Available online at www.sciencedirect.com
ScienceDirect
Signal Processing 88 (2008) 1990–2007
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

Blind source separation for convolutive mixtures based on the joint diagonalization of power spectral density matrices ☆

Tiemin Mei[a,c,*,1], Alfred Mertins[a], Fuliang Yin[b], Jiangtao Xi[d], Joe F. Chicharo[d]

[a]*Institute for Signal Processing, University of Lübeck, Ratzeburger Allee 160, Lübeck 23538, Germany*
[b]*School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116023, China*
[c]*School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110168, China*
[d]*School of Electrical, Computer and Telecommunications Engineering, The University of Wollongong, NSW 2522, Australia*

Received 1 March 2007; received in revised form 21 January 2008; accepted 7 February 2008
Available online 16 February 2008

Abstract

This paper studies the problem of blind separation of convolutively mixed source signals on the basis of the joint diagonalization (JD) of power spectral density matrices (PSDMs) observed at the output of the separation system. Firstly, a general framework of JD-based blind source separation (BSS) is reviewed and summarized. Special emphasis is put on the separability conditions of sources and mixing system. Secondly, the JD-based BSS is generalized to the separation of convolutive mixtures. The definition of a time and frequency dependent characteristic matrix of sources allows us to state the conditions under which the separation of convolutive mixtures is possible. Lastly, a frequency-domain approach is proposed for convolutive mixture separation. The proposed approach exploits objective functions based on a set of PSDMs. These objective functions are defined in the frequency domain, but are jointly optimized with respect to the time-domain coefficients of the unmixing system. The local permutation ambiguity problems, which are inherent to most frequency-domain approaches, are effectively avoided with the proposed algorithm. Simulation results show that the proposed algorithm is valid for the separation of both simulated and real-word recorded convolutive mixtures.

Keywords: Blind source separation; Convolutive mixtures; Joint diagonalization; Power spectral density matrices

1. Introduction

Blind source separation (BSS) is to recover a set of unknown source signals from observations that are unknown mixtures of those sources. A challenging situation for BSS is that mixing processes are convolutive, where observations are the combination of the unknown filtered versions of the source

☆This work is supported by the National Natural Science Foundation of China under Grant Nos. 60172073 and 60372082, the Trans-Century Training Program Foundation for the Talents by the Ministry of Education of China, and the German Research Foundation under Grant No. ME 1170/1, it is also partially supported by Australia Research Council under ARC large Grant No. A00103052.
*Corresponding author at: Institute for Signal Processing, University of Lübeck, Ratzeburger Allee 160, Lübeck 23538, Germany. Tel.: +49 451 5005673; fax: +49 451 5005802.
E-mail addresses: mei@isip.uni-luebeck.de, meitiemin@163.com (T. Mei), mertins@isip.uni-luebeck.de (A. Mertins), flyin@dlut.edu.cn (F. Yin), jiangtao@uow.edu.au (J. Xi), chicharo@uow.edu.au (J.F. Chicharo).
[1]On leave from the School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110168, China.

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2008.02.003

signals. The problem has attracted extensive research work in the research communities due to its many potential applications, such as audio processing, image processing, communication systems and biomedical signal processing [1,2].

People have been trying to solve the convolutive BSS problem by two different types of approaches. One is to achieve BSS directly in the time domain, and the other is to work in the frequency domain. In time-domain BSS, a separation network is directly applied to the observed signals to yield separated source signals, and all variables and objective functions are held in the time domain [3–6]. However, these approaches are not very effective for the cases of long mixing channels, such as those in well-known cocktail party problems, where the mixing channels may have 500–2000 taps or more if modelled by FIR filters.

Frequency-domain-based approaches have been considered as the most promising technique for convolutive BSS, especially for cases of long mixing channels. The frequency-domain-based approaches usually consist of three steps. Firstly, the observed signals are decomposed into narrowband components by means of Fourier transforms. Secondly, BSS approaches developed for instantaneous mixture separation are exploited for each of the frequency bins, and finally, separated signals of all the frequency bins are combined together to form the separated outputs. It has been shown that satisfactory separation can be achieved within all the frequency bins, but combining them together to recover the original sources is a challenging issue due to the unknown permutations associated with BSS of individual frequency bins [2,7–13].

Researchers have done extensive work to remedy the permutation problem, and different ways to overcome the problem have emerged by taking advantage of the following information:

- In the frequency domain, the separation filters should satisfy some smoothness constraint, so that two separation matrices at adjacent frequency bins should be similar to each other [7,8,14,15].
- Smoothness of the separation filters in the frequency domain can be implied by limiting the lengths of their time-domain impulse responses [2].
- Smoothness can also be assumed for the mixing-filter frequency responses [16].
- The separated sub-signals at adjacent frequency bins are more related to each other if they stem from the same source [17,18].
- Contributions from the same source are likely to come from the same spatial direction, so that the position information of sources can be exploited [10,19,20].
- Available time–frequency models of sources can be used [9].
- The separation network is defined in time domain, but the parameters are optimized based on a frequency-domain objective function. Such approaches have been exploited by [21–23].

In [21], the mixing system is identified first, and then, knowing the mixing system, the unmixing system is either determined through matrix inversion, or simpler, as the adjoint of the mixing system. The same authors of [21] proposed a two-stage approach based on the same objective, that is, the frequency-domain optimization of mixing matrices and permutation correction by cross-frequency correlations. This undoubtedly makes the algorithm computationally inefficient [18].

The approach in [22] looks at the multiple-input multiple-output (MIMO) deconvolution problem in a setting where colored, stationary processes are assumed. It is stated that the diagonalization of the power spectral density matrix (PSDM) is a sufficient condition for the MIMO deconvolution of colored and stationary processes if the transfer function matrix is of full column rank for all nonzero values of z in the complex plane and it can be factorized into the multiplication of an irreducible matrix, a unitary matrix, and a diagonal matrix with diagonal entries of monic monomials. This is a very strong constraint on the mixing system. Because the deconvolution of MIMO systems is a much more difficult problem than BSS, these conditions may be necessary, but they are not for BSS.

In [23], integration is applied to the frequency-domain defined Kullback–Leibler divergence and leads to BSS of convolutive mixtures where no local permutations take place.

The approach proposed in this paper is based on frequency-domain second-order statistics (SOS) and the nonstationarity of sources, namely, the joint diagonalization (JD) of a set of PSDMs with respect to the time-domain parameters of the separating system. It is different from the one in [21] in that it directly finds the unmixing system based on a group of entirely different objective functions, and it differs from the one in [22] in that it looks at general nonstationary processes instead of looking at colored stationary ones. It is also different from

[23] as it exploits nonstationarity instead of non-Gaussianity. As the separation is performed by a MIMO system in the time domain whose coefficients are the optimizing variables, the proposed approach does not have the local permutation ambiguity disadvantage.

Throughout the paper we use $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\mathrm{H}}$, $(\cdot)^{-\mathrm{T}}$, $(\cdot)^{-\mathrm{H}}$ and $(\cdot)^{*}$ to denote transpose, Hermitian transpose, transpose and inversion, Hermitian transpose and inversion, and conjugate operation, respectively. The operator diag[·] takes a square matrix as argument and yields a square matrix with diagonal elements equal to the diagonal elements of its argument, and off-diagonal elements equal to zero. E[·] is the expectation operator. The operator det[·] yields the determinant of its argument. Boldface letters are used for vectors and matrices, plain letters are used for scalar variables in both time and frequency domains. Especially, italic boldface uppercase letters are used for time-domain matrices. It is easy to identify them by the context in which they are used.

The paper is organized as follows. Firstly, Section 2 summarizes the basic principle of decorrelation-based JD. Then the frequency-domain JD principle for convolutive mixtures is described in Section 3. In Section 4 we propose a new algorithm on the basis of Section 3. Simulation results are presented in Section 5. Finally, Section 6 concludes the paper.

2. The basic principle of JD

In this section we review the JD principle of cross-correlation matrices for the separation of instantaneously mixed sources. These SOS-based BSS ideas, which have been presented by different researchers [6,24–30], are the basis for our new BSS approach for convolutive mixtures.

We consider the N-by-N case, that is, there are N source signals and N observed signals. We assume that the sources are complex valued and are of zero mean. All the other required properties of sources will be given in Theorem 1. The instantaneous mixing system can be described as follows:

$$\mathbf{x}(n) = \boldsymbol{A}\mathbf{s}(n), \tag{1}$$

where $\mathbf{s}(n) = [s_1(n), s_2(n), \ldots, s_N(n)]^{\mathrm{T}}$ are the source signals, $\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_N(n)]^{\mathrm{T}}$ are the observed signals and $\boldsymbol{A}$ is the mixing matrix which is assumed to be nonsingular and time invariant. The task of BSS is to recover the sources from the observations in the form

$$\mathbf{y}(n) = \boldsymbol{W}\mathbf{x}(n), \tag{2}$$

where $\mathbf{y}(n) = [y_1(n), y_2(n), \ldots, y_N(n)]^{\mathrm{T}}$ is the output of the separation system, and $\boldsymbol{W}$ is the matrix describing the separation network. Combining (1) and (2) gives

$$\mathbf{y}(n) = \boldsymbol{G}\mathbf{s}(n), \tag{3}$$

where $\boldsymbol{G} = [g_{ij}] = \boldsymbol{W}\boldsymbol{A}$, which is the transform matrix from $\mathbf{s}(n)$ to $\mathbf{y}(n)$. Separation is considered to be successful if we can find a matrix $\boldsymbol{W}$ such that $\boldsymbol{G}$ is the product of a diagonal matrix $\boldsymbol{D}$ and a permutation matrix $\boldsymbol{P}$.

Now let us show that the JD of N output cross-correlation matrices leads to the separation of source signals. For this, we follow [30], but consider complex instead of real-valued signals. Firstly we define the characteristic matrix of sources as follows:

$$\boldsymbol{R}_{\mathrm{c}} = \begin{bmatrix} r_{s_1s_1}(n_1, n_1') & r_{s_2s_2}(n_1, n_1') & \cdots & r_{s_Ns_N}(n_1, n_1') \\ r_{s_1s_1}(n_2, n_2') & r_{s_2s_2}(n_2, n_2') & \cdots & r_{s_Ns_N}(n_2, n_2') \\ \vdots & \vdots & \ddots & \vdots \\ r_{s_1s_1}(n_N, n_N') & r_{s_2s_2}(n_N, n_N') & & r_{s_Ns_N}(n_N, n_N') \end{bmatrix}, \tag{4}$$

where $r_{s_js_j}(n_i, n_i') = \mathrm{E}[s_j(n_i)s_j^*(n_i')]$. Note that this matrix potentially includes correlation terms for different time lags (e.g., $n_i' - n_i \neq n_k' - n_k$ for $n_i = n_k$) as well as for different epochs of a nonstationary process (i.e., $n_i' = n_i$ but $n_i \neq n_k$ for different i, k). We have the following theorem.

Theorem 1 (*Yin et al. [30]*). *For N zero mean and complex-valued nonstationary stochastic processes $\{s_1(n), \ldots, s_N(n)\}$, if there exist N different points (n_k, n_k') $(k = 1, 2, \ldots, N)$ in* 2-D *space (n, n'), such that the following cross-correlation matrices of source signals are diagonal, that is*

$$\boldsymbol{R}_{\mathbf{ss}}(n_k, n_k') = \mathrm{diag}[\boldsymbol{R}_{\mathbf{ss}}(n_k, n_k')], \quad k = 1, 2, \ldots, N, \tag{5}$$

where $\boldsymbol{R}_{\mathbf{ss}}(n_k, n_k') = \mathrm{E}[\mathbf{s}(n_k)\mathbf{s}^{\mathrm{H}}(n_k')] = [r_{s_is_j}(n_k, n_k')]$, *and, in addition, the characteristic matrix of sources $\boldsymbol{R}_{\mathrm{c}}$ is of full rank, namely*

$$\det[\boldsymbol{R}_{\mathrm{c}}] \neq 0 \tag{6}$$

then the nonsingularly mixed sources can be separated by Eq. (2) *when the following cross-correlation matrices are simultaneously diagonalized through the*

proper choice of an unmixing matrix $\boldsymbol{W}$:

$$\boldsymbol{R}_{\mathbf{yy}}(n_k, n'_k) = \mathrm{diag}[\boldsymbol{R}_{\mathbf{yy}}(n_k, n'_k)], \quad k = 1, 2, \ldots, N, \tag{7}$$

where $\boldsymbol{R}_{\mathbf{yy}}(n_k, n'_k) = \mathrm{E}[\mathbf{y}(n_k)\mathbf{y}^{\mathrm{H}}(n'_k)] = [r_{y_i y_j}(n_k, n'_k)]$.

The conditions in Theorem 1 are typically satisfied when the sources are statistically independent and the times $n_1, n'_1, \ldots, n_N, n'_N$ have not been very unluckily chosen. However, interestingly, the conditions in Theorem 1 may also be satisfied by partly dependent sources, which indicates that independence of sources is usually a sufficient, but not a necessary condition for BSS through JD of observed PSDMs. Regarding BSS, (5) and (6) are the separability conditions of sources, and (7) is the separation criterion.

3. The JD principle for convolutive mixtures

In this section, we extend the SOS-based JD principle presented in Section 2 to the case of convolutive mixtures.

For stochastic vector processes (in our case, the outputs of the separation system), the power spectral density matrix can be explained as the sub-band correlation matrix of the stochastic vector process, this is why the JD principle can be applied to convolutive mixture separation.

3.1. Convolutive BSS models and assumptions

We still consider the N-by-N case, that is, there are N source signals, N observation signals and N separated signals as well. The mixing channels are assumed to be FIR of length L, and the separation channels are also FIR and their length (M) is chosen so that $M \geqslant (N-1)(L-1)+1$ in order to achieve satisfying performance [21]. Also we have the following assumptions regarding the sources and the mixing processes:

(A1) Source signals $\mathbf{s}(n) = [s_1(n), s_2(n), \ldots, s_N(n)]^{\mathrm{T}}$ are real-valued, zero mean and uncorrelated to each other.

(A2) The source signals $s_k(n)$ are nonstationary, which means that auto-power spectral densities determined within local time intervals as

$$p_{s_k s_k}(n, \omega) = \mathrm{E}[s_k(n, \mathrm{e}^{\mathrm{j}\omega}) s_k^*(n, \mathrm{e}^{\mathrm{j}\omega})],$$

where

$$s_k(n, \mathrm{e}^{\mathrm{j}\omega}) = \sum_m s_k(m) w(m-n) \mathrm{e}^{-\mathrm{j}\omega m}$$

with $w(n)$ being a window function are time-varying in nature. The term $s_k(n, \mathrm{e}^{\mathrm{j}\omega})$ is known as the short-time Fourier transform (STFT) of $s_k(n)$ [31]. In this context, the window $w(n)$ is only assumed to be real-valued, of finite energy, and of finite length. The choice of a particular window $w(n)$ will be further discussed in Section 3.2 and experimentally investigated in Section 5.6.

(A3) The mixing system $\boldsymbol{A}(n) = [a_{ij}(n)]_{N\times N}$ is linear and time invariant (LTI), where $a_{ij}(n)$ is the impulse response of the channel from source $s_j(n)$ to observation $x_i(n)$.

(A4) The transfer matrix of the mixing system

$$\mathbf{A}(z) = \sum_{n=0}^{L-1} \boldsymbol{A}(n) z^{-n}$$

is nonsingular on the unit circle in the complex plane, i.e., $\det[\mathbf{A}(\mathrm{e}^{\mathrm{j}\omega})] \neq 0$.

Assumption (A1) is necessary for decorrelation-based BSS; Assumption (A3) is a basic condition; Assumptions (A2) and (A4) are necessary for the separation of sub-signals at a given frequency.

In practice, noises are always there in observations, but for reasons of conciseness, they are not taken into account in this paper. Noises can be dealt with by using special denoising methods or by adjusting the power spectra of the observations by the estimated amount of noise. The noise-free convolutive mixing model is given as follows:

$$\mathbf{x}(n) = \boldsymbol{A}(n) * \mathbf{s}(n) = \sum_{l=0}^{L-1} \boldsymbol{A}(l)\mathbf{s}(n-l), \tag{8}$$

where $*$ denotes the convolution operation, $\mathbf{s}(n)$ is the source signal vector, $\mathbf{x}(n)$ is the mixture vector, $\boldsymbol{A}(n) = [a_{ij}(n)]_{N\times N}$ is the mixing matrix, and $a_{ij}(n)$ denotes the impulse response of the FIR channel from $s_j(n)$ to mixture $x_i(n)$.

The separation-system output $\mathbf{y}(n) = [y_1(n), y_2(n), \ldots, y_N(n)]^{\mathrm{T}}$ is given as follows:

$$\mathbf{y}(n) = \boldsymbol{H}(n) * \mathbf{x}(n) = \sum_{l=0}^{M-1} \boldsymbol{H}(l)\mathbf{x}(n-l), \tag{9}$$

where $\boldsymbol{H}(n) = [h_{ij}(n)]_{N\times N}$ is the separation matrix and $h_{ij}(n)$ denotes the impulse response of the FIR channel from $x_j(n)$ to output $y_i(n)$. From (8) and (9),

we have

$$\mathbf{y}(n) = \boldsymbol{H}(n) * \boldsymbol{A}(n) * \mathbf{s}(n) = \boldsymbol{G}(n) * \mathbf{s}(n), \tag{10}$$

where $\boldsymbol{G}(n) = \boldsymbol{H}(n) * \boldsymbol{A}(n)$. Equivalently in the z-domain we have

$$\mathbf{Y}(z) = \mathbf{G}(z)\mathbf{S}(z). \tag{11}$$

BSS is considered to be successful if the output $\mathbf{y}(n)$ is at most a permuted and filtered version of the source signals $\mathbf{s}(n)$, in which case $\mathbf{G}(z)$ is a product of a permutation matrix $\mathbf{P}$ and a diagonal matrix $\mathbf{D}(z)$:

$$\mathbf{G}(z) = \mathbf{P}\mathbf{D}(z). \tag{12}$$

3.2. JD principle for convolutive mixtures

As source signals are nonstationary and the mixing system is LTI, we use the STFT to describe the mixing process (8) and the separating process (9) in the time–frequency domain. For this, a vector of source STFTs is defined as

$$\mathbf{S}(n, \mathrm{e}^{\mathrm{j}\omega}) = [s_1(n, \mathrm{e}^{\mathrm{j}\omega}), s_2(n, \mathrm{e}^{\mathrm{j}\omega}), \ldots, s_N(n, \mathrm{e}^{\mathrm{j}\omega})]^{\mathrm{T}},$$

where n is the time index which describes the short-time signal spectra in different time windows. For the window $w(n)$ used to generate the STFTs $s_k(n, \mathrm{e}^{\mathrm{j}\omega})$, we first choose a simple rectangular window. If the original source signals $\mathbf{s}(n)$ are nonstationary, then the source sub-signals $\mathbf{S}(n, \mathrm{e}^{\mathrm{j}\omega})$ are also nonstationary.

The multiplication of $\mathbf{S}(n, \mathrm{e}^{\mathrm{j}\omega})$ with

$$\mathbf{A}(\mathrm{e}^{\mathrm{j}\omega}) = \sum_{n=0}^{L-1} \boldsymbol{A}(n)\mathrm{e}^{-\mathrm{j}\omega n}$$

yields

$$\tilde{\mathbf{X}}(n, \mathrm{e}^{\mathrm{j}\omega}) = \mathbf{A}(\mathrm{e}^{\mathrm{j}\omega})\mathbf{S}(n, \mathrm{e}^{\mathrm{j}\omega}). \tag{13}$$

When taking the inverse Fourier transforms of $\tilde{\mathbf{X}}(n, \mathrm{e}^{\mathrm{j}\omega})$ and adding them up with the corresponding overlaps, one obtains the exact result of the linear convolution in (8). For the case where the Fourier transform is computed for a discrete set of frequencies via the FFT, this is known as the overlap-and-add method of fast convolution [32].

On the other hand, when we compute the STFTs of the mixed signals as

$$\mathbf{X}(n, \mathrm{e}^{\mathrm{j}\omega}) = [x_1(n, \mathrm{e}^{\mathrm{j}\omega}), x_2(n, \mathrm{e}^{\mathrm{j}\omega}), \ldots, x_N(n, \mathrm{e}^{\mathrm{j}\omega})]^{\mathrm{T}}$$

with

$$x_k(n, \mathrm{e}^{\mathrm{j}\omega}) = \sum_m x_k(m)w(m-n)\mathrm{e}^{-\mathrm{j}\omega m},$$

where $w(n)$ is the same rectangular window as before, we observe a small difference between $\tilde{\mathbf{X}}(n, \mathrm{e}^{\mathrm{j}\omega})$ and $\mathbf{X}(n, \mathrm{e}^{\mathrm{j}\omega})$ due to boundary effects. Like in the overlap-and-add method of fast convolution, the inverse Fourier transforms of $\tilde{\mathbf{X}}(n, \mathrm{e}^{\mathrm{j}\omega})$ and $\mathbf{X}(n, \mathrm{e}^{\mathrm{j}\omega})$ will differ at both ends while being exactly equal in the center part. The longer the window the longer the part containing an exact match will be. The number of the samples affected by boundary effects equals the filter length. Thus, when assuming a sufficiently long window $w(n)$ and fast decaying room responses, we have the relationship

$$\mathbf{X}(n, \mathrm{e}^{\mathrm{j}\omega}) \approx \tilde{\mathbf{X}}(n, \mathrm{e}^{\mathrm{j}\omega}) \tag{14}$$

with very good approximation. The unmixing process can be written as

$$\mathbf{Y}(n, \mathrm{e}^{\mathrm{j}\omega}) = \mathbf{H}(\mathrm{e}^{\mathrm{j}\omega})\mathbf{X}(n, \mathrm{e}^{\mathrm{j}\omega}), \tag{15}$$

where relationship (15) is exact, and the filtered outputs $\mathbf{y}(n)$ can be computed from $\mathbf{Y}(n, \mathrm{e}^{\mathrm{j}\omega})$ without error.

Other choices for the window $w(n)$ than the rectangular one, such as Hamming, Hann or Gaussian windows, may yield better time–frequency resolution of the STFTs of the mixtures, however, they typically result in greater approximation errors between $\mathbf{X}(n, \mathrm{e}^{\mathrm{j}\omega})$ and $\tilde{\mathbf{X}}(n, \mathrm{e}^{\mathrm{j}\omega})$. The Tukey window, which contains a constant part in the center and a soft roll off at the ends, is a compromise between approximation properties and time–frequency resolution. In the experimental part, we will present results for different window choices.

For a given time instant n, the instant PSDM of the separation-system output $\mathbf{Y}(n, \mathrm{e}^{\mathrm{j}\omega})$ in (15) can be defined as

$$\mathbf{P}_{\mathbf{YY}}(n, \omega) = \mathrm{E}[\mathbf{Y}(n, \mathrm{e}^{\mathrm{j}\omega})\mathbf{Y}^{\mathrm{H}}(n, \mathrm{e}^{\mathrm{j}\omega})]. \tag{16}$$

For a given frequency ω, this is nothing but the instant correlation matrix of the nonstationary sub-signals $y_k(n, \mathrm{e}^{\mathrm{j}\omega})$ contained in $\mathbf{Y}(n, \mathrm{e}^{\mathrm{j}\omega})$.

In practice, the STFT will be evaluated for a discrete set of frequencies $\omega_k = 2\pi k/K$, and the implementation will be based on the FFT. If the mixing and unmixing systems are FIR and the conditions known for fast convolution algorithms [32] are met, the operations still yield the exact result of linear convolution. If, however, the FFT length is too short in relation to the mixing-filter and analysis-window lengths, the originally linear convolution is replaced by circular convolution (cf. [7]). In the following, we will write ω as a continuous

variable, although discrete implementations will be based on the FFT and a discrete set of frequencies.

For a given frequency ω, the vector $\tilde{\mathbf{X}}(n,e^{j\omega})$ in (13) can be considered as a mixture of complex sources $\mathbf{S}(n,e^{j\omega})$ which have been mixed through an instantaneous complex mixing matrix $\mathbf{A}(e^{j\omega})$. Similarly, $\mathbf{Y}(n,e^{j\omega})$ in (15) can be considered as the output of a separation system modelled by an instantaneous separating matrix $\mathbf{H}(e^{j\omega})$ whose input is $\mathbf{X}(n,e^{j\omega})$. According to assumptions (A1)–(A4) and Theorem 1, and if the approximation (14) is sufficiently close, we can separate the mixtures at frequency ω based on the JD principle for nonstationary sources.

Now we extend the JD principle to the STFT of the outputs of the separation system in the time–frequency domain. Firstly, we define the time–frequency domain characteristic matrix of sources as follows:

$$\mathbf{P}_c(n_1,n_2,\ldots,n_N,\omega)=[p_{s_is_l}(n_i,\omega)]_{N\times N}, \tag{17}$$

where $p_{s_is_l}(n_i,\omega)=\mathrm{E}[s_l(n_i,e^{j\omega})s_l^*(n_i,e^{j\omega})]$ is the ith row and lth column entry of $\mathbf{P}_c(n_1,n_2,\ldots,n_N,\omega)$.

The characteristic matrix $\mathbf{P}_c(n_1,n_2,\ldots,n_N,\omega)$ is composed of the instant power spectral density functions of sources in different epochs, so it shows the time dependance of the power spectra.

According to assumption (A1), the cross-correlation matrices of sources, $\boldsymbol{R}_{\mathbf{ss}}(l,m)=\mathrm{E}[\mathbf{s}(l)\mathbf{s}^{\mathrm{T}}(l-m)]$, are diagonal, and equivalently, the PSDMs are diagonal, that is,

$$\mathbf{P}_{\mathbf{SS}}(n,\omega)=[p_{s_is_j}(n,\omega)]=\mathrm{diag}[\mathbf{P}_{\mathbf{SS}}(n,\omega)], \tag{18}$$

where $p_{s_is_j}(n,\omega)=\mathrm{E}[s_i(n,e^{j\omega})s_j^*(n,e^{j\omega})]$ is the instant cross-power spectral density of sources $s_i(n)$ and $s_j(n)$. The term $p_{s_is_j}(n,\omega)$ can also be seen as the cross correlation of source sub-signals $s_i(n,\omega)$ and $s_j(n,\omega)$ for a given ω.

According to assumption (A2), for a given frequency ω, we should be able to find N different time instants $n_1,n_2,\ldots,n_N$ such that the characteristic matrix $\mathbf{P}_c(n_1,n_2,\ldots,n_N,\omega)$ is of full rank. According to Theorem 1 in Section 2, it is certain that the JD of $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ $(l=n_1,n_2,\ldots,n_N)$ through the choice of $\mathbf{H}(e^{j\omega})$ will lead to the separation of mixtures at the frequency ω. This can be extended to all frequencies, and thus we have the following conclusion: For convolutive BSS models defined above, if there exist N different time instants $n_1,n_2,\ldots,n_N$ such that $\mathbf{P}_c(n_1,n_2,\ldots,n_N,\omega)$ is of full rank for all ω, the convolutive mixtures can be separated if the PSDMs $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ $(l=n_1,n_2,\ldots,n_N)$ are jointly diagonalized for all ω, that is

$$\mathbf{P}_{\mathbf{YY}}(l,\omega)=[p_{y_iy_j}(l,\omega)]=\mathrm{diag}[\mathbf{P}_{\mathbf{YY}}(l,\omega)] \tag{19}$$

for all $l=n_1,n_2,\ldots,n_N$ and ω.

BSS approaches can be developed based on the JD principle (19). A straightforward way is to jointly diagonalize $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ $(l=n_1,n_2,\ldots,n_N$, for all $\omega)$ with respect to $\mathbf{H}(e^{j\omega})$ in the frequency domain [2,7]. However, this will inevitably result in the permutation ambiguity problems.

In order to avoid the permutation ambiguity problem, we aim to find an alternative approach that also jointly diagonalizes $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ $(l=n_1,n_2,\ldots,n_N$, for all $\omega)$ but with respect to other parameters, namely the time-domain filter taps of the unmixing system. First of all, we will introduce a nonnegative function $f_\omega(\mathbf{P}_{\mathbf{YY}}(l,\omega))$, called the diagonalization index function (DIF), to measure how different $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ is from a diagonal matrix. $f_\omega(\mathbf{P}_{\mathbf{YY}}(l,\omega))$ is constructed in such a way that (a) $f_\omega(\mathbf{P}_{\mathbf{YY}}(l,\omega))$ $(\geqslant 0)$ reaches its minimum when $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ is a diagonal matrix, (b) $f_\omega(\mathbf{P}_{\mathbf{YY}}(l,\omega))$ is greater than its minimum if $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ is not a diagonal matrix, and (c) $f_\omega(\mathbf{P}_{\mathbf{YY}}(l,\omega))$ increases as $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ deviates from a diagonal matrix, and it decreases as $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ approaches a diagonal matrix.

The above DIF of $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ is defined with respect to a single frequency ω only. In order to achieve BSS, we would make $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ diagonal for all frequencies. Given that $f_\omega(\mathbf{P}_{\mathbf{YY}}(l,\omega))$ is nonnegative in nature, we can use the integrated diagonalization index function (IDIF) defined as follows:

$$\begin{aligned}\mathrm{IDIF}&=f(l,\boldsymbol{H}(n)|_{n=0,1,2,\ldots,M-1})\\&=\int_{-\pi}^{\pi}f_\omega(\mathbf{P}_{\mathbf{YY}}(l,\omega))\,\mathrm{d}\omega.\end{aligned} \tag{20}$$

To see how the IDIF results in a function of the time-domain coefficients of the separation system and the time index l, one has to consider the relationship

$$\begin{aligned}\mathbf{P}_{\mathbf{YY}}(l,\omega)&=\mathbf{H}(e^{j\omega})\mathbf{P}_{\mathbf{XX}}(l,\omega)\mathbf{H}^{\mathrm{H}}(e^{j\omega})\\&=\left(\sum_{k=0}^{M-1}H(k)e^{-j\omega k}\right)\mathbf{P}_{\mathbf{XX}}(l,\omega)\\&\quad\times\left(\sum_{k=0}^{M-1}H^{\mathrm{T}}(k)e^{j\omega k}\right).\end{aligned} \tag{21}$$

After integration over ω, the remaining variables are $\boldsymbol{H}(k)$, $k=0,1,\ldots,M-1$ and l. The IDIFs for $l=n_1,n_2,\ldots,n_N$ are then optimized jointly.

The above IDIF measures the difference of the PSDMs for all frequencies to diagonal ones, which also has the properties including (a) IDIF ($\geqslant 0$) reaches its minimum if $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ for all ω are diagonal matrices, (b) IDIF is greater than its minimum if $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ for all ω are not diagonal matrices, and (c) the IDIF increases as $\mathbf{P}_{\mathbf{YY}}(l,\omega)$, for all ω as a whole, deviate from diagonal matrices, and it decreases as $\mathbf{P}_{\mathbf{YY}}(l,\omega)$, for all ω as a whole, approaches diagonal matrices. It is obvious that the IDIF defined in (20) can be used as an objective function for optimizing the BSS system.

Based on the above work, the following conclusion regarding the BSS of convolutive mixtures is achieved:

Theorem 2. *For N source signals and a mixing system which satisfies assumptions* (A1)–(A4), *if there exist at least N time instants* $n_1,n_2,\ldots,n_N$ *that make the time–frequency domain characteristic matrix* (17) *be full rank for all* ω, *then the joint minimization of the N IDIFs defined in* (20) ($l=n_1,n_2,\ldots,n_N$), *in the sense that they reach their minima, will result in the separation of the N sources from N convolutively mixed observations.*

Similar to the instantaneous case, assumptions (A1)–(A4) and the time–frequency domain characteristic matrix (17) are the separability conditions of sources and mixing system, the joint minimization of the IDIFs in (20) is the separation criterion for convolutive mixtures.

Note that there are a few points that should be emphasized from Theorem 2. Firstly, source signals must be nonstationary, whether the sources are colored or not does not matter. Secondly, the joint optimization of IDIFs is with respect to the time-domain parameters of the separation system rather than the frequency-domain parameters, this implies that the length of $\boldsymbol{H}(n)$ is predetermined and it sets a smoothness constraint on the corresponding frequency-domain parameters $\mathbf{H}(e^{j\omega})$, just like that in [21–23], so the permutation issue can be avoided effectively. Thirdly, the number of the IDIFs involved in the joint optimization is equal to or greater than the number of sources. In the next section, we will develop our BSS algorithm.

Proof of Theorem 2. Firstly, as the instantaneous mixing model of sub-signals described in (13) concerned, from assumption (A1), the instant PSDMs (they are also the correlation matrices of sub-signals) of sources are diagonal, this means that (18) holds, in other words, sub-signals $s_i(n,e^{j\omega})$ ($i=1,2,\ldots,N$) at frequency ω are uncorrelated to each other. In addition, assumption (A4) points out that matrix $\mathbf{A}(e^{j\omega})$ in (9) is nonsingular, therefore, according to Theorem 1, if there are N time instants (this holds because of the nonstationarity assumption (A2)) which make the characteristic matrix (17) be full column rank for a given frequency ω, then the sub-signals at frequency ω can be separated through the JD of the N PSDMs defined in (16). Secondly, in fact, separating all the sub-signals at different frequencies separately does not automatically lead to the separation of convolutively mixed sources because of the local permutations at different frequencies. This problem is overcome by the definition and the time-domain joint optimization of IDIFs, because the time-domain optimization is essentially that a length constraint has been set to the unmixing filters, or equally, a smoothness constraint, which forbids random local permutation, is set to the unmixing filters in frequency domain. It is proved in Appendix A that if the unmixing filter length M satisfies $M<((N-1)/N^2)K$, where K is the block size of the FFT and N is the number of sources, then there will be no permutation.

4. The BSS algorithm

Theorem 2 given in Section 3 provides a guideline for achieving BSS. However, the explicit form of the DIF and hence IDIF is not given although its properties are defined. In this section we will propose a DIF based on which a BSS algorithm can be developed.

4.1. Selection of the IDIF function

As mentioned above, we should choose a non-negative DIF function that measures the deviation of the PSDMs from diagonal ones. Such a measure is the well-known Hadamard inequality [33], which says that, for a Hermitian and positive definite matrix, the absolute value of the product of its diagonal elements is equal or greater than the absolute value of its determinant, and that the equality holds if and only if the matrix is diagonal. As $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ is Hermitian and positive definite, based on Hadamard's inequality we have

$\det[\mathbf{D_{YY}}(l,\omega)] \geqslant \det[\mathbf{P_{YY}}(l,\omega)]$ where $\mathbf{D_{YY}}(l,\omega) = \mathrm{diag}[\mathbf{P_{YY}}(l,\omega)]$ and $l = n_1, n_2, \ldots, n_N$. Hence we can define the IDIF as follows:

$$f(l, \boldsymbol{H}(n)|_{n=0,1,\ldots,M-1}) = \begin{cases} \dfrac{1}{2}\displaystyle\int_{-\pi}^{\pi} \dfrac{1}{\alpha}\left(\dfrac{\det[\mathbf{D_{YY}}(l,\omega)]}{\det[\mathbf{P_{YY}}(l,\omega)]}\right)^{\alpha} \mathrm{d}\omega & (\alpha \neq 0) \\ \dfrac{1}{2}\displaystyle\int_{-\pi}^{\pi} \log\left(\dfrac{\det[\mathbf{D_{YY}}(l,\omega)]}{\det[\mathbf{P_{YY}}(l,\omega)]}\right) \mathrm{d}\omega & (\alpha = 0) \end{cases} \tag{22}$$

for $l = n_1, n_2, \ldots, n_N$. No matter what value of α is set, the IDIFs always have a minimum at the separation point. For a given α, the joint minimization of IDIFs defined in (22) will lead to the separation of convolutive mixtures. Different values for parameter α will lead to different properties of the algorithm, as the relative contribution of large and small ratios $\det[\mathbf{D_{YY}}(l,\omega)]/\det[\mathbf{P_{YY}}(l,\omega)]$ to the integral changes in accordance with α.

4.2. The BSS algorithm

Now let us derive the BSS algorithm based on the IDIF in (22). The idea is to use the IDIFs as objective functions for a joint optimization with respect to the separating filters' coefficients. For $\alpha \neq 0$, starting with (22) and using the derivation:

$$\begin{aligned} &\frac{\partial\left(\frac{1}{\alpha}\left(\frac{\det[\mathbf{D_{YY}}(l,\omega)]}{\det[\mathbf{P_{YY}}(l,\omega)]}\right)^{\alpha}\right)}{\partial \boldsymbol{H}(n)} \\ &= \left(\frac{\det[\mathbf{D_{YY}}(l,\omega)]}{\det[\mathbf{P_{YY}}(l,\omega)]}\right)^{\alpha-1} \frac{\partial\left(\frac{\det[\mathbf{D_{YY}}(l,\omega)]}{\det[\mathbf{P_{YY}}(l,\omega)]}\right)}{\partial \boldsymbol{H}(n)} \\ &= \left(\frac{\det[\mathbf{D_{YY}}(l,\omega)]}{\det[\mathbf{P_{YY}}(l,\omega)]}\right)^{\alpha} \\ &\quad \times \left(\frac{\frac{\partial \det[\mathbf{D_{YY}}(l,\omega)]}{\partial \boldsymbol{H}(n)}}{\det[\mathbf{D_{YY}}(l,\omega)]} - \frac{\frac{\partial \det[\mathbf{P_{YY}}(l,\omega)]}{\partial \boldsymbol{H}(n)}}{\det[\mathbf{P_{YY}}(l,\omega)]}\right) \\ &= \left(\frac{\det[\mathbf{D_{YY}}(l,\omega)]}{\det[\mathbf{P_{YY}}(l,\omega)]}\right)^{\alpha} \frac{\partial\left(\log\left(\frac{\det[\mathbf{D_{YY}}(l,\omega)]}{\det[\mathbf{P_{YY}}(l,\omega)]}\right)\right)}{\partial \boldsymbol{H}(n)} \end{aligned}$$

we obtain

$$\begin{aligned} &\frac{\partial f(l, \boldsymbol{H}(n)|_{n=0,1,\ldots,M-1})}{\partial \boldsymbol{H}(n)} \\ &= \frac{1}{2}\int_{-\pi}^{\pi} \left(\frac{\det[\mathbf{D_{YY}}(l,\omega)]}{\det[\mathbf{P_{YY}}(l,\omega)]}\right)^{\alpha} \\ &\quad \times \frac{\partial\left(\log\left(\frac{\det[\mathbf{D_{YY}}(l,\omega)]}{\det[\mathbf{P_{YY}}(l,\omega)]}\right)\right)}{\partial \boldsymbol{H}(n)} \mathrm{d}\omega. \end{aligned} \tag{23}$$

First of all, using relationship (21) between $\mathbf{P_{YY}}(l,\omega)$ and $\mathbf{P_{XX}}(l,\omega)$, we have

$$\begin{aligned} &\log(\det[\mathbf{D_{YY}}(l,\omega)]) \\ &= \sum_{i=1}^{N} \log p_{y_i y_i}(l,\omega) \\ &= \sum_{i=1}^{N} \log\left(\sum_{m=1}^{N}\sum_{k=1}^{N} H_{im}(\mathrm{e}^{\mathrm{j}\omega}) H_{ik}(\mathrm{e}^{-\mathrm{j}\omega}) p_{x_m x_k}(l,\omega)\right) \end{aligned} \tag{24}$$

and

$$\begin{aligned} \log(\det[\mathbf{P_{YY}}(l,\omega)]) = {} & \log(\det[\mathbf{H}(\mathrm{e}^{\mathrm{j}\omega})]) \\ & + \log(\det[\mathbf{H}^{\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})]) \\ & + \log(\det[\mathbf{P_{XX}}(l,\omega)]), \end{aligned} \tag{25}$$

where

$$H_{im}(\mathrm{e}^{\mathrm{j}\omega}) = \sum_{n=0}^{M-1} h_{im}(n)\mathrm{e}^{-\mathrm{j}\omega n}. \tag{26}$$

Now we calculate the gradient of the IDIF with respect to separating filters' coefficients. Firstly,

$$\begin{aligned} &\frac{\partial(\log(\det[\mathbf{D_{YY}}(l,\omega)]))}{\partial \boldsymbol{H}(n)} \\ &= \mathbf{D_{YY}^{-1}}(l,\omega)[\mathbf{P_{YY}^{*}}(l,\omega)\mathbf{H}^{-\mathrm{T}}(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{-\mathrm{j}\omega n} \\ &\quad + \mathbf{P_{YY}}(l,\omega)\mathbf{H}^{-\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{\mathrm{j}\omega n}]. \end{aligned} \tag{27}$$

Secondly,

$$\begin{aligned} &\frac{\partial(\log(\det[\mathbf{P_{YY}}(l,\omega)]))}{\partial \boldsymbol{H}(n)} \\ &= \mathbf{H}^{-\mathrm{T}}(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{-\mathrm{j}\omega n} + \mathbf{H}^{-\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{\mathrm{j}\omega n}. \end{aligned} \tag{28}$$

Hence the gradient of the IDIF is calculated as follows:

$$\begin{aligned} &\frac{\partial f(l, \boldsymbol{H}(n)|_{n=0,1,\ldots,M-1})}{\partial \boldsymbol{H}(n)} \\ &= \int_{-\pi}^{\pi} \left\{ \left(\frac{\det[\mathbf{D_{YY}}(l,\omega)]}{\det[\mathbf{P_{YY}}(l,\omega)]}\right)^{\alpha} \right. \\ &\quad \left. \times [\mathbf{D_{YY}^{-1}}(l,\omega)\mathbf{P_{YY}}(l,\omega) - \mathbf{I}]\mathbf{H}^{-\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{\mathrm{j}\omega n} \right\} \mathrm{d}\omega, \end{aligned} \tag{29}$$

where $\mathbf{I}$ is the identity matrix.

A gradient-based algorithm can be obtained using the gradient in (29). We obtain

$$\boldsymbol{H}^{t+1}(n)=\boldsymbol{H}^{t}(n)-\mu\int_{-\pi}^{\pi}\left\{\left(\frac{\det[\mathbf{D}_{\mathbf{YY}}(l,\omega)]}{\det[\mathbf{P}_{\mathbf{YY}}(l,\omega)]}\right)^{\alpha}\times[\mathbf{D}_{\mathbf{YY}}^{-1}(l,\omega)\mathbf{P}_{\mathbf{YY}}(l,\omega)-\mathbf{I}](\mathbf{H}^{t})^{-\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{\mathrm{j}\omega n}\right\}\mathrm{d}\omega, \tag{30}$$

where t is the iteration index for updating the separating filter's coefficients, and $\mu>0$ is the updating step size.

It is shown that, for the case of instantaneous mixture separation, the natural gradient [34] improves the separation performance in terms of convergence properties and enhances the computational efficiency when compared to the normal gradient algorithm. We use the generalized version of the natural gradient from the instantaneous to the convolutive case given in [35]. The natural gradient is as follows:

$$\left.\frac{\partial f(l,\boldsymbol{H}(n)|_{n=0,1,\ldots,M-1})}{\partial\boldsymbol{H}(n)}\right|_{\text{Natural}}=\int_{-\pi}^{\pi}\left\{\left(\frac{\det[\mathbf{D}_{\mathbf{YY}}(l,\omega)]}{\det[\mathbf{P}_{\mathbf{YY}}(l,\omega)]}\right)^{\alpha}\times[\mathbf{D}_{\mathbf{YY}}^{-1}(l,\omega)\mathbf{P}_{\mathbf{YY}}(l,\omega)-\mathbf{I}]\mathbf{H}(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{\mathrm{j}\omega n}\right\}\mathrm{d}\omega, \tag{31}$$

where we have used

$$\sum_{n=-(M-1)}^{0}\boldsymbol{H}^{\mathrm{T}}(-n)\mathrm{e}^{-\mathrm{j}\omega n}=\mathbf{H}^{\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega}).$$

The natural-gradient-based algorithm for updating the separating channel coefficients can then be stated as follows:

$$\boldsymbol{H}^{t+1}(n)=\boldsymbol{H}^{t}(n)-\mu\int_{-\pi}^{\pi}\left\{\left(\frac{\det[\mathbf{D}_{\mathbf{YY}}(l,\omega)]}{\det[\mathbf{P}_{\mathbf{YY}}(l,\omega)]}\right)^{\alpha}\times[\mathbf{D}_{\mathbf{YY}}^{-1}(l,\omega)\mathbf{P}_{\mathbf{YY}}(l,\omega)-\mathbf{I}](\mathbf{H}^{t})(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{\mathrm{j}\omega n}\right\}\mathrm{d}\omega. \tag{32}$$

For the case in which $\alpha=0$, it is easy to check that the gradient of the objective function is just what we obtain if α is set to zero in (29). So if we set $\alpha=0$, then we get the simplified version of (32) as follows:

$$\boldsymbol{H}^{t+1}(n)=\boldsymbol{H}^{t}(n)-\mu\int_{-\pi}^{\pi}[\mathbf{D}_{\mathbf{YY}}^{-1}(l,\omega)\mathbf{P}_{\mathbf{YY}}(l,\omega)-\mathbf{I}]\times\mathbf{H}^{t}(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{\mathrm{j}\omega n}\,\mathrm{d}\omega. \tag{33}$$

Table 1
Off-line implementation

Step 1: The observed signal samples are segmented into $K_m(\geqslant N)$ blocks which can be either overlapping or nonoverlapping;
Step 2: For each block, $\mathbf{P}_{\mathbf{XX}}(l,\omega)$ is obtained based on the observed data;
Step 3: Calculate $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ using (21);
Step 4: Update $\boldsymbol{H}(n)$ and $\mathbf{H}(\mathrm{e}^{\mathrm{j}\omega})$ using (33);
Step 5: For the next block, go to Step 3. If all data blocks have been used before convergence could be achieved, reuse the data by going back to Step 3 and continue the process until convergence is reached.

Table 2
On-line implementation

Step 0: Initializations: $\boldsymbol{H}(n)=\mathbf{0}_{N\times N}$ $(n=1,2,\ldots,M-1)$; $\boldsymbol{H}(0)=\mathbf{I}_{N\times N}$; $\mathbf{P}_{\mathbf{XX}}=\mathbf{0}_{N\times N\times K}$ (K is the block size of FFT).
Step 1: The observed signal samples are arriving with time. A sliding window is used to pick up a block of the most recent observation signal samples;
Step 2: Based on the data within the sliding window, estimate $\mathbf{P}_{\mathbf{XX}}(l,\omega)$ using (34);
Step 3: Calculate $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ using (21);
Step 4: Update $\boldsymbol{H}(n)$ and $\mathbf{H}(\mathrm{e}^{\mathrm{j}\omega})$ using (33);
Step 5: The sliding window is shifted to get new observation signal samples. Go back to Step 2.

The above natural-gradient algorithms can be implemented in two ways. The first one is an off-line approach, described in Table 1. After the convergence of the algorithm (33) as described in Table 1, there are at least $K_m\geqslant N$ different PSDMs $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ being jointly diagonalized, thus, if the separability conditions are fully satisfied, the source signals will be separated. In contrary, if the sources are not separated by the diagonalization of K_m PSDMs, then we re-segment the mixtures into more blocks and do the JD again to separate the sources.

The above off-line mode is suitable for the cases where the observed data are already available and may not be of sufficient duration to allow for on-line adaptation.

The other way is to implement the algorithm in an on-line way. That is, the filter coefficients are updated when new observation data are coming in. The on-line algorithm is described in Table 2.

Obviously the on-line implementation is suitable for cases where real time separation of mixtures is required.

For on-line implementation, the PSDM $\mathbf{P}_{\mathbf{XX}}(l,\omega)$ can be estimated adaptively as follows:

$$\mathbf{P}_{\mathbf{XX}}(l,\omega_k)=\beta\mathbf{P}_{\mathbf{XX}}(l-1,\omega_k)+(1-\beta)\mathbf{X}(l,\mathrm{e}^{\mathrm{j}\omega_k})\mathbf{X}^{\mathrm{H}}(l,\mathrm{e}^{\mathrm{j}\omega_k}), \quad (34)$$

where $0<\beta<1$. After convergence, the on-line algorithm will jointly diagonalize at least $K_m \geqslant N$ different PSDMs $\mathbf{P}_{\mathbf{YY}}(l,\omega)$ of the observed data, and further, for these PSDMs, the corresponding separability conditions are fully satisfied, so that the source signals are separated.

It will be useful to give some consideration on the computational complexity. Because most parts of the computational work are done in the complex domain, only complex multiplication is taken into account here. For each iteration, there are $(\frac{1}{2}N^2+1)(K/2)\log_2 K+N(N^2+\frac{5}{2}N+\frac{1}{2})K/2$ complex multiplications when the conjugate symmetry property of the FFT and the Hermitian property of PSDMs are taken into account. Further, considering that the overlap between two FFT blocks is $K/2$ (i.e., 50%), the computational complexity for each sample will be $(\frac{1}{2}N^2+1)\log_2 K+N(N^2+\frac{5}{2}N+\frac{1}{2})$. This means that the computational complexity will increase with the logarithm of the FFT block size, but it will increase cubically (N^3) with the number of sources. For instance, if the source number is $N=2$ and the FFT block size is $K=8192$, then the computational complexity for each sample is 58 complex multiplications.

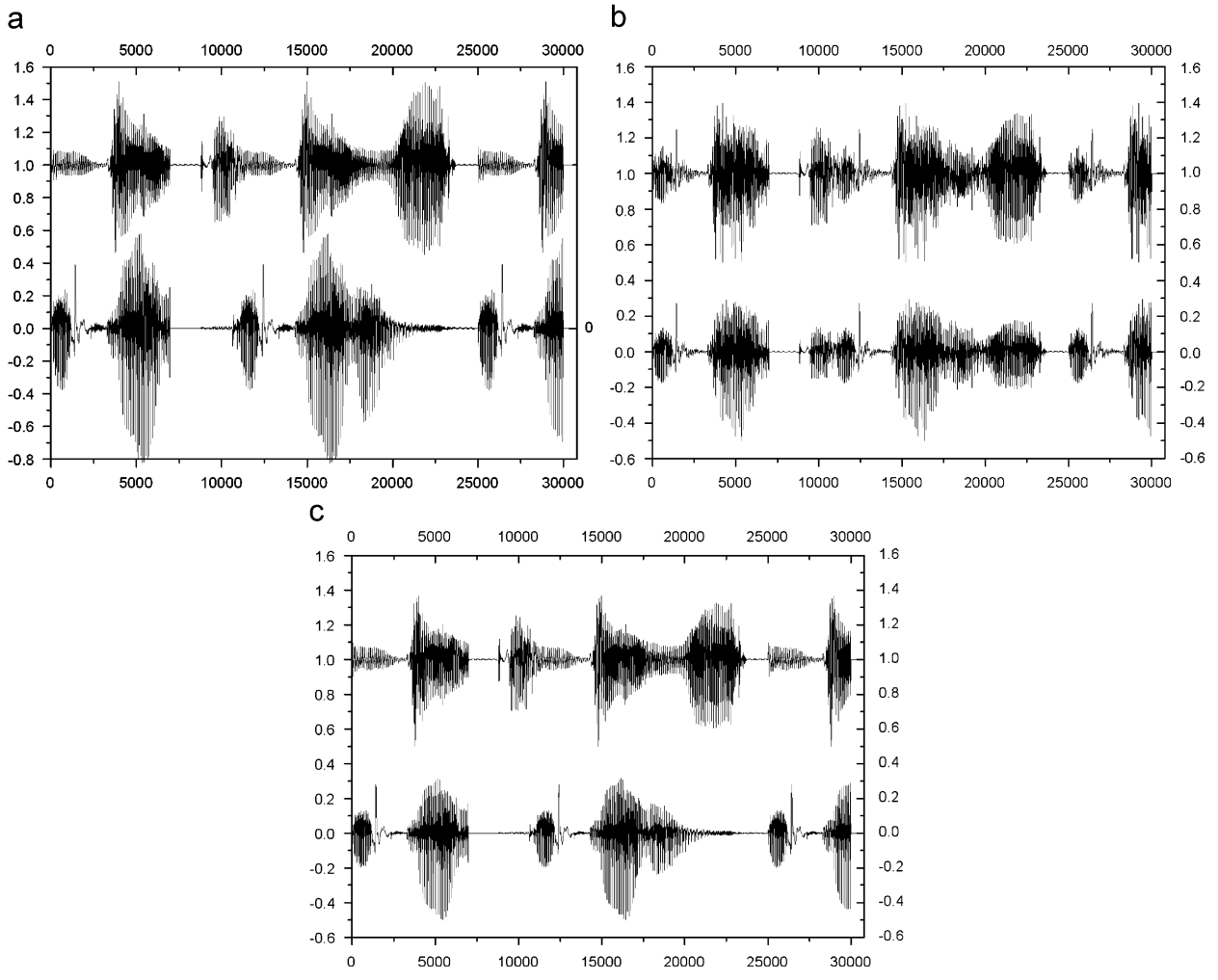

Fig. 1. Separation results for known channels: (a) sources; (b) mixtures; (c) separated sources. Excellent separation results were achieved. Before separation: $\mathrm{SIR}_1=3.53\,\mathrm{dB}$; $\mathrm{SIR}_2=3.72\,\mathrm{dB}$; after separation: $\mathrm{SIR}_1=39.49\,\mathrm{dB}$; $\mathrm{SIR}_2=41.11\,\mathrm{dB}$, respectively.

5. Simulation results

In this section, we present the simulation results for the proposed approach in a two-channel setting. The implementation uses the FFT and evaluates the spectra for frequencies $\omega_k = 2\pi k/K$.

To measure the performance, we use the signal-to-interference ratio (SIR) for the separated outputs, which is defined as follows [23]:

$\mathrm{SIR}_1 = 10\log_{10} p_{s_1}/p_{s_2}$ if the channel is considered to contain source signal 1, or
$\mathrm{SIR}_2 = 10\log_{10} p_{s_2}/p_{s_1}$ if the channel is considered to contain source signal 2,

where p_{s_1} and p_{s_2} are the powers of sources 1 and 2 contained within the output, respectively. We have tested the algorithm (33) against two types of observed signals and the results are presented as follows.

5.1. Mixtures with known channels

In the first situation, the observed mixtures were created by passing two speech signals through four convolutive channels which are given as follows:

$$h_{11}(n) = [1.0, 0.8, 0.7, 0.4, 0.3, 0.25, 0.2, 0.15],$$
$$h_{12}(n) = [0.6, 0.5, 0.5, 0.4, 0.3, 0.2, 0.25, 0.1],$$
$$h_{21}(n) = [0.5, 0.5, 0.4, 0.35, 0.3, 0.3, 0.2, 0.1],$$
$$h_{22}(n) = [1.0, 0.9, 0.8, 0.6, 0.4, 0.35, 0.3, 0.15].$$

Note that for comparison purpose the mixing channels are chosen to be the same as those in [21]. The two source signals all have 30 000 samples and are sampled at 16 000 Hz. The parameters used are as follows. The length of the separation filters is set to be $M = 8$; the FFT block size is chosen to be $K = 4096$. As the observed signals are not long enough for on-line adaptation, we have reused the data for 25 times. The on-line algorithm of Table 2 is used. The parameters $\beta = 0.6$ in (34) and $\mu = 0.01$ in (33) are used. The simulation results are shown in Fig. 1. The SIRs before separation are $\mathrm{SIR}_1 = 3.53$ dB; $\mathrm{SIR}_2 = 3.72$ dB, respectively. After separation the SIRs are obtained as $\mathrm{SIR}_1 = 39.49$ dB; $\mathrm{SIR}_2 = 41.11$ dB. Hence excellent separation is achieved with the proposed approach. The results show that the proposed approach is better

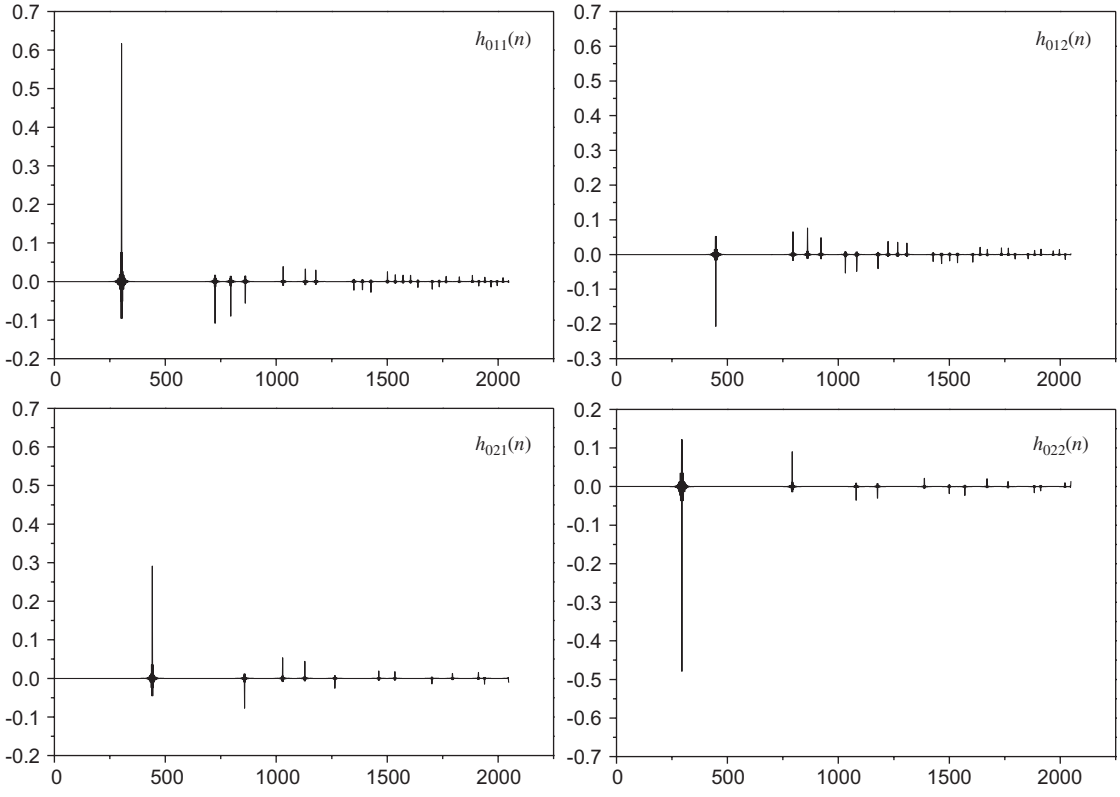

Fig. 2. The simulated responses of mixing channels: $L = 2048$; positions of microphones: [4, 6, 5] and [6, 5, 5]; positions of sources: [0, 5, 5] and [10, 5, 5].

than those in [21], by which the SIRs after separation were reported as 24.12 and 19.03 dB, respectively.

5.2. Separation of simulated mixtures

This experiment is performed with simulated mixtures in a big hall whose size is 10 m × 10 m × 10 m. We would like to thank the authors who provided the simulation Matlab code (simroommix.m) at website [36]. With this Matlab code, we first generated the impulse responses of the mixing channels according to the positions of the sources (source 1 at [0, 5, 5]; source 2 at [10, 5, 5]) and microphones (mic. 1 at [4, 6, 5]; mic. 2 at [6, 5, 5]). The impulse responses were generated for a sampling frequency of 44.1 kHz, but the speech signals were recorded with a sampling frequency of 22.05 kHz. Therefore, we first decimated the impulse responses to 22.05 kHz. The actual length of the impulse responses of the hall at sampling frequency 22.05 kHz was set to be $L = 2048$. The impulse responses of the mixing channels are shown in Fig. 2.

This experiment is performed with the on-line algorithm of Table 2 using the update Eq. (33). The remaining parameters were set as follows: Length of the separation filters: $M = 2048$; FFT block size: $K = 8192$; overlapping of blocks: 7168 samples; $\beta = 0.6$; $\mu = 0.01 - (0.01 - 0.0001)t/t_{\max}$ (where t is the iteration index and $t_{\max}$ is the maximum number of iterations) in (33).

As the sources and mixing filters are known, the SIRs can be evaluated precisely before and after separation. Before separation, the SIRs of the mixtures are 11.56 and 0.69 dB, respectively; after separation, the SIRs of the output signals of the separating system are 22.62 and 22.35 dB, respectively. Obviously, remarkable improvements have been achieved by the proposed approach. For further illustration, the responses of the separating filters and the global channel responses are shown in Figs. 3 and 4, respectively.

The above mentioned set of data was also used to investigate the influence of parameter α on the convergence behavior of algorithm (32). α is set to be $\alpha = 0.0, 0.5, 1.0$, the term $(\det[\mathbf{D}_{\mathbf{YY}}(l,\omega)]/\det[\mathbf{P}_{\mathbf{YY}}(l,\omega)])$ in (32) is limited by 10. The dynamic

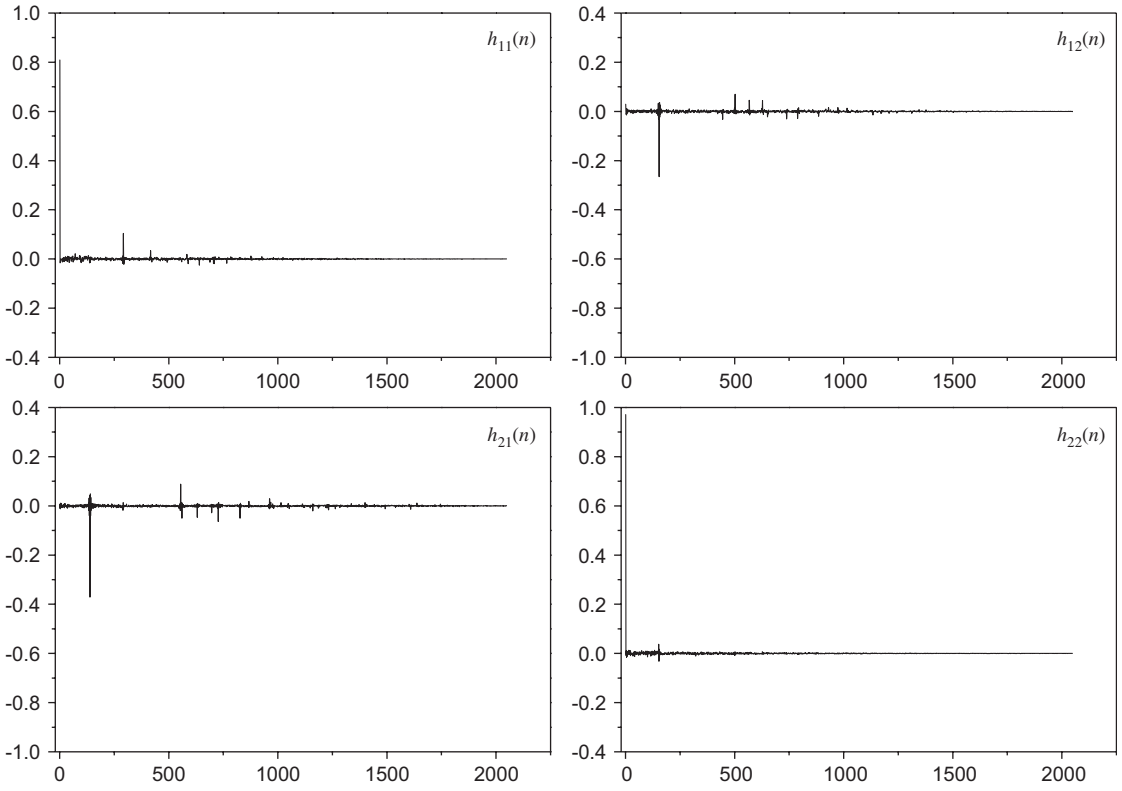

Fig. 3. The responses of separating filters after convergence: $L = 2048$.

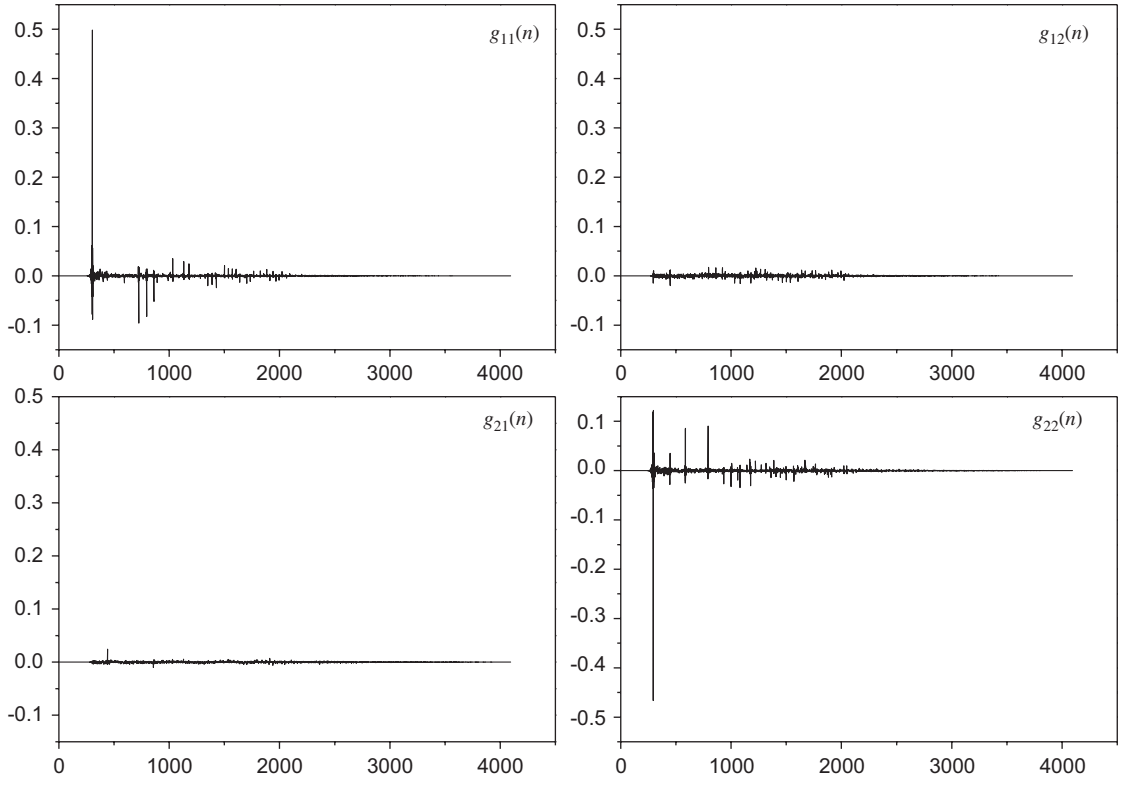

Fig. 4. The global channel responses: the global cross-channel responses are greatly decreased.

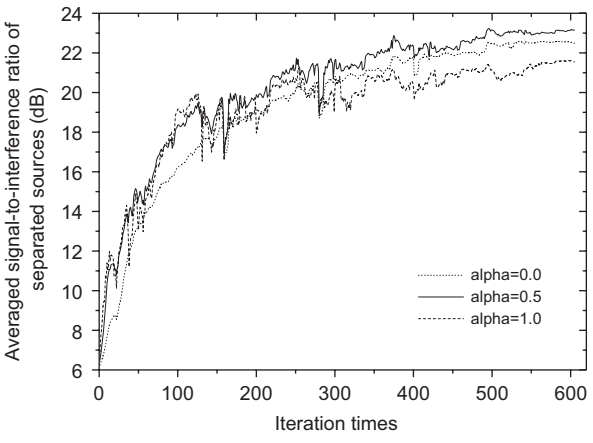

Fig. 5. The relationship between the convergence behavior and parameter α in (32).

convergence behaviors are demonstrated with the averaged SIRs of separated sources. As shown in Fig. 5, the convergence speed is slightly increased when $\alpha = 0.5$ and 1.0. After convergence, the separation performance is better when $\alpha = 0.5$ than for $\alpha = 0.0$ and 1.0.

5.3. Mixtures of speech recorded in a room

The second experiment is based on two practical test recordings of speech in a room, which were provided to the delegates of ICA'99 conference (case 1B) [37] with the on-line algorithm. The convolutive mixtures were recorded with an omni-directional microphone, and the sampling frequency is 16 000 Hz. We used the first 131 072 samples for our simulation. The parameters are as follows. The length of the separation filters is 512; the FFT block size is chosen to be $K = 8192$. We also reused the data for 20 times. The parameters were selected such that $\beta = 0.3$ in (34) and $\mu = 0.01 - (0.01 - 0.0001)\, t/t_{\max}$. The mixtures and the separated sources are shown in Fig. 6, where the mixtures and the separated sources are normalized to the range $[-0.5, 0.5]$. Listening tests showed that very good separation has been achieved. Hence we consider that output 1 contains one source (denoted as source 1) and output 2 contains the other (denoted as source 2). As the original sources are unknown, we use the following approach to estimate the SIRs for each of the two outputs [23].

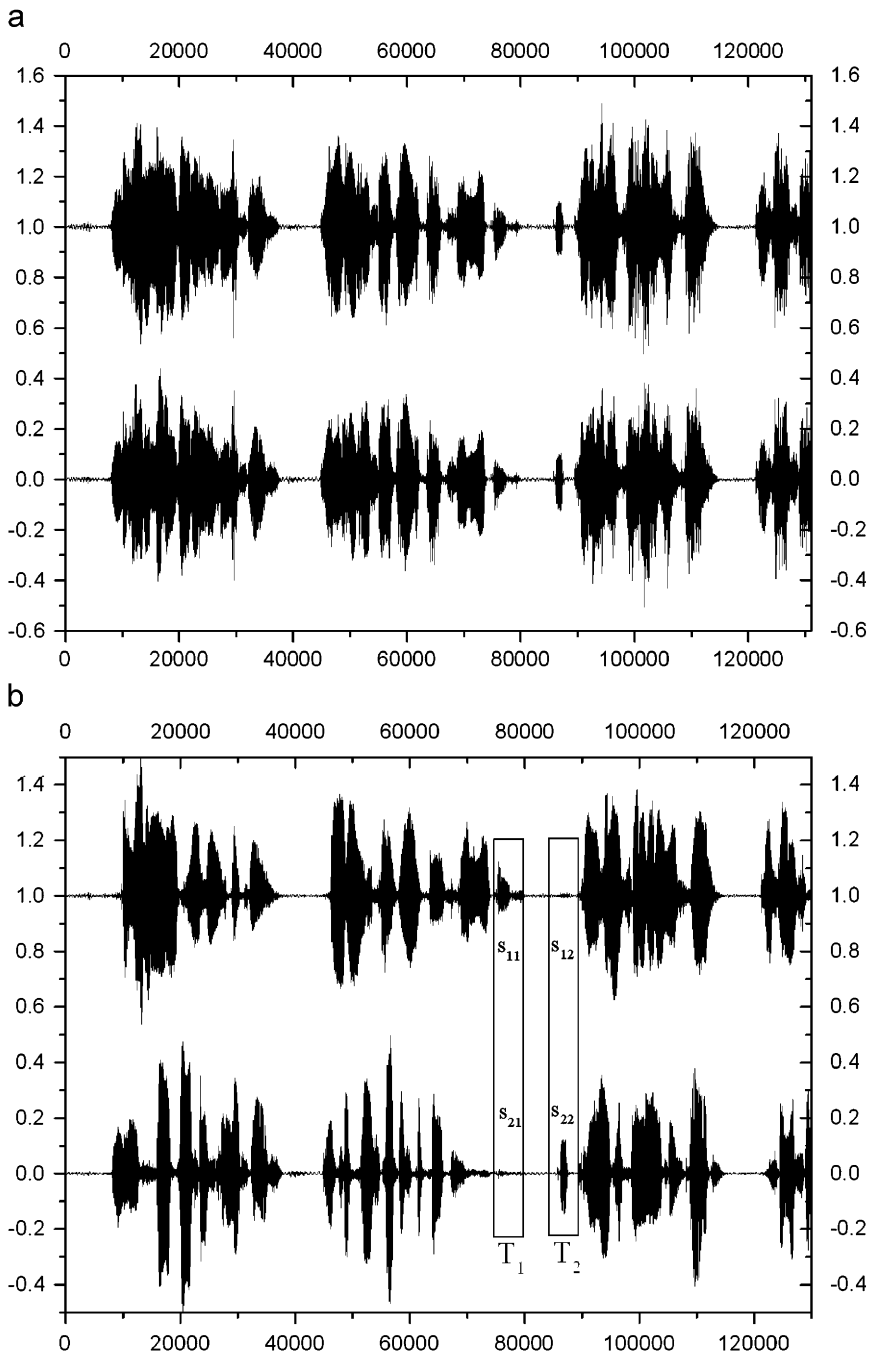

Fig. 6. The real-world recorded speech sequences and its separation results: (a) mixed speech sequences; (b) the separated speech sequences; the two segments of the separated speech sequences T_1 and T_2 ($T_1 = T_2$), which contain 5000 samples, respectively, are used to evaluate the separation performance. The estimated SIRs of separated results are 21.81 and 19.97 dB, respectively.

Find a time interval T_1, during which the waveform of output 1 has a peak and output 2 exhibits low (silent) level. Denote the segment of samples in outputs 1 and 2 as s_{11} and s_{21}, respectively. It is reasonable to believe that s_{11} is the contribution of source 1 only, and that s_{21} is the leakage of source 1 to output 2.

Similarly we could find a time interval T_2, during which output 2 exhibits a peak s_{22} but output 1 is low (silent) s_{12}. Similarly s_{22} can be considered as

the contribution of source 2 only, and s_{12} the leakage of source 2 to output 1. The SIRs for outputs 1 and 2 are calculated as $10\log_{10}p_{s_{11}}/p_{s_{12}}$ and $10\log_{10}p_{s_{22}}/p_{s_{21}}$, respectively.

Based on the above approach, SIRs for channels 1 and 2 are measured as 21.81 and 19.97 dB, respectively. Note that the two mixtures have almost the same amplitude during T_1 and T_2, respectively, which means that the SIRs before separation are about 0 dB. Therefore the two output SIRs show a significant improvement by the proposed algorithm.

5.4. On the block size of the FFT

The block size/length of the FFT plays a key role in frequency-domain BSS algorithms. There are different conclusions concerning this problem [2,8]. We performed simulations to investigate this issue and the results are shown in Fig. 7. Similar to that in [23], it is found that longer FFT blocks provide better performance for BSS. As depicted in Fig. 7, the average SIR of the separation results increases almost linearly with respect to the logarithm of the block size of the FFT. The first reason for this phenomenon seems to be the presence of boundary artifacts due to finite FFT sizes that have less impact for longer FFTs. The second reason seems to be that longer FFT block makes BSS with each frequency bin closer to the instantaneous situation. Lastly, if the FFT size is not big enough such that $K<(N^2/(N-1))M$, then local permutation may take place, which will degenerate the separation performance too.

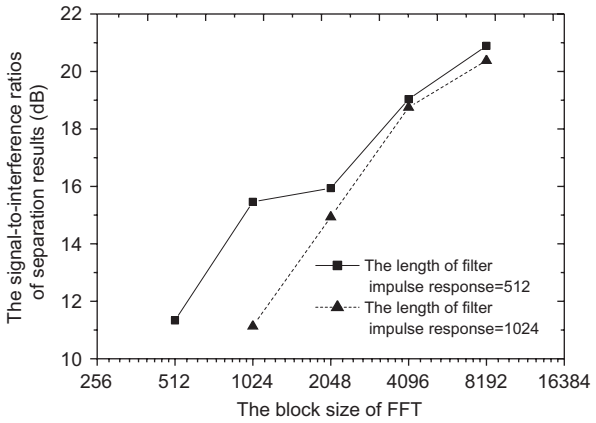

Fig. 7. The averaged signal-to-interference ratio of separation results in dependence of the FFT length. The SIR increases almost linearly in terms of the logarithm of the block size of the FFT.

5.5. Initialization strategies

For the implementation of the learning rule, the initialization of parameters is another problem we must face. It does not make sense if we initialize $\boldsymbol{H}(n)$ with completely random values. A reasonable way is to set the first values of the responses of the direct channels to be nonzeros and all others to zero, in order to guarantee that nonzero outputs of the separation system are obtained during the first iteration, e.g., $h_{ii}(0)=\theta$ $(i=1,2,\ldots,N)$ with θ being a nonzero constant. Simulations show that different choices for θ do not affect the SIRs of the outputs of the separation system, but they affect the amplitudes of the outputs. The smaller the value of θ is, the lower the amplitudes of the outputs will be. In practice, to overcome this shortage, the matrix sequence $\boldsymbol{H}(n)$ is normalized with the largest contained value in each iteration. Simulation shows that it works well. If the geometric setup is known a priori, good initializations can also be obtained from beamformer concepts, as used for example in [38].

5.6. The effect of STFT windows on the separation performance

To investigate the influence of the STFT windows on the separation performance, we performed BSS with different STFT windows which are widely used in signal processing. These windows are: Rectangular window (Rct); Triangular window (Tri); Bartlett window (Bart); Blackman window (Blkn); Blackman-Harris window (Hrrs); Bohman window (Bhm); Chebyshev window ($r=100.0$ dB) (Chb); Flat top window (flttp); Gaussian window ($\alpha=2.5$) (Gss); Hamming window (Hmm); Hann window (Hnn); Kaiser window ($\beta=2.5$) (Ksr); Nuttall window (Ntt); Parzen window (Pzn); Tukey window ($r=0.5$) (Tky). The results are listed in Table 3. We see that the Tukey window is better than the other windows for BSS, on the other hand, the flattop window is the worst one.

5.7. Comparison with other approaches

The proposed algorithm (33) is compared with the algorithms proposed by Sabala and Cichocki [35], Smaragdis [7] and Parra and Spence [2], of

Table 3
The effect of STFT windows on the separation performance (dB)

	Rct	Tri	Bart	Blkn	Hrrs	Bhm	Chb	flttp	Gss	Hmm	Hnn	Ksr	Ntt	Pzn	Tky
SIR_1	23.24	22.45	22.35	20.85	16.97	20.37	19.54	14.71	22.10	23.01	22.87	24.29	19.48	19.55	25.64
SIR_2	20.53	22.04	21.90	20.76	20.24	20.90	20.62	18.14	21.13	22.05	21.16	22.23	19.62	19.00	22.26

Table 4
Comparison with Sabala, Smaragdis and Parra's algorithms

	Sabala's Alg.	Smaragdis' Alg.	Parra's Alg.	Alg. (33)
SIR_1 (dB)	1.14	2.76	9.69	21.81
SIR_2 (dB)	8.29	16.67	12.42	19.97

which the first two exploit non-Gaussianity, and the last one relies on nonstationarity.

For Sabala's algorithm, the activation function is $f(y(n)) = \tanh(\gamma y(n))$ with $\gamma = 15$; for Smaragdis' algorithm, the activation function is $f(z) = \tanh(\mathrm{Re}\{z\}) + \mathrm{j}\tanh(\mathrm{Im}\{z\})$.

The data used in Section 5.3 were employed for algorithm comparison. The SIRs of the separated results are listed in Table 4.

It can be clearly seen in Table 4 that the proposed algorithm has a better performance.

6. Conclusions

In this paper we studied the BSS of convolutive mixtures based on the principle of JD of output PSDMs. By theoretical analysis we provided a general framework for the use of JD of output PSDMs for convolutive mixture separation. For the convolutive mixtures of nonstationary source signals, we proposed a new approach based on the introduced JD framework. The proposed approach employs a group of frequency-domain objective functions to measure to which degree the output PSDMs are diagonal, but the optimizing parameters are the separation-channel coefficients in the time domain. The proposed approach can effectively overcome the local permutation ambiguity which is usually faced in frequency-domain approaches. In addition, the proposed algorithm involves mostly the DFT and IDFT, which can be efficiently implemented by FFT algorithms. Simulation results showed that the new method is very efficient for the separation of convolutively mixed speech in both simulated and real-world environments.

Appendix A. Proof of part of Theorem 2

We take a 2-by-2 unmixing system as an example. Later, the findings are generalized to an N-by-N system.

Consider the unmixing system

$$\mathbf{H}(\mathrm{e}^{\mathrm{j}\omega}) = \begin{bmatrix} H_{11}(\mathrm{e}^{\mathrm{j}\omega}) & H_{12}(\mathrm{e}^{\mathrm{j}\omega}) \\ H_{21}(\mathrm{e}^{\mathrm{j}\omega}) & H_{22}(\mathrm{e}^{\mathrm{j}\omega}) \end{bmatrix},$$

where we assume that the impulse responses $h_{ik}(n)$ are limited to length M and that

$$\mathbf{H}(\mathrm{e}^{\mathrm{j}\omega})\mathbf{P}_{\mathbf{XX}}(m,\omega)\mathbf{H}^{\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega}) = \lambda(m,\omega)$$

is satisfied for $m = 1, 2$. That is, we assume that the length-M unmixing filters exactly diagonalize the two PSDMs. We will show that a permuted unmixing matrix $\tilde{\mathbf{H}}(\mathrm{e}^{\mathrm{j}\omega}) = \mathbf{D}(\omega)\mathbf{P}(\omega)\mathbf{H}(\mathrm{e}^{\mathrm{j}\omega})$ of length-M filters does not exist if the FFT block size K and filter length M satisfy $M < K/4$, where $\mathbf{D}(\omega) = \mathrm{diag}([D_1(\omega), D_2(\omega)]^{\mathrm{H}})$ is a diagonal matrix and $\mathbf{P}(\omega)$ is a permutation matrix.

Let

$$H_{ik}(\mathrm{e}^{\mathrm{j}\omega}) = A_{ik}(\mathrm{e}^{\mathrm{j}\omega}) + B_{ik}(\mathrm{e}^{\mathrm{j}\omega}),$$

where $A_{ik}(\omega)$ are the frequency components that are not permuted and $B_{ik}(\omega)$ are the frequency components which are to be permuted. Here, for given indices i, k, either $A_{ik}(\mathrm{e}^{\mathrm{j}\omega})$ or $B_{ik}(\mathrm{e}^{\mathrm{j}\omega})$ will be zero. Thus,

$$\tilde{H}_{11}(\omega) = D_1(\omega)[A_{11}(\mathrm{e}^{\mathrm{j}\omega}) + B_{21}(\mathrm{e}^{\mathrm{j}\omega})],$$

$$\tilde{H}_{12}(\omega) = D_1(\omega)[A_{12}(\mathrm{e}^{\mathrm{j}\omega}) + B_{22}(\mathrm{e}^{\mathrm{j}\omega})],$$

$$\tilde{H}_{21}(\omega) = D_2(\omega)[A_{21}(\mathrm{e}^{\mathrm{j}\omega}) + B_{11}(\mathrm{e}^{\mathrm{j}\omega})],$$

$$\tilde{H}_{22}(\omega) = D_2(\omega)[A_{22}(\mathrm{e}^{\mathrm{j}\omega}) + B_{12}(\mathrm{e}^{\mathrm{j}\omega})].$$

Because of the conjugate symmetry of the FFT of a real-valued impulse response, we can gather all independent frequency variables (i.e., the potentially nonzero real and imaginary parts of $H_{ik}(\mathrm{e}^{\mathrm{j}2\pi l/K})$ for $l = 0, 1, \ldots, \lceil K/2 \rceil$) in a length-$K$ real-valued vector $\mathbf{h}_{ik}$. The IFFT can then be written as a matrix

multiplication with a real-valued $K \times K$ matrix $\mathbf{W}$:

$$\mathbf{h}_{ik} = \mathbf{W}\mathbf{h}_{ik}.$$

The first M elements of $\mathbf{h}_{ik}$ will be potentially nonzero, whereas the last $K-M$ entries will be zero. With $\mathbf{C}$ containing the last $K-M$ rows of $\mathbf{W}$, we thus have

$$\mathbf{0} = \mathbf{C}\mathbf{h}_{ik}.$$

Similarly, for $\tilde{H}_{ik}(e^{j2\pi l/K})$ we define vectors $\tilde{\mathbf{h}}_{ik}$ that need to satisfy

$$\mathbf{0} = \mathbf{C}\tilde{\mathbf{h}}_{ik}.$$

Each vector $\mathbf{h}_{ik}$ can be split into two vectors $\mathbf{a}_{ik}$ and $\mathbf{b}_{ik}$, where $\mathbf{a}_{ik}$ has length M_a and $\mathbf{b}_{ik}$ has length $M_b = K - M_a$, they are collections of the components of $\mathbf{h}_{ik}$ which will not and will be permuted, respectively. Without loss of generality, we can assume that $M_b \leqslant M_a$. With matrices $\mathbf{G}_a$ and $\mathbf{G}_b$ containing $K-1$ zeros and a single one in each column, we can write

$$\mathbf{h}_{ik} = \mathbf{G}_a\mathbf{a}_{ik} + \mathbf{G}_b\mathbf{b}_{ik}.$$

Similarly we have

$$\tilde{\mathbf{h}}_{ik} = \mathbf{D}_i\mathbf{G}_a\mathbf{a}_{ik} + \mathbf{D}_i\mathbf{G}_b\mathbf{b}_{i'k},$$

where $\mathbf{D}_i$ are diagonal matrices containing the frequency scaling introduced with $D_i(\omega)$ and $i' \neq i$. Alternatively can write

$$\tilde{\mathbf{h}}_{ik} = \mathbf{G}_a\mathrm{diag}(\mathbf{a}_{ik})\mathbf{d}_{i,a} + \mathbf{G}_b\mathbf{c}_{ik},$$

where $\mathbf{d}_{i,a}$ is of length M_a and $\mathbf{c}_{ik}$ is of length M_b.

Instead of trying all possible permutations, we investigate the degrees of freedom for choosing $D_1(\omega)$, $D_2(\omega)$, and $B_{ik}(e^{j\omega})$ for arbitrarily selected $A_{ik}(e^{j\omega})$.

We obtain two independent sets of homogeneous equations for $i = 1, 2$:

$$\mathbf{0} = \mathbf{C}\mathbf{G}_a\mathrm{diag}(\mathbf{a}_{i1})\mathbf{d}_{i,a} + \mathbf{C}\mathbf{G}_b\mathbf{c}_{i1},$$

$$\mathbf{0} = \mathbf{C}\mathbf{G}_a\mathrm{diag}(\mathbf{a}_{i2})\mathbf{d}_{i,a} + \mathbf{C}\mathbf{G}_b\mathbf{c}_{i2}.$$

Each set contains $2(K-M)$ equations. The number of unknowns is $M_a + 2M_b = K + M_b$. Taking into account that, in order to avoid the trivial solution, one of the unknowns can be chosen arbitrarily, this gives $K + M_b - 1$ unknowns in total. Assuming that all $2(K-M)$ equations are linearly independent, we can find filters $\tilde{\mathbf{h}}_{ik} \neq \mathbf{h}_{ik}$ if

$$K + M_b - 1 \geqslant 2K - 2M.$$

Taking into account that $M_b \leqslant K/2$ this yields

$$K + K/2 - 1 \geqslant 2K - 2M.$$

Hence,

$$2M - 1 \geqslant K/2$$

and finally

$$M > K/4.$$

In other words, if $M \leqslant K/4$, the solution $H_{ik}(e^{j\omega})$ is unique and $\tilde{H}_{ik}(e^{j\omega}) = H_{ik}(e^{j\omega})$.

By deducing in the same way, we will obtain the following inequality that ensures unique solutions for the N-by-N case:

$$M \leqslant \frac{N-1}{N^2}K.$$

References

[1] P. Comon, Independent component analysis, a new concept?, Signal Processing 36 (1994) 287–314.

[2] L. Parra, C. Spence, Convolutive blind separation of non-stationary sources, IEEE Trans. Speech Audio Process. 8 (3) (2000) 320–327.

[3] S. Amari, A. Cichocki, Adaptive blind signal processing—neural network approaches, Proc. IEEE 86 (10) (1998) 2026–2048.

[4] D. Yellin, E. Weinstein, Criteria for multichannel signal separation, IEEE Trans. Signal Process. 42 (8) (1994) 2156–2168.

[5] M. Kawamoto, K. Matsuoka, N. Ohnishi, A method of blind separation for convolved non-stationary signals, Neurocomputing 22 (1998) 157–171.

[6] H. Bousbia-Salah, A. Belouchrani, K. Abed-Meraim, Blind separation of convolutive mixtures using joint block diagonalization, in: Sixth International Symposium on Signal Processing and its Applications, vol. 1, 2001, pp. 13–16.

[7] P. Smaragdis, Blind separation of convolved mixtures in the frequency domain, Neurocomputing 22 (1998) 21–34.

[8] S. Araki, S. Makino, T. Nishikawa, H. Saruwatari, Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech, IEEE Trans. Speech Audio Process. 11 (2) (2003) 109–116.

[9] N. Mitianoudis, M.E. Davies, Audio source separation of convolutive mixtures, IEEE Trans. Speech Audio Process. 11 (5) (2003) 489–497.

[10] H. Saruwatari, T. Kawamura, K. Sawai, K. Shikano, Evaluation of fast-convergence algorithm for blind source separation of real convolutive mixture, in: IEEE Proceedings of the ICSP'02, The Sixth International Conference on Signal Processing, 2002, pp. 346–349.

[11] A. Ciaramella, R. Tagliaferri, Amplitude and permutation indeterminacies in frequency domain convolved ICA, in: Proceedings of the International Joint Conference on Neural Networks, vol. 1, 2003, pp. 708–713.

[12] J.C. Principe, H.-C. Wu, Blind separation of convolutive mixtures, in: Proceedings of the IJCNN'99, International Joint Conference on Neural Networks, vol. 2, 1999, pp. 1054–1058.

[13] T. Lee, A.J. Bell, R. Orglmeister, Blind source separation of real world signals, in: International Conference on Neural Networks, vol. 4, 1997, pp. 2129–2134.

[14] Y. Zhou, B. Xu, Blind source separation in frequency domain, Signal Processing 83 (2000) 2037–2046.
[15] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, N. Kitawaki, Combined approach of array processing and independent component analysis for blind separation of acoustic signals, IEEE Trans. Speech Audio Process. 11 (3) (2003) 204–215.
[16] D.-T. Pham, C. Servière, H. Boumaraf, Blind separation of convolutive audio mixtures using nonstationarity, in: Proceedings of the ICA'03, Fourth International Symposium on Independent Component Analysis and Blind Signal Separation, Nara, Japan, April 2003, pp. 981–986.
[17] J. Anemüller, B. Kollmeier, Amplitude modulation decorrelation for convolutive blind source separation, in: Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, 2000, pp. 215–220.
[18] K. Rahbar, J. Reilly, A frequency domain method for blind source separation of convolutive audio mixtures, IEEE Trans. Speech Audio Process. 13 (5) (2005) 832–844.
[19] H. Saruwatari, T. Kawamura, K. Shikano, Fast-convergence algorithm for ICA-based blind source separation using array signal processing, in: Proceedings of the IEEE WASPAA, Workshop on Applications of Signal Processing to Audio and Acoustics, New York, October 2001, pp. 91–94.
[20] H. Sawada, R. Mukai, S. Araki, S. Makino, A robust and precise method for solving the permutation problem of frequency-domain blind source separation, IEEE Trans. Speech Audio Process. 12 (5) (September 2004) 530–538.
[21] K. Rahbar, J. Reilly, Blind source separation of convolved sources by joint approximate diagonalization of cross-spectral density matrices, in: Proceedings of the ICASSP'01, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 5, 2001, pp. 2745–2748.
[22] M. Kawamoto, Y. Inouye, Blind deconvolution of MIMO-FIR systems with colored inputs using second-order statistics, IEICE Trans. Fundamentals E86-A (3) (2003) 597–604.
[23] T. Mei, J. Xi, F. Yin, A. Mertins, J.F. Chicharo, Blind source separation based on time-domain optimization of a frequency-domain independence criterion, IEEE Trans. Audio Speech Language Process. 14 (6) (2006) 2075–2085.
[24] L. Tong, R. Liu, Blind estimation of correlated source signals, in: Proceedings of the ACSSC, The twenty-fourth Asilomar Conference on Signal, Systems and Computers, vol. 1, 1990, pp. 258–262.
[25] A. Belouchrani, K. Abed-Meraim, J.F. Cardoso, E. Moulines, A blind source separation technique using second-order statistics, IEEE Trans. Signal Process. 45 (2) (1997) 434–443.
[26] C. Chang, Z. Ding, S.F. Yau, F.H.Y. Chan, A matrix-pencil approach to blind separation of colored nonstationary signals, IEEE Trans. Signal Process. 48 (3) (2000) 900–907.
[27] S. Choi, A. Cichocki, Blind separation of nonstationary sources in noisy mixtures, Electronics Lett. 36 (9) (2000) 848–849.
[28] J.F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, Radar Signal Process. IEE Proc. F 140 (6) (1993) 362–370.
[29] L. Tong, R. Liu, V.C. Soon, Y.-F. Huang, Indeterminacy and identificability of blind identification, IEEE Trans. Circuits Systems 38 (5) (1991) 499–509.
[30] F. Yin, T. Mei, J. Wang, Blind source separation based on decorrelation and nonstationarity, IEEE Trans. Circuits Systems I 54 (5) (2007) 1150–1158.
[31] P.P. Vaidyanathan, Multirate Systems and Filter Banks, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[32] J.G. Proakis, C.M. Rader, F. Ling, C.L. Nikias, Advanced Digital Signal Processing, Macmillan, New York, 1992.
[33] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, 1985, p. 477.
[34] S. Amari, Natural gradient works efficiently in learning, Neural Comput. 10 (1998) 251–276.
[35] I. Sabala, A. Cichocki, S. Amari, Relationships between instantaneous blind source separation and multichannel blind deconvolution, in: IEEE World Congress on Computational Intelligence, Neural Networks Proceedings, vol. 1, 1998, pp. 39–44.
[36] See ⟨http://sound.media.mit.edu/ica-bench/⟩.
[37] See ⟨http://www2.ele.tue.nl/ica99/realworld2.html⟩ Case 1B.
[38] M. Gupta, S.C. Douglas, Beamforming initialization and data prewhitening in natural gradient convolutive blind source separation of speech mixtures, in: Independent Component Analysis and Signal Separation, vol. 4666, Springer, Berlin, 2007, pp. 512–519.