter} = 2\%. \tag{22}$$

That is, the computation for the permutation algorithm is quite negligible compared with that for instantaneous BSS. In other words, the permutation correction algorithm only increases the number of calculations slightly.

The execution time of an algorithm depends on a lot of factors such as computational complexity, program structure, hardware pipeline, etc. In terms of computational complexity, the proposed frequency-domain BSS is promising for real-time application. After code optimization, the execution of it should be fast.

## VI. Conclusion

Studying frequency-domain convolutive blind source separation, this paper proposes a new permutation alignment method

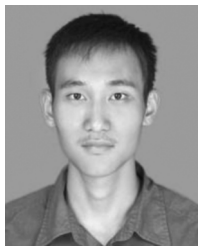
based on an inter-frequency dependence of separated signal powers. With a region-growing permutation alignment style, the proposed method minimizes the spreading of misalignment at isolated frequency bins to others. Good separation results are observed in simulations with synthetic and real data. Last but not least, the proposed method is computationally efficient so it is promising for real-time applications.

## ACKNOWLEDGMENT

The authors would like to thank Dr. T. Mei for his invaluable discussions. The authors would also like to thank Dr. H. Sawada for fruitful discussions and providing information on the original work.

## REFERENCES

[1] A. Hyvarien, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.
[2] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook on Speech Processing and Speech Communication*. New York: Springer, 2007, pp. 1–34.
[3] S. C. Douglas and X. Sun, "Convolutive blind separation of speech mixtures using the natural gradient," *Speech Commun.*, vol. 39, pp. 65–78, 2003.
[4] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," *Signal Process.*, vol. 86, no. 6, pp. 1260–1277, 2006.
[5] S. C. Douglas, M. Gupta, H. Sawada, and S. Makino, "Spatio-temporal FastICA algorithms for the blind separation of convolutive mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1511–1520, Jul. 2007.
[6] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.
[7] H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Blind Speech Separation*. New York: Springer, 2007, pp. 47–78.
[8] 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 Process.*, vol. 22, no. 2, pp. 109–116, Feb. 2003.
[9] T. Melia and S. Rickard, "Underdetermined blind source separation in echoic environments using DESPRIT," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, pp. 90–109, 2007.
[10] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.
[11] T. Mei, A. Mertins, F. Yin, J. Xi, and J. F. Chicharo, "Blind source separation for convolutive mixtures based on the joint diagonalization of power spectral density matrices," *Signal Process.*, vol. 88, no. 8, pp. 1990–2007, 2008.
[12] C. Serviere and D. T. Pham, "Permutation correction in the frequency domain in blind separation of speech mixtures," *EURASIP J. Appl. Signal Process.*, vol. 2006, no. 1, pp. 177–193, 2006.
[13] 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, 2001.
[14] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. IEEE Int. Symp. Circuits Syst.*, New Orleans, LA, 2007, pp. 3247–3250.
[15] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, Jan. 2007.
[16] I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Process.*, vol. 87, no. 8, pp. 1859–1871, 2007.
[17] 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 Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.
[18] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: Investigation and solutions," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 1–13, Jan. 2005.
[19] E. Bingham and A. Hyvarien, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Syst.*, vol. 10, pp. 1–8, 2000.
[20] A. J. Bell and T. J. Sejonwski, "An information maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.
[21] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 1996, vol. 8, pp. 757–763.
[22] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. 2001 Int. Workshop Ind. Compon. Anal. Blind Signal Separat.*, San Diego, CA, 2001, pp. 722–727.
[23] S. C. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Honolulu, HI, 2007, vol. 2, pp. 637–640.
[24] H. Sawada, R. Mukai, S. Kethulle, S. Araki, and S. Makino, "Spectral smoothing for frequency-domain blind source separation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Kyoto, Japan, 2003, pp. 311–314.
[25] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speechsignals in a reverberant environment," in *Proc. IEEE Int. Conf. Acoust. , Speech, Signal Process.*, Istanbul, Turkey, 2000, vol. 2, pp. 1041–1044.
[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.



**Lin Wang** was born in Anhui, China, in 1981. He received the B.S. degree in electronic engineering from Tianjin University, Tianjin, China, in 2003 and the Ph.D. degree in signal processing from Dalian University of Technology, Dalian, China, in 2010.

He has been a Visiting Fellow with the Institute for Microstructural Sciences, National Research Council Canada, for 13 months, from 2008 to 2009. His research interests include video and audio compression, blind source separation, and 3-D audio processing.

**Heping Ding** (SM'03) received the Ph.D. degree in electrical engineering from Queen's University, Kingston, ON, Canada, in 1992 and the M.Sc. degree in information physics from Nanjing University, Nanjing, China, in 1984.

He is currently a Research Council Officer with the Institute for Microstructural Sciences, National Research Council, Ottawa, ON, Canada. Since 2003, he has been an Adjunct Professor and a part-time Professor with the School of Information Technology and Engineering (SITE) at the University of Ottawa, Canada, and an Adjunct Professor at the School of Graduate Studies, Ryerson University, Canada. He has also served in Bell Northern Research and Nortel, Canada, as a Senior Audio Signal Processing Specialist. He holds 13 patents and has over 40 major publications in fields of subjective audio, telephony, adaptive filtering, echo control, non-stationary signal processing, noise reduction, audio coding, and electronic equipment.

Dr. Ding is a member of the Signal Processing Oriented Technologies (SPOT) group, University of Ottawa. He has received 17 awards from industry, academia, and science and technology communities.

**Fuliang Yin** was born in Fushun city, Liaoning province, China, in 1962. He received the B.S. degree in electronic engineering and the M.S. degree in communications and electronic systems from Dalian University of Technology (DUT), Dalian, China, in 1984 and 1987, respectively.

He joined the Department of Electronic Engineering, DUT, as a Lecturer in 1987 and became an Associate Professor in 1991. He has been a Professor at DUT since 1994, and the Dean of the School of Electronic and Information Engineering of DUT since 2000. His research interests include digital signal processing, speech processing, image processing, broadband wireless communication, and integrated circuit design.