Blind Source Separation Exploiting Higher-Order Frequency Dependencies

Taesu Kim, *Student Member, IEEE*, Hagai T. Attias, Soo-Young Lee, *Member, IEEE*, and Te-Won Lee, *Member, IEEE*

***Abstract*—Blind source separation (BSS) is a challenging problem in real-world environments where sources are time delayed and convolved. The problem becomes more difficult in very reverberant conditions, with an increasing number of sources, and geometric configurations of the sources such that finding directionality is not sufficient for source separation. In this paper, we propose a new algorithm that exploits higher order frequency dependencies of source signals in order to separate them when they are mixed. In the frequency domain, this formulation assumes that dependencies exist between frequency bins instead of defining independence for each frequency bin. In this manner, we can avoid the well-known frequency permutation problem. To derive the learning algorithm, we define a cost function, which is an extension of mutual information between multivariate random variables. By introducing a source prior that models the inherent frequency dependencies, we obtain a simple form of a multivariate score function. In experiments, we generate simulated data with various kinds of sources in various environments. We evaluate the performances and compare it with other well-known algorithms. The results show the proposed algorithm outperforms the others in most cases. The algorithm is also able to accurately recover six sources with six microphones. In this case, we can obtain about 16-dB signal-to-interference ratio (SIR) improvement. Similar performance is observed in real conference room recordings with three human speakers reading sentences and one loudspeaker playing music.**

***Index Terms*—Blind source separation (BSS), cocktail party problem, convolutive mixture, frequency domain, higher order dependency, independent component analysis, permutation problem.**

I. Introduction

In RECENT years, recovering the original source signals from observed signals without knowing the mixing process, so called blind source separation (BSS), has attracted a number of researchers. BSS is relevant to many applications including speech enhancement for noise robust speech recognition, crosstalk separation in telecommunication, high-quality hearing aids equipment, analyzing biological signals such as electroencephalograph (EEG) and magnetoencephalograph (MEG). The fundamental assumption in the BSS problem is that the source signals are statistically independent.

Independent component analysis (ICA) is the method to find statistically independent sources from mixtures of sources by utilizing higher order statistics [1]–[3]. In its simplest form, the ICA model assumes linear instantaneous mixing without sensor noise and the number of sources being equal to the number of sensors. When trying to solve the problem of separating source signals mixed in a real environment, those assumptions are not valid and model extensions are required. In those cases, observed signals are not instantaneous mixtures of sources, but convolutive mixtures, which mean that they are mixed with time delays and convolutions. In order to deal with convolved mixtures, the ICA model formulation and the learning algorithm have been extended to convolutive mixtures in both the time and the frequency domains [4]–[9]. Those models are known as solutions to the multichannel blind deconvolution problem. In the case of the time domain approach, solutions usually require intensive computations with long de-reverberation filters, and the resulting unmixed source signals are whitened due to the independent and identically distributed (i.i.d.) assumption [5]. Slow convergence speed especially for colored signals has been observed. The computational load and slow convergence can be overcome by the frequency domain approach, in which multiplication at each frequency bin replaces convolution operation in the time domain. Thus, one can apply the ICA algorithm to instantaneous mixtures in each frequency bin. Although this may be attractive, the main problem then is the permutation of the ICA solutions over different frequency bins due to the indeterminacy of permutation inherent in the ICA algorithm. One should correct the permutations of separating matrices at each frequency so that the separated signal in the time domain is reconstructed properly.

Various approaches have been proposed to solve the permutation problem. A popular approach is to impose a smoothness constraint of the source that translates into smoothing the separating filter. This approach has been realized by several techniques such as averaging separating matrices with adjacent frequencies [9], limiting the filter length in the time domain [10], or considering the coherency of separating matrices at adjacent frequencies [11]. Another related approach is based on direction of arrival (DOA) estimation which is much used in array signal processing. By analyzing the directivity patterns formed by a

Manuscript received February 1, 2005; revised December 6, 2005. The work of T. Kim and S.-Y. Lee were supported in part by the Chung Moon Soul Center for Bioinformation Bioelectronics and in part by the Brain Neuroinformatics Research Program, Korean Ministry of Science and Technology. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Bhiksha(GE) Raj.

T. Kim is with the Deptartment of Biosystems, Korea Advanced Institute of Science and Technology (KAIST), Dajeon 305-701, Korea and also with the Institute for Neural Computation, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: taesu.kim@kaist.ac.kr; taesu@ucsd.edu).

H. T. Attias is with Golden Metallic, Inc., San Francisco, CA 94147 USA (htattias@goldenmetallic.com).

S.-Y. Lee is with the Deptartment of Biosystems, Korea Advanced Institute of Science and Technology, Dajeon 305-701, Korea (e-mail: sylee@kaist.ac.kr).

T.-W. Lee is with the Institute for Neural Computation, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: tewon@ucsd.edu).

Digital Object Identifier 10.1109/TASL.2006.872618

separating matrix, source directions can be estimated and, therefore, permutations can be aligned [12]–[14]. When the sources are nonstationary signals, one can employ the interfrequency correlations of signal envelopes to align permutations [15], [16]. Although these methods perform well under certain specific conditions, there is no method that performs well in general conditions. Moreover, in the case of an ill-posed problem, e.g., the case that each mixing filter of the source is very similar, the sources are located close to each other, or DOA of the sources are similar, the methods developed so far fail to separate the source signals.

In this paper, we propose a novel approach for BSS by focusing on a new cost function and a dependency model which captures interfrequency dependencies in data. These dependencies are related to an improved model for the source signal prior. While the source priors are defined as independent priors at each frequency bin in conventional algorithms, we utilize higher order dependencies across frequency. In this manner, we define each source prior as a multivariate super-Gaussian distribution, which is only a simple extension of the independent Laplacian distribution. The algorithm itself is able to preserve higher order dependencies and structures of frequencies. Therefore, the permutation problem is completely avoided, and the separation performances are comparably high even in severely ill-posed conditions.

II. Proposed Method

The proposed method consists of the mixing and separating procedure in a convolutive environment, the definition of a cost function, and an algorithm for learning the parameters of the separating filters. We define a new cost function for the frequency domain BSS, and derive its learning algorithm by minimizing it. Notation used in this paper is defined below.[1]

A. Model

We define the relationship between the sources and observations. Let $x_i(t)$ be the ith observation signal at time t

$$x_i(t) = \sum_{j=1}^{L} \sum_{\tau=0}^{T-1} h_{ij}(\tau) s_j(t-\tau) \tag{1}$$

where $h_{ij}(t)$ is a time domain transfer function from the jth source to the ith observation, which has T length in time, $s_j(t)$ is the jth source signal at time t, and L is the number of sources.

[1]We use plain lower-cased characters to denote scalar variables, bold-faced, lower-cased characters to denote vector variables, and upper-cased characters to denote matrix variables. Superscript indicates a frequency bin, and subscript indicates a source or an observation. For example, $\mathbf{x}_i$ is the ith observation vector that consists of K frequency bins, $[x_i^{(1)} \cdots x_i^{(K)}]^{\mathsf{T}}$. $\mathbf{x}^{(k)}$ is an observation vector at the kth frequency bin, which consists of M observations at the kth frequency bin, $[x_1^{(k)} \cdots x_M^{(k)}]^{\mathsf{T}}$. $H^{(k)} \equiv \{h_{ij}^{(k)}\}$ means that $h_{ij}^{(k)}$ is the ith row, jth column element of the matrix $H^{(k)}$. $x_i^{(k)}[n]$ denotes the nth sample of random variable $x_i^{(k)}$. $x_i^{\star(k)}$ denotes the complex conjugate of $x_i^{(k)}$, and $\mathbf{x}_i^{\dagger}$ denotes the conjugate transpose of $\mathbf{x}_i$.

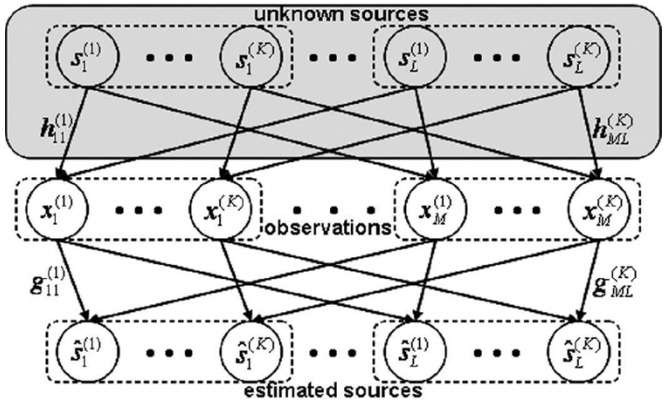


Fig. 1. Mixing and separating model for the frequency domain BSS. Observed signals at the kth frequency bin and a certain time frame are denoted by a vectorized form as $\mathbf{x}^{(k)} = H^{(k)}\mathbf{s}^{(k)}$, and estimated sources are denoted as $\hat{\mathbf{s}}^{(k)} = G^{(k)}\mathbf{x}^{(k)}$, where $H^{(k)} \equiv \{h_{ij}^{(k)}\}$ is the mixing matrix, and $G^{(k)} \equiv \{g_{ij}^{(k)}\}$ is the separating matrix at the kth frequency bin. Thus, the joint probability of estimated sources is given as $p(\hat{\mathbf{s}}^{(1)} \cdots \hat{\mathbf{s}}^{(K)}) = p(\mathbf{x}^{(1)} \cdots \mathbf{x}^{(K)}) / \prod_k |\det G^{(k)}|$.

By executing short-time Fourier transform, the time domain signal $x_i(t)$ is converted to the frequency domain signal $x_i^{(k)}[n]$

$$x_i^{(k)}[n] = \sum_{t=0}^{K-1} w(t) x_i(nJ+t) e^{-j\omega_k t} \tag{2}$$

where $\omega_k = 2\pi(k-1)/K$, $k = 1, 2 \cdots K$, J is shift size, and $w(t)$ is a window function. If the window length K is sufficiently longer than the length of the mixing filter $h_{ij}(t)$, the convolution in the time domain is approximately converted to multiplication in the frequency domain as follows:

$$x_i^{(k)}[n] \approx \sum_{j=1}^{L} h_{ij}^{(k)} s_j^{(k)}[n]. \tag{3}$$

If the separating filter matrices exist, that is, the inverses or pseudoinverses of mixing matrices at each frequency exist ($L \leq M$), then the separated ith source signal is given as

$$\hat{s}_i^{(k)}[n] = \sum_{j=1}^{M} g_{ij}^{(k)} x_j^{(k)}[n] \approx s_i^{(k)}[n] \tag{4}$$

where $g_{ij}^{(k)}$ is the separating filter at the kth frequency bin, and M is the number of observed signals. Mixing and separating model is shown in Fig. 1, where both sources and observations are multivariate.

B. Cost Function

In order to separate multivariate sources from multivariate observations, we need to define a cost function for multivariate random variables. Here, we define Kullback–Leibler divergence between two functions as the measure of independence. One is an exact joint probability density function $p(\hat{\mathbf{s}}_1 \cdots \hat{\mathbf{s}}_L)$, and the other is a nonlinear function which is the product of approximated probability density functions of individual source vectors

$\prod_{i=1}^{L} q(\hat{\mathbf{s}}_i)$. We would think of it as an extension of mutual information between multivariate random variables

$$\begin{aligned}\mathcal{C} &= \mathcal{KL}\left(p(\hat{\mathbf{s}}_1 \cdots \hat{\mathbf{s}}_L) \| \prod_{i=1}^{L} q(\hat{\mathbf{s}}_i)\right) \\ &= \int p(\hat{\mathbf{s}}_1 \cdots \hat{\mathbf{s}}_L) \log \frac{p(\hat{\mathbf{s}}_1 \cdots \hat{\mathbf{s}}_L)}{\prod_{i=1}^{L} q(\hat{\mathbf{s}}_i)} \mathrm{d}\hat{\mathbf{s}}_1 \cdots \mathrm{d}\hat{\mathbf{s}}_L \\ &= \int p(\mathbf{x}_1 \cdots \mathbf{x}_M) \log p(\mathbf{x}_1 \cdots \mathbf{x}_M) \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_M \\ &\quad - \sum_{k=1}^{K} \log \left|\det G^{(k)}\right| - \sum_{i=1}^{L} \int p(\hat{\mathbf{s}}_i) \log q(\hat{\mathbf{s}}_i) \mathrm{d}\hat{\mathbf{s}}_i \\ &= const. - \sum_{k=1}^{K} \log \left|\det G^{(k)}\right| - \sum_{i=1}^{L} E \log q(\hat{\mathbf{s}}_i). \end{aligned} \quad (5)$$

$\int p(\mathbf{x}_1 \cdots \mathbf{x}_M) \log p(\mathbf{x}_1 \cdots \mathbf{x}_L) \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_M$ is the entropy of the given observations, which is a constant. Note that the random variables in above equations are multivariate. The interesting parts of this cost function are that each source is multivariate and it would be minimized when the dependency between the source vectors is removed but the dependency between the components of each vector does not need to be removed. Therefore, the cost function preserves the inherent frequency dependency within each source, but it removes dependency between the sources.

C. Learning Algorithm: A Gradient Descent Method

Now that we defined the cost function, derivation of the learning algorithm is straightforward. Here, we are using a gradient descent method to minimize the cost function. By differentiating the cost function $\mathcal{C}$ with respect to the coefficients of the separating matrices $g_{ij}^{(k)}$, we can obtain the gradients for the coefficients as follows:

$$\Delta g_{ij}^{(k)} = -\frac{\partial \mathcal{C}}{\partial g_{ij}^{(k)}} = g_{ij}^{-\dagger(k)} - E\varphi^{(k)}\left(\hat{s}_i^{(1)} \cdots \hat{s}_i^{(K)}\right) x_j^{\star(k)} \quad (6)$$

where $(G^{(k)^{-1}})^{\dagger} \equiv \{g_{ij}^{-\dagger(k)}\}$. By multiplying scaling matrices $G^{(k)^{\dagger}} G^{(k)}$ to the gradient matrices $\Delta G^{(k)} \equiv \{\Delta g_{ij}^{(k)}\}$, we can obtain the natural gradient, which is well known as a fast convergence method [17]

$$\Delta g_{ij}^{(k)} = \sum_{l=1}^{L} \left(I_{il} - E\varphi^{(k)}\left(\hat{s}_i^{(1)} \cdots \hat{s}_i^{(K)}\right) \hat{s}_l^{\star(k)}\right) g_{lj}^{(k)} \quad (7)$$

where I_{il} is 1 only when $i = l$, otherwise 0, and the nonlinear function $\varphi^{(k)}(\cdot)$ is given as

$$\varphi^{(k)}\left(\hat{s}_i^{(1)} \cdots \hat{s}_i^{(K)}\right) = -\frac{\partial \log q\left(\hat{s}_i^{(1)} \cdots \hat{s}_i^{(K)}\right)}{\partial \hat{s}_i^{(k)}}. \quad (8)$$

We would term it multivariate score function corresponding to the score function in the conventional ICA. Section III discusses the multivariate score function in detail. We can update the coefficients of separating matrices with either the batch update rule or the online update rule. The batch update rule is given as

$$g_{ij}^{(k)^{new}} = g_{ij}^{(k)^{old}} + \eta \Delta g_{ij}^{(k)} \quad (9)$$

where η is learning rate. The online update rule can be obtained by omitting the expectation in (7) and updating at every sample time.

D. Scaling Problem and Overlap Add

Although our approach avoids the permutation problem by exploiting the higher order frequency dependencies, the scaling problem still needs to be solved. If the sources are stationary and the variances of the sources are known in all frequency bins, one can solve it by adjusting the variances to the known values. However, natural signal sources are dynamic, nonstationary in general, as well as we do not know the variances. Instead of adjusting the source variances, we can solve the scaling problem by adjusting the learned separating filter matrix. A well-known method is obtained by the minimal distortion principle [18]. Once the learning algorithm is finished, the learned separating filter matrix is an arbitrary scaled version of the exact one, which is given as

$$G^{(k)} = D^{(k)} H^{-1(k)} \quad (10)$$

where $D^{(k)}$ is an arbitrary diagonal matrix. Therefore, by replacing the separating filter matrix as

$$G^{(k)} \leftarrow \mathrm{diag}\left(G^{-1(k)}\right) G^{(k)} \quad (11)$$

where $\mathrm{diag}(X)$ denotes the diagonal matrix of the matrix X, we can obtain the separating filter matrix that has reasonable scales

$$G^{(k)} = \mathrm{diag}\left(H^{(k)}\right) H^{-1(k)}. \quad (12)$$

After solving the scaling problem, we calculate the finally separated sources in the frequency domain by (4). Then, we perform an inverse Fourier transform and overlap add to reconstruct the time domain signal as follows:

$$\hat{s}_i(t) = \sum_{n=0}^{N-1} \sum_{k=1}^{K} \hat{s}_i^{(k)}[n] e^{j\omega_k(t-nJ)} \quad (13)$$

where ω_k, K, and J are same as those in (2). In the case of using a Hanning window, the window effect can be avoided by setting shift size J to 1/4 of the window length K.

III. Multivariate Score Function

In the previous section, one can notice that the only difference between our approach and the conventional ICA is the form of the score function. If we define the multivariate score function $\varphi^{(k)}(\hat{s}_i^{(1)} \cdots \hat{s}_i^{(K)})$ as a single-variate score function $\varphi(\hat{s}_i^{(k)})$, the algorithm is converted to the same as the conventional ICA. Therefore, the fact that the score function

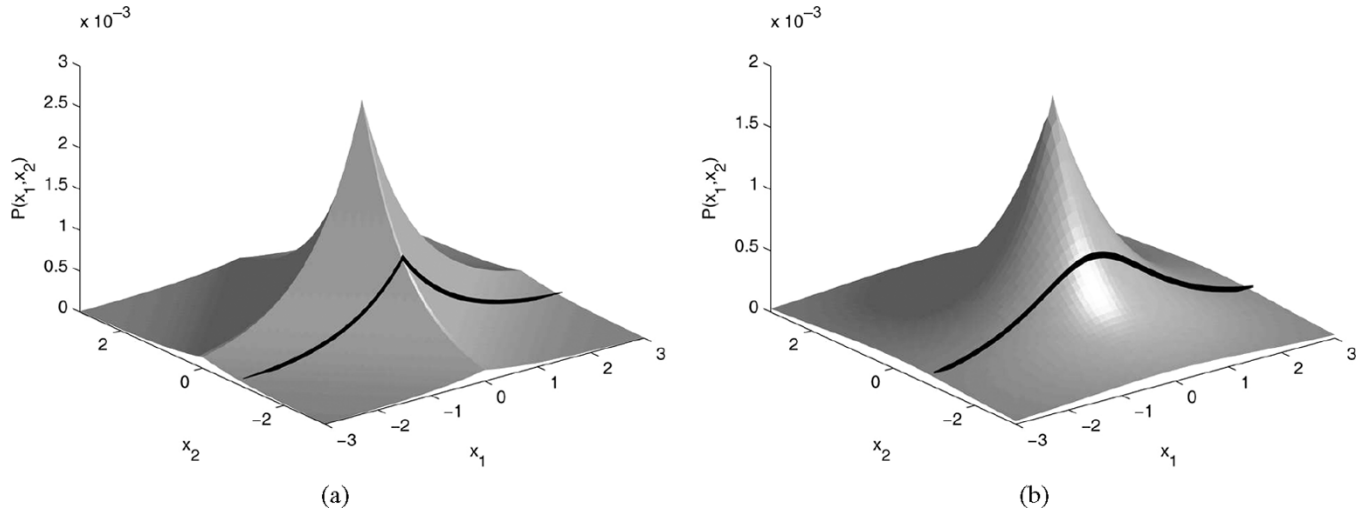


Fig. 2. Comparison between (a) an independent Laplacian distribution and (b) a dependent super-Gaussian distribution. The figure shows the dependency between only two arbitrary elements of a multidimensional variable $\mathbf{s} = [s^{(1)} \cdots s^{(K)}]^{\mathsf{T}}$. x_1 can be considered as either real or imaginary part of $s^{(1)}$, and also x_2 can be considered as either real or imaginary part of $s^{(2)}$. The black line indicates $p(x_1|x_2 = 1)$. In (a), the probability of x_1 always has Laplacian distribution regardless of x_2. In (b), however, the probability of x_1 given $x_2 = 1$ does not have Laplacian distribution even though the probability of x_1 given $x_2 = 0$ has Laplacian distribution.

is a multivariate function is the most important point in our approach. According to many ICA literatures, a score function is closely related to a source prior. For example, when the sources have super-Gaussian distribution, Laplacian distribution is widely used as a source prior. Here, a multivariate score function is also closely related to a source prior, because the cost function in Section II-B includes $q(\hat{\mathbf{s}}_i)$, which is an approximated probability density function of a source vector, that is, $q(\mathbf{s}_i) \approx p(\mathbf{s}_i)$. Thus, as shown in (8), a multivariate score function can be obtained by differentiating the log prior with respect to each element of a source vector.

In most BSS approaches, the source prior for a super-Gaussian signal is defined by a Laplacian distribution. Thus, suppose that the source prior of a vector is independent Laplacian distribution in each frequency bin. This can be written as

$$p(\mathbf{s}_i) = \prod_{k=1}^{K} p\left(s_i^{(k)}\right) = \alpha \prod_{k=1}^{K} \exp\left(-\frac{\left|s_i^{(k)} - \mu_i^{(k)}\right|}{\sigma_i^{(k)}}\right) \quad (14)$$

where α is a normalization term, and $\mu_i^{(k)}$ and $2(\sigma_i^{(k)})^2$ are a mean and a variance of the ith source signal at the kth frequency bin, respectively. Assuming zero mean and unit variance, the score function is given as

$$\varphi^{(k)}\left(\hat{s}_i^{(1)} \cdots \hat{s}_i^{(K)}\right) = \frac{\partial \sum_{k=1}^{K} \left|\hat{s}_i^{(k)}\right|}{\partial \hat{s}_i^{(k)}} = \frac{\hat{s}_i^{(k)}}{\left|\hat{s}_i^{(k)}\right|} = \exp\left(j \cdot \arg\left(\hat{s}_i^{(k)}\right)\right). \quad (15)$$

Indeed, (15) is not a multivariate function, because the function depends on only a single variable $\hat{s}_i^{(k)}$. Therefore, instead of using an independent prior, we have to define a new prior, which is highly dependent on the other elements of a source vector.

In our approach, we defined the source prior as a dependent multivariate super-Gaussian distribution, which can be written as

$$p(\mathbf{s}_i) = \alpha \exp\left(-\sqrt{(\mathbf{s}_i - \mu_i)^{\dagger} \Sigma_i^{-1} (\mathbf{s}_i - \mu_i)}\right) \quad (16)$$

where μ_i and Σ_i are a mean vector and a covariance matrix of the ith source signal, respectively. Fig. 2 shows the difference between the assumption of independent Laplacian distribution and a dependent multivariate super-Gaussian distribution. In Fig. 2(b), the joint distribution of x_1 and x_2 does not display any directionality which means x_1 and x_2 are uncorrelated. However, the marginal distribution of x_1 is different from the joint distribution of x_1 given x_2, that is, x_1 and x_2 are highly dependent. In contrast to the distribution shown in Fig. 2(a), Fig. 2(b) has a radial shape, which is similar to Gaussian distribution, but has higher peak and heavier tail. Thinking in a different way, one can notice that the distribution shown in Fig. 2(b) can be obtained by a scale mixture of Gaussians with a fixed mean and a variable variance, as we describe next.

Suppose that there is a K-dimensional random variable, which is defined by

$$\mathbf{s}_i = \sqrt{v} \cdot \mathbf{z}_i + \mu_i \quad (17)$$

where v is a scalar random variable, $\mathbf{z}_i$ is a K-dimensional random variable, and μ_i is a K-dimensional deterministic variable. Here, the random variable, $\mathbf{z}_i$, has Gaussian distribution with zero mean and Σ_i covariance matrix

$$p(\mathbf{z}_i) = \alpha_z \exp\left(-\frac{\mathbf{z}_i^{\dagger} \Sigma_i^{-1} \mathbf{z}_i}{2}\right) \quad (18)$$

where α_z is a normalization term. Suppose that v has a kind of Gamma distribution as follows:

$$p(v) = \alpha_v v^{\frac{(K-1)}{2}} \exp\left(-\frac{v}{2}\right) \tag{19}$$

where α_v is a normalization term. Then, the random variable $\mathbf{s}_i$ given v has Gaussian distribution. Its mean and covariance are μ_i and $v\Sigma_i$, respectively. In this model, the distribution we used can be obtained by integrating joint distribution of $\mathbf{s}_i$ and v over v as follows:

$$\begin{aligned} p(\mathbf{s}_i) &= \int_0^\infty p(\mathbf{s}_i|v)p(v)\mathrm{d}v \\ &= \hat{\alpha}\int_0^\infty \sqrt{v}\exp\left(-\frac{1}{2}\left(\frac{(\mathbf{s}_i-\mu_i)^\dagger\Sigma_i^{-1}(\mathbf{s}_i-\mu_i)}{v}+v\right)\right)\mathrm{d}v \\ &= \alpha\exp\left(-\sqrt{(\mathbf{s}_i-\mu_i)^\dagger\Sigma^{-1}(\mathbf{s}_i-\mu_i)}\right). \end{aligned} \tag{20}$$

Therefore, each component of $\mathbf{s}_i$ is not only correlated to others caused by Σ_i, but also has variance dependency generated by v. Even though we assume the covariance matrix Σ_i is identity, that is, each component of $\mathbf{s}_i$ is uncorrelated, the components are dependent on each other. Most natural signals have inherent dependencies between frequency bins such as variance dependency we modeled above. In other words, when one frequency component has a larger variance, the other frequency components have larger variances as well. Nonetheless, each frequency bin is uncorrelated to the others, because the Fourier bases are orthogonal bases. Thus, we can set the covariance term Σ_i as a diagonal matrix. Since Fourier outputs have zero means, we can rewrite (16) as follows:

$$p(\mathbf{s}_i) = \alpha\exp\left(-\sqrt{\sum_k\left|\frac{s_i^{(k)}}{\sigma_i^{(k)}}\right|^2}\right) \tag{21}$$

where $\sigma_i^{(k)}$ is the variance of the ith source at the kth frequency bin, which determines the scale of each element of a source vector. In the algorithm, we set $\sigma_i^{(k)}$ to 1, because we adjust the scale as in Section II-D after learning separating filters. Consequently, the multivariate score function we used is given as

$$\varphi^{(k)}\left(\hat{s}_i^{(1)}\cdots\hat{s}_i^{(K)}\right) = \frac{\partial\sqrt{\sum_{k=1}^K\left|\hat{s}_i^{(k)}\right|^2}}{\partial\hat{s}_i^{(k)}} = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^K\left|\hat{s}_i^{(k)}\right|^2}}. \tag{22}$$

Although we used a fixed form of a multivariate score function such as (22), we do not claim that only this form is appropriate for separating source signals. Since the form of a multivariate score function is related to dependency of sources, the proper form of a multivariate score function might vary with different types of dependency. Designing proper multivariate score functions for various dependency models would be promising future work.

IV. Experimental Results

We evaluated the performance of the proposed algorithm using both simulated and real data. Simulated data were obtained by simulating impulse responses of a rectangular room based on the image model technique [19]–[21]. To generate the microphone signals, we used 8-s-long audio signals sampled at 8 kHz, and they were convolved with corresponding room impulse responses. The proposed algorithm was compared with two well-known frequency domain BSS algorithms, such as ICA algorithm with permutation correction and Parra and Spence's algorithm [10].[2] The former method was obtained by using the conventional score function given as (15). To solve the permutation problem in this method, we used the method to consider DOA and interfrequency correlations together [14]. Parra and Spence's algorithm avoids the permutation problem by limiting the length of the filter in the time domain to smooth the shape of the filter in the frequency domain, while learning the separating filters. The performances were measured by signal-to-interference ratio (SIR) in decibels defined as

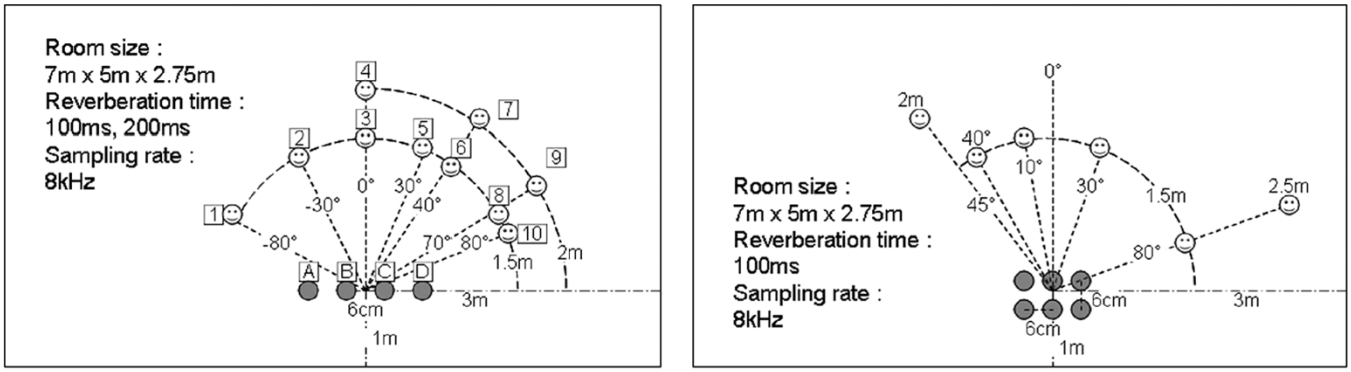
$$\mathrm{SIR}_{\mathrm{in}} = 10\log\left(\frac{\sum_{n,k}\left|\sum_i h_{iq(i)}^{(k)}\hat{s}_{q(i)}^{(k)}[n]\right|^2}{\sum_{n,k}\left|\sum_{i\neq j} h_{iq(j)}^{(k)}\hat{s}_{q(j)}^{(k)}[n]\right|^2}\right) \tag{23}$$

$$\mathrm{SIR}_{\mathrm{out}} = 10\log\left(\frac{\sum_{n,k}\left|\sum_i r_{iq(i)}^{(k)}\hat{s}_{q(i)}^{(k)}[n]\right|^2}{\sum_{n,k}\left|\sum_{i\neq j} r_{iq(j)}^{(k)}\hat{s}_{q(j)}^{(k)}[n]\right|^2}\right) \tag{24}$$

where $q(i)$ indicates separated source index that the ith source appears, and $r_{iq(j)}^{(k)}$ is an overall impulse response, which is defined by $\sum_m g_{im}^{(k)}h_{mq(j)}^{(k)}$. Real data were obtained in an ordinary conference room, where human speakers read several sentences and a loudspeaker played music. In all experiments, we used a 1024-point fast Fourier transform (FFT) and Hanning window to convert time domain signals to the frequency domain. The length of window was 1024 samples and shift size was 256 samples. Initial values for both the proposed and the conventional ICA based algorithm were chosen as whitening matrix in each frequency bin. The algorithm ran until the decrement of the cost function was less than 10^{-6}. For Parra and Spence's algorithm, we used the same number of FFT points and limited the length of time domain filter to 256, which provided the best performances.

First, the proposed algorithm was applied to the problem with two microphones and two sources in simulated room environments. We assumed that the room size was $7 \times 5 \times 2.75$ m. For an intensive analysis, we evaluated the performances with a number of source locations and reverberation times. All the heights of sources and microphones were 1.5 m. The environments are shown in Fig. 3(a), in which microphone B and C were used, and we chose seven pairs of source locations. Although two cases of locations such as 1 and 8, and 2 and 6 are comparably easier cases, 5 and 6, and 8 and 10 are more difficult cases because the sources are located on the same side and have similar DOAs. The other three cases such as 3 and 4, 6 and

[2]The code was downloaded from http://ida.first.gmd.de/~harmeli/download/download_convbss.html



(a) (b)

Fig. 3. Simulated room environments. All the heights of sources and microphones were 1.5 m. Several combinations of source locations were selected in (a). Two microphones (B,C) were used for the case of two sources, and four microphones (A,B,C,D) were used for the case of four sources. Environment in (b) is a more challenging case. There were six microphones and six sources, in which some of the sources located very close to another or have the same DOAs.

TABLE I
COMPUTATIONAL TIME

	Time per iter.	# of iter.	Learning time
Proposed	0.1689s	200	33.78s
ICA	0.1623s	130	21.1s

7, and 8 and 9 are the most difficult cases, because the sources are located closely, and they also have the same DOAs. Fig. 4 shows the results of all cases with varying reverberation times, when one source was a male speech, and the other was a female speech. The average $\mathrm{SIR}_{\mathrm{in}}$ of all cases was about 0 dB. To evaluate the computational complexity, we measured the computational time when the sources were located 1 and 8, and the reverberation time was 100 ms. The proposed algorithm was coded in MATLAB and executed on an Intel Pentium IV 3-GHz processor. Table I shows the CPU time per iteration and the number of iterations until convergence. The computational time per iteration is almost similar to the ICA algorithm. Since the learning rule in each frequency bin is related to others, the convergence speed is a little slower.

Second, we tested the algorithms with four microphones and four sources. The environments were the same as the previous experiments, except for the locations of the sources and microphones. In this experiment, all the microphones in Fig. 3(a) were used, and five combinations of source locations were chosen. The sources were two male speeches and two female speeches. Fig. 5 shows the results of all cases, where the average $\mathrm{SIR}_{\mathrm{in}}$ of all cases was about -5 dB. As shown in Figs. 4 and 5, the proposed algorithm outperforms the others in most cases. Even in the worst cases, the others did not exceed the proposed algorithm by more than 2 dB. The ICA algorithm with permutation correction based on interfrequency correlation is not robust, because a misalignment of a permutation at a certain frequency bin may cause consecutive misalignments of neighboring frequency bins. By combining DOA and correlation, robustness can be improved a little. However, the DOA-based permutation correction is not precise, as the reverberation time increases or the source locations are close. Moreover, it completely fails when the sources have the same DOAs. So, the algorithm combining DOA and correlation is still not robust enough in some cases. Its performance can be severely bad in some cases although it performs best in certain cases. The severe disadvantage of Parra and Spence's algorithm is that it cannot use the full length of the filter, because it limits the filter length to avoid the permutation problem. Thus, the effective filter length in the time domain was 256, even though we used a 1024-point FFT. The performances of their algorithm were degraded more than others, as the source locations were difficult or the reverberation time was long. However, the proposed algorithm overcomes these disadvantages. Therefore, it does not limit the filter length, and it is very robust.

In addition to the previous experiments, we performed another experiment that shows how the performances are affected by the kind of sources. Instead of using only speech signals, we also used babble noise sound and classical music as source signals. We selected four different pairs of sources, male speech and female speech, male speech and classical music, female speech and babble noise, and babble noise and classical music. As shown in Fig. 6, the proposed algorithm outperformed the others in most cases. Therefore, the source model discussed in the Section III is appropriate not only to separate speeches but also to other signals that have interfrequency dependencies. To test the algorithm when there are many sources and microphones, we set up a six sources separation problem. The simulated room condition was the same as the previous experiments. Fig. 3(b) shows the room condition and the locations of the sources and microphones, in which some sources were located very closely, and other sources had the same DOAs. In this experiment, $\mathrm{SIR}_{\mathrm{in}}$ was -7 dB, and $\mathrm{SIR}_{\mathrm{out}}$ of the proposed algorithm was 9.5 dB. However, $\mathrm{SIR}_{\mathrm{out}}$ of the other algorithms did not exceed 0 dB, that is, they could not separate the sources. Fig. 7 shows separated source signals in the time domain, and Fig. 8 shows overall impulse responses.

Finally, we recorded real data in an ordinary conference room that has long reverberation time. Four microphones were located in a line. The sources consisted of three human speakers reading sentences, and a hip-hop music from a loud speaker, which were located approximately 1$\sim$2 m from the microphones. Three human speakers were located approximately 1$\sim$ 2 m from the microphones, and read several sentences. The

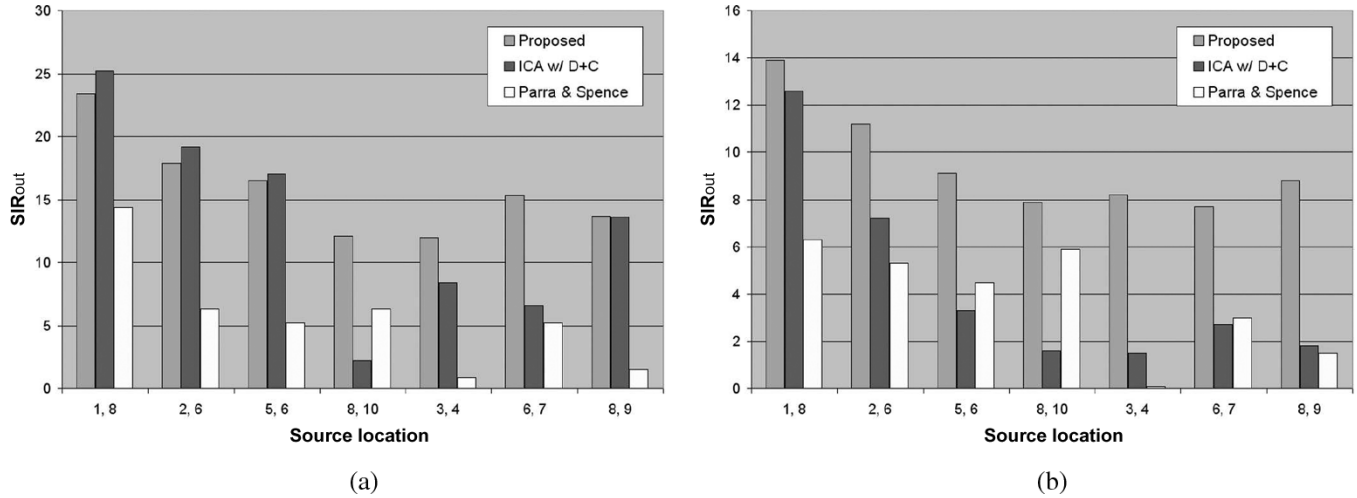


Fig. 4. Experimental results of two sources separation. $\mathrm{SIR}_{\mathrm{out}}$ was compared in various pairs of source locations and reverberation time with two sources and two microhpones (B,C), which are shown in Fig. 3(a). (a) 100-ms reverberation. (b) 200-ms reverberation.

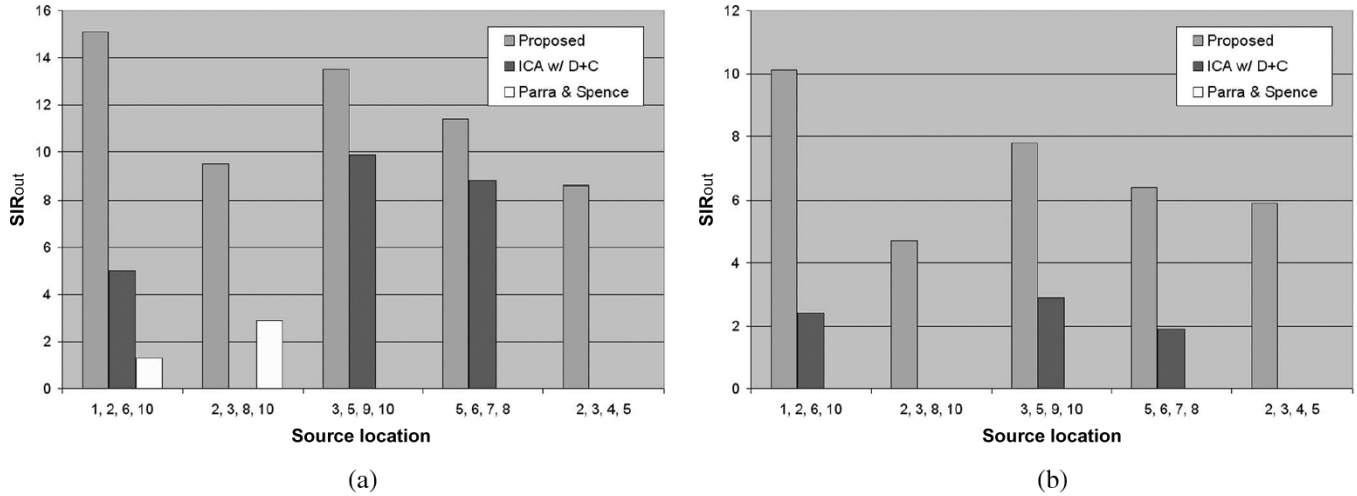


Fig. 5. Experimental results of four sources separation. $\mathrm{SIR}_{\mathrm{out}}$ was compared in various pairs of source locations and reverberation time with four sources and four microphones (A,B,C,D), which are shown in Fig. 3(a). In some cases, the other two algorithms failed to separate sources, where the bar was not shown. (a) 100-ms reverberation. (b) 200-ms reverberation.

approximate SIR improvement was about 14 dB. Audio files and more information are available on our web page.[3]

V. Discussion

So far, we have proposed a new cost function and a new source prior to derive the algorithm. Now that we have the algorithm, we can think about it in some interesting ways. There are several interesting viewpoints to this approach. On one hand, one can argue that a more precise source prior is helpful in finding a BSS solution. The defined source prior model though is still rough and assumes only a simple dependency between all frequencies. This prior model is therefore applicable to many natural signals since they all display certain dependencies and are not random. On the other hand, we can show that this approach tries to preserve higher order dependencies in data.

[3]http://ergo.ucsd.edu/~taesu/source_separation.html

Preserving those signal dependencies has shown its significance in applications where the independence assumption of sources across frequency is too strong and may not be realistic. Several approaches have been proposed that perform a variation of ICA by defining dependencies of the components [22]–[28]. Most of these approaches are to extract interesting features from data by unsupervised learning. None of those approaches considered modeling dependencies of sources in a convolved scenario. Interestingly, Hyvärinen and Hoyer's work [23], [24] is closely related to our source definition model. They defined the norm of each subspace output as a super-Gaussian distribution. In their approach, they were interested in modeling dependencies in image subspaces. Their results provide grouping of subspaces or features. A common feature of the dependency models is that they measure the variance of the source signal to approximate higher order dependencies in data.

Although it seems that we take two viewpoints in explaining our approach, namely the source prior and the dependency model, it is important to note that this model cannot be simply

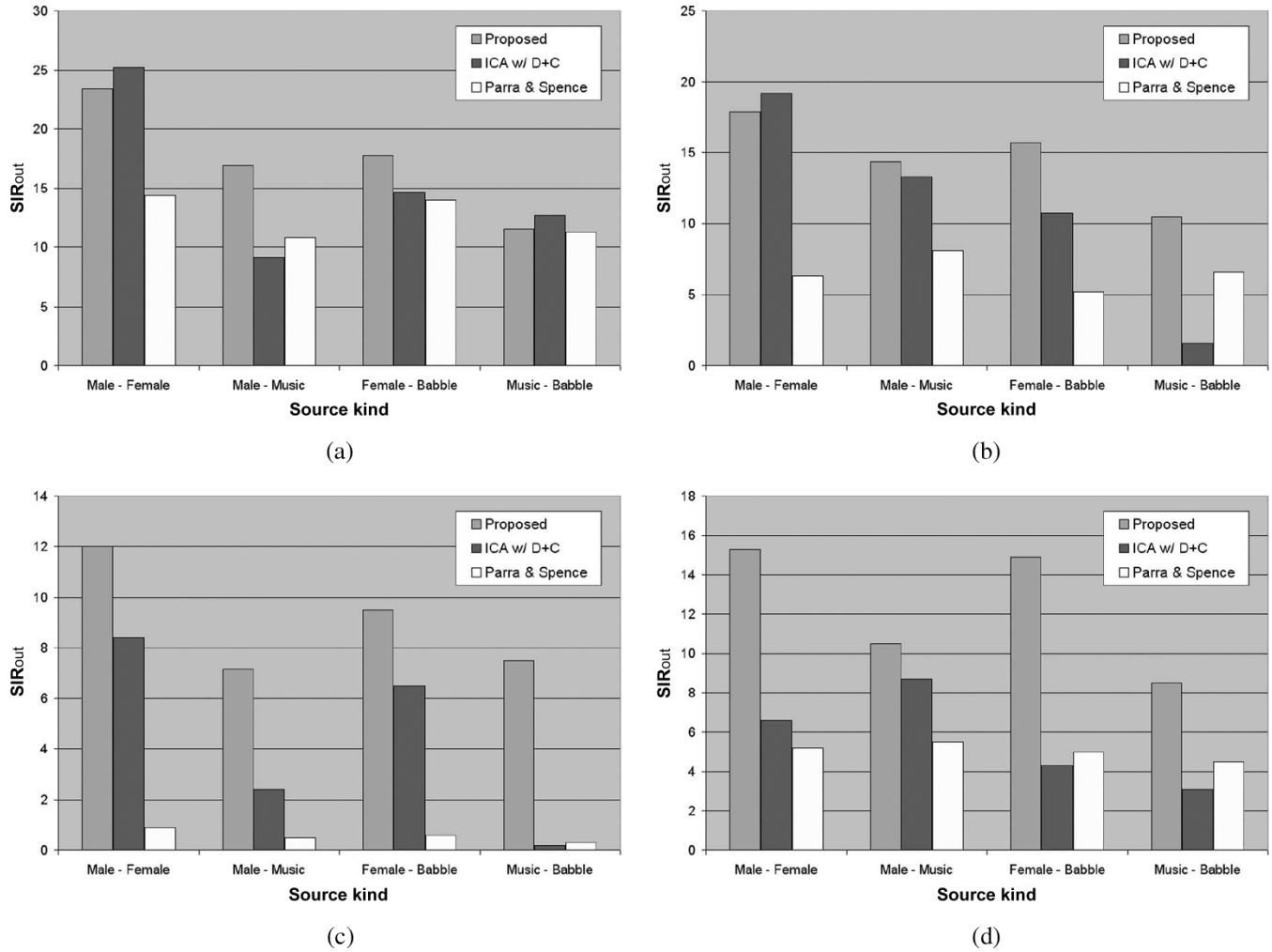


Fig. 6. Experimental results with simulated data (other sound sources). $\mathrm{SIR}_{\mathrm{out}}$ was compared varying kinds of sources. The environment was the same as Fig. 3(a) and the reverberation time was 100 ms. (a) Source location 1, 8. (b) Source location 2, 6. (c) Source location 3, 4. (d) Source location 6, 7.

reduced to a use of a different source prior. Our approach is better understood by utilizing higher order dependencies in data. For a given source estimate, our multivariate score function in the learning rule does not only depend on one frequency, but it includes all frequencies in a nonlinear way. This is again similar to nonlinear dependency models such as [23], [24], [27], and [28], where more precisely the nonlinear dependencies are considered.

One can also view this approach as a form of ICA for multivariate components. We have several observations which are mixed with independent source components, and each observation is a vector such as the output of the Fourier transform. Each source component is also a vector which has the same dimension as each observation. In this sense, we clearly exploit interfrequency dependencies inherent in the source signal. In terms of the subspace interpretation, we can consider each source vector as independent of the others, but the vector components of each source are highly dependent on each other. Therefore, the proposed algorithm may be considered as a generalization of the ICA algorithm to a vectorized form of the observations and sources. In a vector domain, especially the Fourier domain, BSS of a convolutive mixture in the time domain equals now BSS of an instantaneous mixture, which does not cause additional problems such as the permutation problem. A nice consequence of our approach in the frequency domain for BSS is that the use of dependent prior information avoids the permutation problem.

VI. Conclusion

We have proposed a new algorithm for BSS that exploits higher order frequency dependencies, leading to a generalization of the ICA algorithm to a vectorized form of observations and sources. Instead of defining independence for each frequency bin, we have assumed that source signals in the frequency domain have dependencies, which caused a multivariate score function. Simply, the only difference of the proposed algorithm from the conventional ICA algorithm is the fact that the score function is a multivariate function. The only additional computation is calculation of the multivariate score function. Additionally, it does not need to correct the permutation problem any more. Thus, the complexity of the algorithm is very low, and the computational load is not increased much. The experimental results have shown that the proposed algorithm is very robust and precise in most cases. Also, the

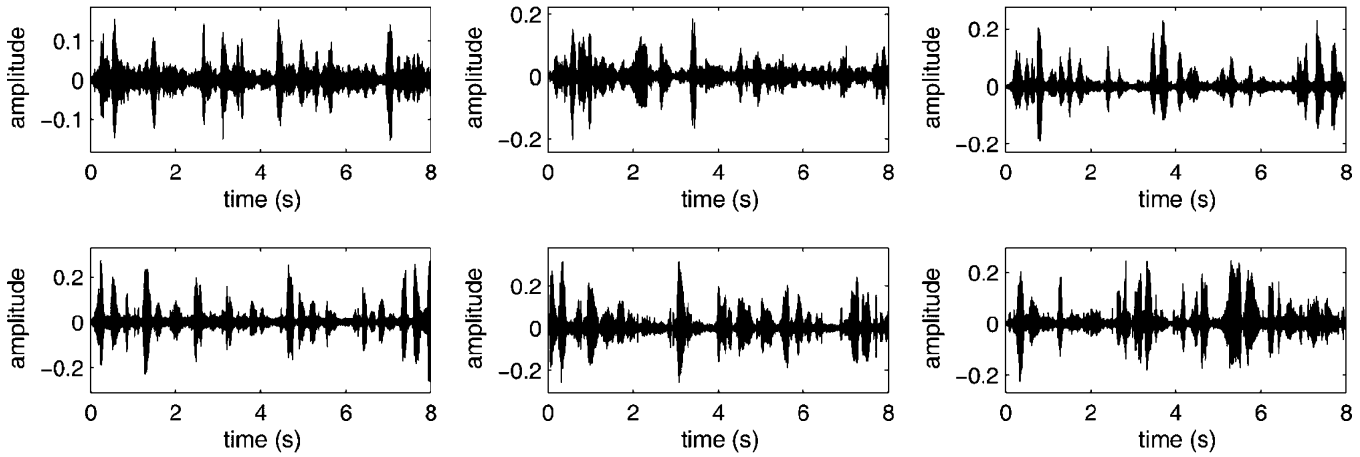


Fig. 7. Separated output signals from six input signals in the environment of Fig. 3(b). A 1024 sample sized Hanning window and 1024 FFT point was used.

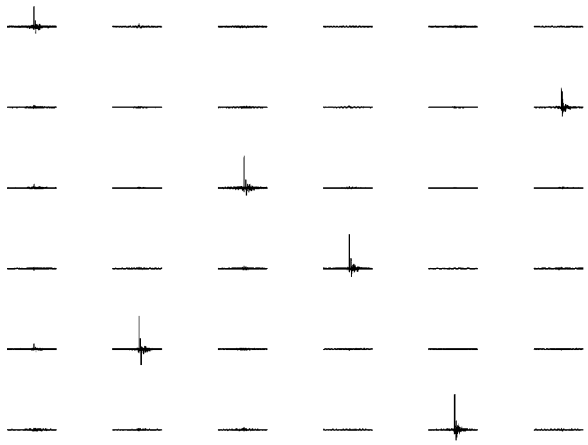


Fig. 8. Overall impulse responses, the impulse response located at ith row and jth column is $\sum_m \sum_\tau g_{i,m}(\tau) h_{m,j}(t-\tau)$, which implies the impulse response for jth source in ith output. For example, the third source is separated at the first output as shown in the first row.

proposed algorithm has shown good performance when separated six sources with six observations. Similar performance has been observed in real world recordings of four sources with four observations mixed in a conference room environment. The results suggest that exploiting higher order dependencies may be a key to solving challenging BSS problems.

ACKNOWLEDGMENT

The authors would like to thank all the reviewers for their constructive comments and feedback which improved the presentation of the paper significantly.

REFERENCES

[1] T.-W. Lee, *Independent Component Analysis: Theory and Applications*. Boston, MA: Kluwer, 1998.
[2] S. Roberts and R. Everson, *Independent Component Analysis: Principles and Practice*. Cambridge, U.K.: Cambridge Univ. Press, 2001.
[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley Interscience, 2001.
[4] D. Yellin and E. Weinstein, "Multichannel signal separation: methods and analysis," *IEEE Trans. Signal Process.*, vol. 44, no. 1, pp. 106–118, Jan. 1996.
[5] R. Lambert, "Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures," Ph.D. dissertation, Univ. Southern California, Los Angeles, 1996.
[6] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proc. IEEE Int. Workshop Neural Netw. Signal Process.*, 1996, pp. 423–432.
[7] T.-W. Lee, A. J. Bell, and R. Lambert, "Blind separation of convolved and delayed sources," *Adv. Neural Inf. Process. Syst.*, pp. 758–764, 1997.
[8] S. Weiß, "On adaptive filtering on oversampled subbands," Ph.D. dissertation, Signal Process. Div., Univ. Strathclyde, Glasgow, U.K., 1997.
[9] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomput.*, vol. 22, pp. 21–34, 1998.
[10] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.
[11] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "A combined approach of array processing and independent component analysis for blind separation of acoustic signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 2729–2732.
[12] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, pp. 3140–3143.
[13] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 881–884.
[14] 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, p. 530, Sep. 2004.
[15] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation*, 2000, pp. 215–220.
[16] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomput.*, vol. 41, pp. 1–24, 2001.
[17] S.-I. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," *Adv. Neural Inf. Process. Syst.*, vol. 8, pp. 752–763, 1996.
[18] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation*, 2001, pp. 722–727.
[19] R. B. Stephens and A. E. Bate, *Acoustics and Vibrational Physics*. London, U.K.: E. Arnold, 1966.
[20] 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.

[21] W. G. Gardner, "The Virtual Acoustic Room," Master's thesis, Massachusetts Inst. Technol., Cambridge, 1992.
[22] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1998, pp. 1941–1944.
[23] A. Hyvärinen and P. O. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces," *Neural Comput.*, vol. 12, no. 7, pp. 1705–1720, 2000.
[24] A. Hyvärinen, P. O. Hoyer, and M. Inki, "Topographic independent component analysis," *Neural Comput.*, vol. 13, no. 7, pp. 1527–1558, 2001.
[25] T.-W. Lee, M. S. Lewicki, and T. Sejnowski, "ICA mixture models for unsupervised classification of non-Gaussian classes and automatic context switching in blind separation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1078–1089, Oct. 2000.
[26] T.-W. Lee and M. S. Lewicki, "Unsupervised image classification, segmentation, and enhancement using ICA mixture models," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 270–279, Mar. 2002.
[27] Y. Karklin and M. S. Lewicki, "Learning higher order structures in natural images," *Netw.: Comput. Neural Syst.*, vol. 14, no. 3, pp. 483–499, 2003.
[28] H.-J. Park and T.-W. Lee, "Modeling nonlinear dependencies in natural images using mixture of Laplacian distribution," *Adv. Neural Inf. Process. Syst.*, pp. 1041–1048, 2005.

Taesu Kim (S'05) was born in Korea in 1978. He received the B.S. degree from Hanyang University, Seoul, Korea, in 2001 and the M.S. degree from the Korea Advanced Institute of Science and Technology (KAIST), Dajeon, Korea, in 2003, both in electrical engineering. He is currentlypursuing the Ph.D. degree in the Department of BioSystems, KAIST, and doing his thesis research at the Institute for Neural Computation at the University of California at San Diego, La Jolla.

His research interests include machine learning for signal processing, probabilistic method for unsupervised or reinforcement learning, biologically plausible learning algorithm and their applications. Recently, he has researched on the blind source separation algorithm and probabilistic model for capturing dependencies.

Hagai T. Attias received the Ph.D. degree in statistical physics from Yale University, New Haven, CT.

He is the Chief Scientist at Golden Metallic, Inc. He has authored or coauthored over 50 scientific publications in machine learning, signal processing, and related areas, and has several issued patents. Prior to Golden Metallic, he was a Researcher at the Machine Learning and Applied Statistics Group, Microsoft Research, Redmond, WA. Before that, he was a Sloan Postdoctoral Fellow at the Sloan Center for Theoretical Neurobiology, University of California, San Francisco. In between he had a stint as a Senior Research Fellow at the Gatsby Machine Learning and Computational Neuroscience Unit, University of London.

Soo-Young Lee (M'83) received the B.S. degree from Seoul National University, Seoul, Korea, in 1975, the M.S. degree from the Korea Advanced Institute of Science, Daejon, Korea, in 1977, and the Ph.D. degree from the Polytechnic Institute of New York in 1984.

From 1977 to 1980, he was with the Taihan Engineering Company, Seoul. From 1982 to 1985, he was also with the General Physics Corporation, Columbia, MD. In early 1986, he joined the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), as an Assistant Professor and is now a Full Professor in the Department of BioSystems and also the Department of Electrical Engineering and Computer Science. In 1997, he established the Brain Science Research Center, which is the main research organization for the Korean Brain Neuroinformatics Research Program. The research program is one of the Korean Brain Research Promotion Initiatives sponsored by the Korean Ministry of Science and Technology from 1998 to 2008, and currently about 35 Ph.D. researchers have joined the research program from many Korean universities. His research interests have resided in the artificial brain, the human-like intelligent systems based on biological information processing mechanism in our brain. He has worked on the auditory models from the cochlea to the auditory cortex for noisy speech processing, the unsupervised developmental models of human knowledge with multimodal man–machine interactions, and the top-down selective attention models for superimposed pattern recognitions. Both computational algorithm and VLSI implementation are investigated. Especially, he had developed a System-on-Chip (SoC) for speech recognition based on his auditory model, and a digital chip for active noise canceling and blind signal separation based on independent component analysis.

Dr. Lee is a Past-President of Asia–Pacific Neural Network Assembly, and has contributed to the International Conference on Neural Information Processing as Conference Chair (2000), Conference Vice Co-Chair (2003), and Program Co-Chair (1994, 2002). He is the Editor-in-Chief of the newly established online/offline journal with a double-blind review process, *Neural Information Processing—Letters and Reviews*, and is on the Editorial Board for two international journals, *Neural Processing Letters* and *Neurocomputing*. He received the Leadership Award and Presidential Award from the International Neural Network Society in 1994 and 2001, respectively, and the APPNA Service Award in 2004.

Te-Won Lee (M'03) received the diploma degree and the Ph.D. degree (summa cum laude) in electrical engineering from the University of Technology Berlin in 1995 and 1997, respectively.

He is an Associate Research Professor at the Institute for Neural Computation, University of California at San Diego, La Jolla, and a Collaborating Professor in the Biosystems Department, Korea Advanced Institute of Science and Technology (KAIST), Dajeon, Korea. His research interests include machine learning algorithms with applications in signal and image processing. Recently, he has worked on variational Bayesian methods for independent component analysis, algorithms for speech enhancement and recognition, models for computational vision, and classification algorithms for medical informatics.

Dr. Lee received the Erwin–Stephan prize for excellent studies from the University of Technology Berlin and the Carl-Ramhauser prize for excellent dissertations from the Daimler–Chrysler Corporation. He was a Max-Planck Institute fellow (1995–1997) and a Research Associate at the Salk Institute for Biological Studies (1997–1999).