1-1-2006

Blind source separation based on time-domain optimization of a frequency-domain independence criterion

Tiemin Mei
Dalian University of Technology, China, tiemin@uow.edu.au

Jiangtao Xi
University of Wollongong, jiangtao@uow.edu.au

Fuliang Yin
Dalian University of Technology

A. Mertins
University of Wollongong, mertins@uow.edu.au

Jose F. Chicharo
University of Wollongong, chicharo@uow.edu.au

Blind Source Separation Based on Time-Domain Optimization of a Frequency-Domain Independence Criterion

Tiemin Mei, Jiangtao Xi, *Member, IEEE*, Fuliang Yin, Alfred Mertins, *Senior Member, IEEE*, and Joe F. Chicharo, *Senior Member, IEEE*

***Abstract*—A new technique for the blind separation of convolutive mixtures is proposed in this paper. Inspired by the works of Amari, Sabala , and Rahbar, we firstly start from the application of Kullback–Leibler divergence in frequency domain, and then we integrate Kullback–Leibler divergence over the whole frequency range of interest to yield a new objective function which turns out to be time-domain variable dependent. In other words, the objective function is derived in frequency domain which can be optimized with respect to time domain variables. The proposed technique has the advantages of frequency domain approaches and is suitable for very long mixing channels, but does not suffer from the local permutation problem as the separation is achieved in time-domain.**

***Index Terms*—Blind source separation (BSS), convolutive mixtures, frequency domain, integrated objective function, Kullback–Leibler divergence.**

I. Introduction

BLIND SOURCE separation (BSS) has been an active research topic during the past decade due to its potential applications in many areas. As a special case, the separation of instantaneous mixtures has been intensively studied and many approaches have been proposed with very good performance [1]–[5]. However, a more challenging situation is the so-called convolutive mixing problem, where observation signals are the mixtures of signal sources via a multiple-input–multiple-output (MIMO) system. Although extensive work has been conducted, convolutive BSS is still an open issue as there is still not a good solution for some practical situations with long mixing channels, such as audio or speech signals in a reverberant conference room [6]–[17].

A general way for solving the convolutive BSS problem is to extend the approaches for instantaneous mixtures to the case of convolutive mixtures, which can be done in either time or frequency domain. The corresponding methods are referred to as time-domain and frequency-domain approaches, respectively. With time-domain approaches, an objective function that measures the independence of the outputs of the separation system is usually defined as a function of the impulse responses of the unmixing system. Examples of time-domain approaches are [6], [7], [11], [18]–[20], and [33]. An advantage associated with the time-domain approaches is that they usually do not suffer from the so-called unknown local permutation problem at frequency level. However, the time-domain approaches are usually not able to achieve good separation for long mixing channels.

Frequency-domain approaches are considered as promising techniques for BSS in the cases of very long mixing channels. The scenario behind is that convolutive mixtures in the time domain are equivalent to instantaneous mixtures in frequency domain. Hence, by transferring mixtures into frequency domain, the approaches for instantaneous mixture separation can be applied to every individual frequency bin, and good separation can be achieved. However, the frequency-domain approach suffers from the local permutation problem in that the separated subsignals can be misaligned which makes the restoration of signals very difficult [8], [9], [24].

People have carried out extensive work to remedy the local permutation problem. The general way is to identify the local permutation based on source signal and/or BSS system properties. Examples are those making use of the fact that two successive frequency-domain separating matrices should be more similar to each other when they are in the same local permutation [8], [26]–[28]. Applying smoothness constraints to the separating system in the frequency domain is another way to overcome the local permutation problem [9]. Some researchers investigated this problem in other ways [21], [23], [29].

Despite extensive efforts so far, the local permutation ambiguity problem is still a challenging issue. A better way would be to avoid local permutation rather than to identify it. The idea in this paper is, therefore, to build objective functions in the frequency domain that keep the advantages of the frequency-domain approaches, but capture the optimizing parameters in the

Manuscript received February 16, 2005; revised October 6, 2005. This work was supported in part by the National Natural Science Foundation of China under Grant 60172073 and Grant 60372082, and the Trans-Century Training Program Foundation for the Talents by the Ministry of Education of China and in part by the Australia Research Council under ARC large Grant A00103052, and in part by the German Research Foundation under Grant ME1170/1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hong-Goo Kang.

T. Mei is with the School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116023, China, on leave from the School of Information Science and Engineering, Shenyang Institute of Technology, Shenyang 110168, China (e-mail: meitiemin@163.com).

J. Xi and J. F. Chicharo are with the School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong, Wollongong NSW 2522, Australia (e-mail: jiangtao@uow.edu.au; joe_chicharo@uow.edu.au).

F. Yin is with the School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116023, China (e-mail: flyin@dlut.edu.cn).

A. Mertins is with the Signal Processing Group, Institute of Physics, University of Oldenburg, Oldenburg 26111, Germany (e-mail: alfred.mertins@uni-oldenburg.de).

Digital Object Identifier 10.1109/TASL.2006.872623

time domain. In other words, the separation network is still defined in the time domain, but the parameters of the separation system are optimized based on a frequency-domain objective function. Note that such approaches have also been exploited in [12] and [15]. However, in [12], the method is based on the identification of the mixing system, followed by a step where the identified mixing system is inverted to achieve BSS. In [15], the approach is based on second-order statistics and considers the deconvolution of MIMO finite-impulse response (FIR) systems in a setting where colored stationary source processes are assumed.

In this paper, we propose an approach for convolutive BSS which does not suffer from the frequency-domain local permutation problem. The new method is based on the existing work of using Kullback–Leibler divergence (KL-divergence) for instantaneous BSS [2]. An objective function, based on the integration of KL-divergence applied to each frequency bin, is defined in the frequency domain. As a function of the time-domain parameters of the separation system, the objective function is optimized in the time domain, so that the local permutation problem at each frequency bin is avoided.

Note that similar cost functions as the one used in the present work have been considered in [33] and [17]. While the method in [33] is a direct time-domain approach, the work in [17] considers both time- and frequency-domain methods. The frequency-domain method in [17] determines the gradient of the cost function with respect to the independent demixing matrices at different frequency bins and then truncates the resulting (longer) time-domain impulse responses to a predefined maximum length, similar to the method in [9]. By contrast, we determine the gradient of the frequency-domain cost function directly with respect to the time-domain parameters. A further distinction between our approach and that of [33] and [17] is that our optimization is done under the assumption of phase-independent probability density functions in the frequency domain, which appears to be more robust than the standard gradient.

This paper is organized as follows. Section II describes the BSS problem considered in this paper. In Section III, the KL-divergence and the work by Amari *et al.* [2], [7], [30] are reviewed with the purpose of deriving the objective function for the new algorithm. Section IV gives the details of the new approach. Simulation results for the proposed algorithm are given in Section V, and Section VI concludes the paper.

In this paper, the following notations will be used. Matrices and vectors are printed in boldface letters. The term det(.) denotes the determinant operator, and the superscript $^{\mathrm{T}}$ means matrix transposition. In addition, the superscript $^{-\mathrm{T}}$ will be used as the inversion and transpose operator, the superscript $^{\mathrm{H}}$ denotes Hermitian transposition, and the superscript $^{-\mathrm{H}}$ means Hermitian transposition and inversion of a matrix. The superscript * stands for complex conjugation.

II. Problem Statement

In this paper, we only consider N-by-N case, where we have N signal sources, N observation signals, and N separated signals. The mixing channels are assumed to be FIR of length L, and the separating channels are also FIR with length $M \geq (N-1)(L-1)+1$ [12]. We assume that the sources are real, zero mean, and independent of each other. Moreover, the mixing system is considered to be linear and time invariant. We will use $\mathbf{s}(n) = [s_1(n), s_2(n), \ldots, s_N(n)]^{\mathrm{T}}$ to denote the signal sources, $\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_N(n)]^{\mathrm{T}}$ to denote the observations, and $\mathbf{y}(n) = [y_1(n), y_2(n), \ldots, y_N(n)]^{\mathrm{T}}$ to denote the separated outputs.

The noise-free convolutive mixing model is given as follows:

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

where $*$ denotes the convolution operation. The matrices $\boldsymbol{A}(n)$ are given by $\boldsymbol{A}(n) = [a_{ij}(n)]_{N\times N}$, where $a_{ij}(n)$ is the impulse response of the channel from source $s_j(n)$ to observation $x_i(n)$. We also assume that the transfer function matrix of the mixing system, $\mathbf{A}(z) = \sum_{n=0}^{L-1} \boldsymbol{A}(n)z^{-n}$, is nonsingular on the unit circle of the complex plane, which guarantees that the sources are separable at each frequency bin.

The separation network is also a MIMO system with FIR channels denoted by $\mathbf{H}(n) = [h_{ij}(n)]_{N\times N}$, where $h_{ij}(n)$ denotes the impulse response of the channel from $x_j(n)$ to output $y_i(n)$. $\mathbf{H}(z) = \sum_{n=0}^{M-1} \boldsymbol{H}(n)z^{-n}$ is the transfer function matrix of the separation system.

Given the definitions above, the separation system output is given as follows:

$$\mathbf{y}(n) = \boldsymbol{H}(n) * \mathbf{x}(n) = \sum_{l=0}^{M-1} \boldsymbol{H}(l)\mathbf{x}(n-l). \tag{2}$$

From (1) and (2) we have

$$\mathbf{y}(n) = \boldsymbol{H}(n) * \boldsymbol{A}(n) * \mathbf{s}(n) = \boldsymbol{G}(n) * \mathbf{s}(n) \tag{3}$$

where $\boldsymbol{G}(n) = [g_{ij}(n)]_{N\times N} = \boldsymbol{H}(n) * \boldsymbol{A}(n)$. Equation (3) can be rewritten in the z-domain as follows:

$$\mathbf{Y}(z) = \mathbf{G}(z)\mathbf{S}(z). \tag{4}$$

BSS is considered to be successful if the output $\mathbf{y}(n)$ is a permuted and filtered version of the signal sources $\mathbf{s}(n)$, which implies that the global transfer function matrix $\mathbf{G}(z)$ is of the following form:

$$\mathbf{G}(z) = \mathbf{P}\mathbf{D}(z) \tag{5}$$

where $\mathbf{P}$ is a permutation matrix and $\mathbf{D}(z)$ is a diagonal transfer function matrix.

III. BSS Based on the Kullback–Leibler Divergence

In this section, we present a brief review of the KL-divergence and the BSS algorithms developed based on it so far.

A. KL-Divergence

The KL-divergence is a fundamental means to measure the independency of a set of random variables. It is defined as follows:

$$KL\left(p_{\mathbf{y}}(\boldsymbol{y})|p_m(\boldsymbol{y})\right) = \int_{-\infty}^{\infty} p_{\mathbf{y}}(\boldsymbol{y}) \log \frac{p_{\mathbf{y}}(\boldsymbol{y})}{p_m(\boldsymbol{y})} d\boldsymbol{y} \tag{6}$$

where $p_{\mathbf{y}}(\mathbf{y})$ is the joint probability density function (pdf) of random vector $\mathbf{y} = [y_1, y_2, \ldots, y_N]^{\mathrm{T}}$ and $p_m(y) = \prod_{i=1}^{N} p_{y_i}(y_i)$ is the marginal distribution, where $p_{y_i}(y_i)$ is the pdf of y_i. Note that the KL-divergence in (6) is a nonnegative function which exhibits its minimum value of zero when the components of the random vector are independent [2].

B. KL-Divergence-Based BSS Algorithms in the Time Domain

For the instantaneous mixing cases ($L = 1$ in (1)), Amari *et al.* [2], [31] proposed an algorithm based on the KL-divergence, in which the objective function is given as

$$\phi(\boldsymbol{H}) = -\frac{1}{2} \log \left(\det(\boldsymbol{H}^{\mathrm{T}} \boldsymbol{H}) \right) - \sum_{i=1}^{N} \log p_i(y_i) \tag{7}$$

or

$$\phi(\boldsymbol{H}) = -\log |\det(\boldsymbol{H})| - \sum_{i=1}^{N} \log p_i(y_i) \tag{8}$$

where $\boldsymbol{H}$ is the separation matrix.

Based on the above objective function, a natural-gradient based approach for instantaneous mixtures was derived as follows [2]:

$$\boldsymbol{H}^{l+1} = \boldsymbol{H}^{l} + \mu \left(\mathbf{I} - \mathbf{f}\left(\mathbf{y}(l)\right) \mathbf{y}^{\mathrm{T}}(l) \right) \boldsymbol{H}^{l} \tag{9}$$

where l is a time index and iteration indicator, and $\mathbf{I}$ is the identity matrix. The term

$$\mathbf{f}\left(\mathbf{y}(l)\right) = \left[f_1\left(y_1(l)\right), f_2\left(y_2(l)\right), \ldots, f_N\left(y_N(l)\right) \right]^{\mathrm{T}}$$

with

$$f_i\left(y_i(l)\right) = -\frac{d \log\left(p_i\left(y_i(l)\right)\right)}{dy_i(l)} = -\frac{p_i'\left(y_i(l)\right)}{p_i\left(y_i(l)\right)}$$

which depends on the pdf of the sources, is referred to as the activation function.

The KL-divergence-based objective function (8) was generalized to the separation of convolutive mixtures in the time domain as follows [7]

$$\phi\left(\boldsymbol{H}(n)|_{n=0,1,\ldots,M-1}\right) = -\frac{1}{j2\pi} \oint \log \left|\det\left(\mathbf{H}(z,k)\right)\right| z^{-1} dz - \sum_{i=1}^{N} \log p_i\left(y_i(k)\right). \tag{10}$$

The corresponding natural gradient algorithm becomes:

$$\boldsymbol{H}^{l+1}(n) = \boldsymbol{H}^{l}(n) + \mu \Big(\boldsymbol{H}^{l}(n) - \mathbf{f}\left(\mathbf{y}(l - M + 1)\right) \mathbf{u}^{\mathrm{H}}(l - n) \Big) \tag{11}$$

where $\mathbf{u}(l) = \sum_{n=0}^{M-1} \boldsymbol{H}^{\mathrm{H}}(n) \mathbf{y}(l-n)$. An improved version of this method, which takes special care of a bias that occurs as a result of windowing effects, has been proposed in [33].

Alternatively, Sabala *et al.* generalized Amari's algorithm (9) to the separation of convolutive mixtures on the basis of abstract algebra theory [30]. The corresponding algorithm is as follows:

$$\boldsymbol{H}^{l+1}(n) = \boldsymbol{H}^{l}(n) + \mu \left(\boldsymbol{H}^{l}(n) - \mathbf{f}\left(\mathbf{y}(l)\right) * \mathbf{u}^{\mathrm{H}}(-l) \right) \tag{12}$$

where $\mathbf{u}(l) = \boldsymbol{H}^{\mathrm{H}}(-l) * \mathbf{y}(l)$. If the Fourier transform is applied to the two sides of (12), we will find that (12) is nothing but the approach proposed in [14].

The advantage of (11) and (12) is that they do not suffer from the frequency-domain local permutation ambiguity. However, they usually do not work very well for the separation of signals with long mixing channels.

C. KL-Divergence-Based BSS Algorithms in the Frequency Domain

With the frequency-domain approaches, observation signals are decomposed into a set of narrowband components via the short time Fourier transform (STFT), and the separation is performed for each frequency bin. The separation process can be described by the following equation:

$$\mathbf{Y}(l, e^{j\omega}) = \mathbf{H}(e^{j\omega}) \mathbf{X}(l, e^{j\omega}) \tag{13}$$

where l is a time index and

$$\mathbf{Y}(l, e^{j\omega}) = \left[y_1(l, e^{j\omega}), \ldots, y_N(l, e^{j\omega}) \right]^{\mathrm{T}}$$
$$\mathbf{X}(l, e^{j\omega}) = \left[x_1(l, e^{j\omega}), \ldots, x_N(l, e^{j\omega}) \right]^{\mathrm{T}}$$

where $y_i(l, e^{j\omega})$ and $x_i(l, e^{j\omega})$ are the STFTs of $y_i(n)$ and $x_i(n)$, respectively.

Note that (13) results from applying the STFT to (2). Strictly speaking, (13) only holds approximately because an error will be introduced when linear convolution is replaced by circular convolution. This error can be reduced by a suitable window and far larger block size of the discrete Fourier transform than the length of system impulse response. As $\mathbf{H}(e^{j\omega})$ is an instantaneous mixing matrix for any specified ω, (13) implies that instantaneous BSS approaches can be used for all the individual frequency bins.

As the mixing is instantaneous in nature for each frequency, the objective function in (7) can be directly applied to every frequency bin. Sources will be recovered from the subsignals obtained at the frequency bins. This work was done by Smaragdis [8], and the resulting objective function is as follows:

$$\phi\left(l, \mathbf{H}(e^{j\omega})\right) = -\frac{1}{2} \log \left(\det\left(\mathbf{H}^{\mathrm{H}}(e^{j\omega}) \mathbf{H}(e^{j\omega}) \right) \right) - \sum_{i=1}^{N} \log p_i\left(y_i(l, e^{j\omega}) \right). \tag{14}$$

The corresponding natural-gradient based algorithm is as follows:

$$\mathbf{H}^{l+1}(e^{j\omega}) = \mathbf{H}^{l}(e^{j\omega}) + \mu \times \left[\mathbf{I} - \mathbf{f}\left(\mathbf{Y}(l,e^{j\omega})\right)\mathbf{Y}^{\mathrm{H}}(l,e^{j\omega})\right]\mathbf{H}^{l}(e^{j\omega}) \quad (15)$$

where

$$\mathbf{f}\left(\mathbf{Y}(l,e^{j\omega})\right) = \left[f_1\left(y_1(l,e^{j\omega})\right),\ldots,f_N\left(y_N(l,e^{j\omega})\right)\right]^{\mathrm{T}}$$

is the complex-valued activation function and ω has discrete values.

The frequency-domain algorithm (15) suffers from the local permutation ambiguity. Although measures are taken to eliminate this problem, separation results are not always guaranteed [8], [9], [24].

IV. New Approach

In this section, we will construct an objective function based on the KL-divergence in the frequency domain, but the variables are the time-domain parameters of the separation system. These parameters will be obtained directly through the optimization of this objective function, so that the local permutation ambiguity for each frequency bin can be avoided. This is because the frequency-domain parameters of different frequency bins of the separation system are coupled to each other, and when the time-domain parameters are changed, they will be automatically aligned in an identical manner. In addition, we will use the phase-independent pdf definition and the polar-coordinate activation function for complex-valued variables, which was proposed in [34] and [35] and further analyzed in [25]. It was proved by [25] that polar type nonlinear activation functions behave better than the Cartesian type for complex-valued sources.

We integrate the frequency-domain objective function (14) with respect to the frequency ω and replace $p_i(y_i(l,e^{j\omega}))$ in (14) with the phase-independent pdf $p_i(|y_i(l,e^{j\omega})|)$, which yields an objective function whose variables are just the time-domain parameters of the separation channels. That is

$$\psi\left(l,\boldsymbol{H}(n)|_{n=0,1,\ldots,M-1}\right) = -\frac{1}{2}\int_{-\pi}^{\pi}\log\left(\det\left(\mathbf{H}^{\mathrm{H}}(e^{j\omega})\mathbf{H}(e^{j\omega})\right)\right)d\omega - \sum_{i=1}^{N}\int_{-\pi}^{\pi}\log p_i\left(\left|y_i(l,e^{j\omega})\right|\right)d\omega. \quad (16)$$

The integration in the objective function (16) makes it different from that in (14). It converts the objective function with respect to $\mathbf{H}(e^{j\omega})$ in (14) into a new objective function with time-domain parameters $\boldsymbol{H}(n)$ as its variables. In addition, it is showed that it is reasonable to assume that the pdf of a complex-valued signal is independent of its phase when the natural gradient based algorithm (9) is applied. The replacement of $p_i(y_i(l,e^{j\omega}))$ with $p_i(|y_i(l,e^{j\omega})|)$ will make the new algorithm more robust than (15) [25].

When comparing the objective functions (10) and (16), it is evident that there are two important differences. Firstly, the first terms in these two functions are in fact the same, but the second terms are different. In (16), the second term results from the application of KL-divergence to the sub-signals rather than the whole signals as that in (10). Second, the nonstationarity property of the sources is taken into account through the time index l.

Now, we evaluate the gradient of the objective function in (16) with respect to the channel coefficient matrix $\boldsymbol{H}(n)$

$$\frac{\partial\psi\left(l,\boldsymbol{H}(n)|_{n=0,1,\ldots,M-1}\right)}{\partial\boldsymbol{H}(n)} = -\frac{1}{2}\int_{-\pi}^{\pi}\frac{\partial\log\left(\det\left(\mathbf{H}^{\mathrm{H}}(e^{j\omega})\mathbf{H}(e^{j\omega})\right)\right)}{\partial\boldsymbol{H}(n)}d\omega - \int_{-\pi}^{\pi}\frac{\partial\left(\sum_{i=1}^{N}\log p_i\left(\left|y_i(l,e^{j\omega})\right|\right)\right)}{\partial\boldsymbol{H}(n)}d\omega. \quad (17)$$

Let us evaluate the first term of (17). As the partial differentiation in (17) is with respect to the FIR filter matrix $\boldsymbol{H}(n)$, we consider the nth coefficient of its element in the pth row and qth column. First

$$\begin{aligned}\frac{\partial\log\left(\det\left(\mathbf{H}^{\mathrm{H}}(e^{j\omega})\mathbf{H}(e^{j\omega})\right)\right)}{\partial h_{pq}(n)} &= \frac{\partial\log\left(\det\left(\mathbf{H}^{\mathrm{H}}(e^{j\omega})\right)\right)}{\partial h_{pq}(n)} + \frac{\partial\log\left(\det\left(\mathbf{H}(e^{j\omega})\right)\right)}{\partial h_{pq}(n)}\\ &= \frac{1}{\det\left(\mathbf{H}^{\mathrm{H}}(e^{j\omega})\right)}\frac{\partial\left(\det\left(\mathbf{H}^{\mathrm{H}}(e^{j\omega})\right)\right)}{\partial h_{pq}(n)} + \frac{1}{\det\left(\mathbf{H}(e^{j\omega})\right)}\frac{\partial\left(\det\left(\mathbf{H}(e^{j\omega})\right)\right)}{\partial h_{pq}(n)}.\end{aligned} \quad (18)$$

In order to evaluate the above gradient, taking the Laplacian expansion of $\det[\mathbf{H}(e^{j\omega})]$ with respect to its pth row entries gives

$$\det\left[\mathbf{H}(e^{j\omega})\right] = \sum_{l=1}^{N}H_{pl}(e^{j\omega})\det\left[\mathbf{H}_{pl}^{\mathrm{adjoint}}(e^{j\omega})\right] \quad (19)$$

where $\mathbf{H}_{pl}^{\mathrm{adjoint}}(e^{j\omega})$ is the adjoint matrix of the entry $H_{pl}(e^{j\omega})$ of matrix $\mathbf{H}(e^{j\omega})$.

The derivative of $\det[\mathbf{H}(e^{j\omega})]$ with respect to $h_{pq}(k)$ is

$$\frac{\partial\left(\det\left[\mathbf{H}(e^{j\omega})\right]\right)}{\partial h_{pq}(n)} = \det\left[\mathbf{H}_{pq}^{\mathrm{adjoint}}(e^{j\omega})\right]e^{-j\omega n}. \quad (20)$$

Deducing in the same way for $\partial\det[\mathbf{H}^{\mathrm{H}}(e^{j\omega})]/\partial h_{pq}(n)$, we get

$$\frac{\partial\left(\det\left[\mathbf{H}^{\mathrm{H}}(e^{j\omega})\right]\right)}{\partial h_{pq}(n)} = \det\left[\mathbf{H}_{pq}^{\mathrm{adjoint}}(e^{j\omega})\right]^{*}e^{j\omega n}. \quad (21)$$

Thus, (18) becomes

$$\frac{\partial\log\left(\det\left(\mathbf{H}^{\mathrm{H}}(e^{j\omega})\mathbf{H}(e^{j\omega})\right)\right)}{\partial h_{pq}(n)} = \frac{\det\left[\mathbf{H}_{pq}^{\mathrm{adjoint}}(e^{j\omega})\right]}{\det\left[\mathbf{H}(e^{j\omega})\right]}e^{-j\omega n} + \frac{\det\left[\mathbf{H}_{pq}^{\mathrm{adjoint}}(e^{j\omega})\right]^{*}}{\det\left[\mathbf{H}^{\mathrm{H}}(e^{j\omega})\right]}e^{j\omega n}. \quad (22)$$

Writing (22) in matrix form, we obtain

$$\frac{\partial\log\left(\det\left(\mathbf{H}^{\mathrm{H}}(e^{j\omega})\mathbf{H}(e^{j\omega})\right)\right)}{\partial\boldsymbol{H}(n)} = \mathbf{H}^{-\mathrm{T}}(e^{j\omega})e^{-j\omega n} + \mathbf{H}^{-\mathrm{H}}(e^{j\omega})e^{j\omega n}. \quad (23)$$

Now, we consider the second term of (17). Also taking its element in the pth row and qth column, we get

$$\begin{aligned}&\frac{\partial\left(\sum_{i=1}^{N}\log p_i\left(\left|y_i(l,e^{j\omega})\right|\right)\right)}{\partial h_{pq}(n)}\\&\quad=\frac{\partial\left(\log p_p\left(\left|y_p(l,e^{j\omega})\right|\right)\right)}{\partial h_{pq}(n)}\\&\quad=\frac{1}{p_p\left(\left|y_p(l,e^{j\omega})\right|\right)}\frac{\partial p_p\left(\left|y_p(l,e^{j\omega})\right|\right)}{\partial h_{pq}(n)}\end{aligned}\tag{24}$$

where

$$\begin{aligned}&\frac{\partial p_p\left(\left|y_p(l,e^{j\omega})\right|\right)}{\partial h_{pq}(n)}\\&=\frac{\partial p_p\left(\left|y_p(l,e^{j\omega})\right|\right)}{\partial\left|y_p(l,e^{j\omega})\right|}\frac{\partial\left|y_p(l,e^{j\omega})\right|}{\partial h_{pq}(n)}\\&=\frac{1}{2\left|y_p(l,e^{j\omega})\right|}\frac{\partial p_p\left(\left|y_p(l,e^{j\omega})\right|\right)}{\partial\left|y_p(l,e^{j\omega})\right|}\frac{\partial\left(\left|y_p(l,e^{j\omega})\right|^2\right)}{\partial h_{pq}(n)}\\&=\frac{1}{2\left|y_p(l,e^{j\omega})\right|}\frac{\partial p_p\left(\left|y_p(l,e^{j\omega})\right|\right)}{\partial\left|y_p(l,e^{j\omega})\right|}\frac{\partial\left(y_p(l,e^{j\omega})y_p^*(l,e^{j\omega})\right)}{\partial h_{pq}(n)}\\&=\frac{1}{2\left|y_p(l,e^{j\omega})\right|}\frac{\partial p_p\left(\left|y_p(l,e^{j\omega})\right|\right)}{\partial\left|y_p(l,e^{j\omega})\right|}\left(y_p^*(l,e^{j\omega})x_q(l,e^{j\omega})\right.\\&\quad\left.\times e^{-j\omega n}+y_p(l,e^{j\omega})x_q^*(l,e^{j\omega})e^{j\omega n}\right).\end{aligned}\tag{25}$$

Substituting (25) into (24), we obtain

$$\begin{aligned}&\frac{\partial\left(\sum_{i=1}^{N}\log p_i\left(\left|y_i(l,e^{j\omega})\right|\right)\right)}{\partial h_{pq}(n)}=-\frac{1}{2}\left(f_p^*\left(y_p(l,e^{j\omega})\right)\right.\\&\quad\left.\times x_q(l,e^{j\omega})e^{-j\omega n}+f_p\left(y_p(l,e^{j\omega})\right)x_q^*(l,e^{j\omega})e^{j\omega n}\right)\end{aligned}\tag{26}$$

where

$$\begin{aligned}f_p\left(y_p(l,e^{j\omega})\right)&=-\frac{1}{p_p\left(\left|y_p(l,e^{j\omega})\right|\right)}\frac{\partial p_p\left(\left|y_p(l,e^{j\omega})\right|\right)}{\partial\left|y_p(l,e^{j\omega})\right|}\\&\quad\times\frac{y_p(l,e^{j\omega})}{\left|y_p(l,e^{j\omega})\right|}\\&=-\frac{\partial\left(\log p_p\left(\left|y_p(l,e^{j\omega})\right|\right)\right)}{\partial\left|y_p(l,e^{j\omega})\right|}\\&\quad\times e^{j\theta\left(y_p(l,e^{j\omega})\right)}\end{aligned}\tag{27}$$

is the activation function with

$$\theta\left(y_p(l,e^{j\omega})\right)=\arg\left(y_p(l,e^{j\omega})\right)$$

being the phase of $y_p(l,e^{j\omega})$. Note that (27) is also referred to as a polar-coordinate activation function [25].

Now, the second term of (17) can be obtained by rewriting (26) in matrix form

$$\begin{aligned}&\frac{\partial\left(\sum_{i=1}^{N}\log p_i\left(\left|y_i(l,e^{j\omega})\right|\right)\right)}{\partial\boldsymbol{H}(n)}\\&\quad=-\frac{1}{2}\left(\mathbf{F}^*\left(\mathbf{Y}(l,e^{j\omega})\right)\mathbf{X}^{\mathrm{T}}(l,e^{j\omega})e^{-j\omega n}\right.\\&\qquad\left.+\mathbf{F}\left(\mathbf{Y}(l,e^{j\omega})\right)\mathbf{X}^{\mathrm{H}}(l,e^{j\omega})e^{j\omega n}\right)\\&\quad=-\frac{1}{2}\left(\mathbf{F}^*\left(\mathbf{Y}(l,e^{j\omega})\right)\mathbf{Y}^{\mathrm{T}}(l,e^{j\omega})\mathbf{H}^{-\mathrm{T}}(e^{j\omega})e^{-j\omega n}\right.\\&\qquad\left.+\mathbf{F}\left(\mathbf{Y}(l,e^{j\omega})\right)\mathbf{Y}^{\mathrm{H}}(l,e^{j\omega})\mathbf{H}^{-\mathrm{H}}(e^{j\omega})e^{j\omega n}\right)\end{aligned}\tag{28}$$

with

$$\mathbf{F}\left(\mathbf{Y}(l,e^{j\omega})\right)=\left[f_1\left(y_1(l,e^{j\omega})\right),\ldots,f_N\left(y_N(l,e^{j\omega})\right)\right]^{\mathrm{T}}.$$

Substituting (23) and (28) into (17), the gradient of the objective function can be obtained as

$$\begin{aligned}&\frac{\partial\psi\left(\boldsymbol{H}(n)|_{n=0,1,\ldots,M-1}\right)}{\partial\boldsymbol{H}(n)}\\&\quad=-\frac{1}{2}\left\{\int_{-\pi}^{\pi}\left[\mathbf{I}-\mathbf{F}\left(\mathbf{Y}(l,e^{j\omega})\right)\mathbf{Y}^{\mathrm{H}}(l,e^{j\omega})\right]\mathbf{H}^{-\mathrm{H}}(e^{j\omega})\right.\\&\qquad\left.\times e^{j\omega n}d\omega+\int_{-\pi}^{\pi}\left[\mathbf{I}-\mathbf{F}^*\left(\mathbf{Y}(l,e^{j\omega})\right)\mathbf{Y}^{\mathrm{T}}(l,e^{j\omega})\right]\right.\\&\qquad\left.\times\mathbf{H}^{-\mathrm{T}}(e^{j\omega})e^{-j\omega n}d\omega\right\}\\&\quad=-\int_{-\pi}^{\pi}\left[\mathbf{I}-\mathbf{F}\left(\mathbf{Y}(l,e^{j\omega})\right)\mathbf{Y}^{\mathrm{H}}(l,e^{j\omega})\right]\\&\qquad\times\mathbf{H}^{-\mathrm{H}}(e^{j\omega})e^{j\omega n}d\omega.\end{aligned}\tag{29}$$

Based on the definition in [30] and [31], the corresponding natural gradient is as follows:

$$\begin{aligned}&\left.\frac{\partial\psi\left(\boldsymbol{H}(n)|_{n=0,1,\ldots,M-1}\right)}{\partial\boldsymbol{H}(n)}\right|_{\text{Natural}}\\&\quad=\frac{\partial\psi\left(\boldsymbol{H}(n)|_{n=0,1,\ldots,M-1}\right)}{\partial\boldsymbol{H}(n)}*\boldsymbol{H}^{\mathrm{T}}(-n)*\boldsymbol{H}(n)\\&\quad=-\int_{-\pi}^{\pi}\left[\mathbf{I}-\mathbf{F}\left(\mathbf{Y}(l,e^{j\omega})\right)\mathbf{Y}^{\mathrm{H}}(l,e^{j\omega})\right]\\&\qquad\times\mathbf{H}(e^{j\omega})e^{j\omega n}d\omega.\end{aligned}\tag{30}$$

Therefore, the natural-gradient based adaptive learning rule can be obtained as follows:

$$\begin{aligned}\boldsymbol{H}^{l+1}(n)=\boldsymbol{H}^{l}(n)+\mu\int_{-\pi}^{\pi}&\left[\mathbf{I}-\mathbf{F}\left(\mathbf{Y}(l,e^{j\omega})\right)\mathbf{Y}^{\mathrm{H}}(l,e^{j\omega})\right]\\&\times\mathbf{H}^{l}(e^{j\omega})e^{j\omega n}d\omega.\end{aligned}\tag{31}$$

To find out the activation function, we assume that the STFT of the source signals has the generalized Gaussian distribution of the form [32]:

$$p_p\left(\left|y_p(l,e^{j\omega})\right|\right)=\frac{r_p}{2\sigma_p\Gamma\left(\frac{1}{r_p}\right)}e^{-\frac{1}{r_p}\left(\frac{\left|y_p(l,e^{j\omega})\right|}{\sigma_p}\right)^{r_p}} \quad (32)$$

where $\Gamma(\cdot)$ is the gamma function and $\sigma_p^r(l,\omega)=E[|y_p(l,e^{j\omega})|^r]$ is the generalized measure of variance, known as the dispersion of the distribution.

Based on (32), the activation function $f_p(y_p(l,e^{j\omega}))$ in (27) can be obtained as follows:

$$f_p\left(y_p(l,e^{j\omega})\right)=\frac{\left|y_p(l,e^{j\omega})\right|^{r_p-1}}{\sigma_p^{r_p}(l,\omega)}e^{j\theta\left(y_p(l,e^{j\omega})\right)}. \quad (33)$$

Substituting (33) into (31), we obtain

$$\boldsymbol{H}^{l+1}(n)=\boldsymbol{H}^{l}(n)+\mu\int_{-\pi}^{\pi}\left[\mathbf{I}-\mathbf{D}^{-1}(l,\omega)\mathbf{P}(l,\omega)\right]\times\mathbf{H}^{l}(e^{j\omega})e^{j\omega n}d\omega \quad (34)$$

where

$$\mathbf{D}(l,\omega)=\operatorname{diag}\left([\sigma_1^{r_1}(l,\omega),\ldots,\sigma_N^{r_N}(l,\omega)]^{\mathrm{T}}\right)$$

is a diagonal matrix with $\sigma_p^{r_p}(l,\omega)(p=1,2,\ldots,N)$ being its diagonal entries, and

$$\mathbf{P}(l,\omega)=\mathbf{Y}^{r-1}(l,e^{j\omega})\mathbf{Y}^{\mathrm{H}}(l,e^{j\omega})$$

with

$$\mathbf{Y}^{r-1}(l,e^{j\omega})=\left[\left|y_1(l,e^{j\omega})\right|^{r_1-1}e^{j\theta\left(y_1(l,e^{j\omega})\right)},\ldots,\left|y_N(l,e^{j\omega})\right|^{r_N-1}e^{j\theta\left(y_N(l,e^{j\omega})\right)}\right]^{\mathrm{T}}.$$

If the sources are nonstationary, such as speech signals, the diagonal matrix $\mathbf{D}(l,\omega)$ will change with time l, so algorithm (34) can also be seen as the joint optimization of many different objective functions $\psi(l,\boldsymbol{H}(n)|_{n=0,1,\ldots,M-1})\ (l=0,1,2,\ldots)$. Therefore, the nonstationarity of sources is not a problem, and it can even make the KL-divergence-based algorithm (34) more robust than that of stationary sources.

In the implementation, $\sigma_p^{r_p}(l,\omega)$ is computed as follows:

$$\sigma_p^{r_p}(l,\omega)=\beta\sigma_p^{r_p}(l,\omega)+(1-\beta)\left|y_p(l,e^{j\omega})\right|^{r_p} \quad (35)$$

where $0<\beta<1$ is the moving average parameter.

TABLE I
PSEUDO CODES FOR THE IMPLEMENTATION OF ALGORITHM (34)

1. Initialization: $l=0$; $H^l(0)=\mathbf{I}$, $H^l(n)=\mathbf{0}$ $(n=1,2,...,M-1)$, where $\mathbf{I}$ is identity matrix.
2. $\mathbf{x}_l=[\mathbf{x}(lk_0),...,\mathbf{x}(lk_0-K+1)]$, where k_0 indicates that the window is moved forward k_0 samples every time; K is the block size of the FFT.
3. $\mathbf{X}(l,e^{j\omega_k})=\mathbf{FFT}(\mathbf{x}_l)$.
4. $\mathbf{H}^l(e^{j\omega_k})=\mathbf{FFT}(H^l(n))$.
5. $\mathbf{Y}(l,e^{j\omega_k})=\mathbf{H}^l(e^{j\omega_k})\mathbf{X}(l,e^{j\omega_k})$, computing $y(n)$ $(n=lk_0,\ ...,\ (l+1)k_0-1)$ by the overlap-and-add technique.
6. Compute $\mathbf{D}(l,\omega_k)$ and $\mathbf{P}(l,\omega_k)$ accordingly.
7. $\Delta H^{l+1}(n)=\operatorname{win}(\mathbf{IFFT}[(\mathbf{I}-\mathbf{D}^{-1}(l,\omega_k)\mathbf{P}(l,\omega_k))\mathbf{H}^l(e^{j\omega_k})])$, where function $\operatorname{win}(.)$ takes only the first M values of $\mathbf{IFFT}[.]$.
8. $H^{l+1}(n)=H^l(n)+\mu\Delta H^{l+1}(n)$.
9. $l=l+1$. If a maximum number of iterations is reached, stop; otherwise, goto 2.
10. End.

The algorithm can be implemented with the FFT and inverse FFT (IFFT), so it has very good computational efficiency. Pseudocodes for the implementation of (34) are given in Table I.

V. SIMULATIONS

In this section, we present simulation results on the proposed algorithm with simulated and real-world recorded mixture data. We also study the influence of the FFT block size on the separation performance of the proposed algorithm. Comparisons with other approaches are presented as well.

For the purpose of evaluating the performance of our new algorithm, we define the signal-to-interference-ratio (SIR) as follows (refer to (3), and suppose that no global permutation happens):

$$\mathrm{SIR}_{y_i}=10\log_{10}\left\{\frac{\mathrm{E}\left[(g_{ii}(n)*s_i(n))^2\right]}{\mathrm{E}\left[\left(\sum_{j=1,j\neq i}^{N}g_{ij}(n)*s_j(n)\right)^2\right]}\right\} \quad (36)$$

where $*$ is the convolution operator; E[.] gives the expectation which is replaced with time average in practice.

A. Separation of Simulated Mixtures

In this first experiment, we simulated the mixing process of two speech signals in a relatively large space whose size is $10\times10\times10$ m. We would like to thank the authors who provided the simulation MATLAB code (simroommix.m) at the website.[1] With the said 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]) in the hall. 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

[1][Online]. Available: http://sound.media.mit.edu/ica-bench.

to be $L = 2048$. The impulse responses of the mixing channels are shown in Fig. 1.

In this experiment, the remaining parameters were set as follows: length of the separation filters: $M = 2048$; FFT block size: $K = 8192$; iteration times: 5; $\beta = 0.3$; $\mu = 0.01$, and $r_p = 2\,(p = 1, 2, \ldots, N)$ in (35).

As the sources and mixing filters are known, then SIRs can be evaluated precisely before and after separation. We found that, before separation, the SIRs are 11.56 and 0.69 dB, respectively, and after separation, they are 23.98 and 18.46 dB, respectively. Obviously, remarkable improvements have been achieved by the proposed approach.

The impulse responses of global channels (including the mixing and separating systems) are shown in Fig. 2.

B. Separation of Real-World Recordings

In this section, two sets of real-world recordings were used to demonstrate the performance of the proposed approach. One set of the data is very clean and contains little noise; the other contains some background noise which usually makes separation much harder. The separation results are as follows.

1) Separation of Clean Speech Mixtures: This experiment was based on two sequences of speech mixtures recorded in a room with dimensions $3.4 \times 3.8 \times 5.2$ meters (Height $\times$ Width $\times$ Depth), as provided to the delegates of the ICA'99 conference.[2] In these sequences, two male speakers are speaking simultaneously and there is no background noise. These mixtures were recorded with omnidirectional microphones, and the sampling frequency was 16 kHz. We used the first 131 072 samples for our simulation. In our experiment, the parameters of our algorithm were as follows: length of the separation filters: $M = 512$; FFT block size: $K = 4096$; iteration times: 20; $\beta = 0.3$; $\mu = 0.01$; and $r_p = 2\,(p = 1, 2, \ldots, N)$ in (35). The mixtures and the separated sources are shown in Fig. 3, where the mixtures and the separated sources have been 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 source (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.

- Find a time interval T_1 [refer to Fig. 3(b)] during which the waveform of output 1 has a peak and output 2 exhibits low (silent) samples. Denote the segment of samples in outputs 1 and 2 as $s_{11}(n)$ and $s_{21}(n)$, respectively. It is reasonable to believe that $s_{11}(n)$ is the contribution of source 1 only, and that $s_{21}(n)$ 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 $s_{22}(n)$ but output 1 is low (silent) $s_{12}(n)$. Similarly, $s_{22}(n)$ can be considered as the contribution of source 2 only, and $s_{12}(n)$ the leakage of source 2 to output 1.
- Define the average powers as $p_{s_{ij}} = \sum_{(n \in T_j)} s_{ij}^2(n)/T_j$ $(i, j \in [1, 2])$, the SIRs for outputs 1 and 2 are then calculated as $10 \log_{10}(p_{s_{11}}/p_{s_{12}})$ and $10 \log_{10}(p_{s_{22}}/p_{s_{21}})$, respectively.

[2][Online]. Available: http://www2.ele.tue.nl/ica99//realworld2.html. Case 1B.

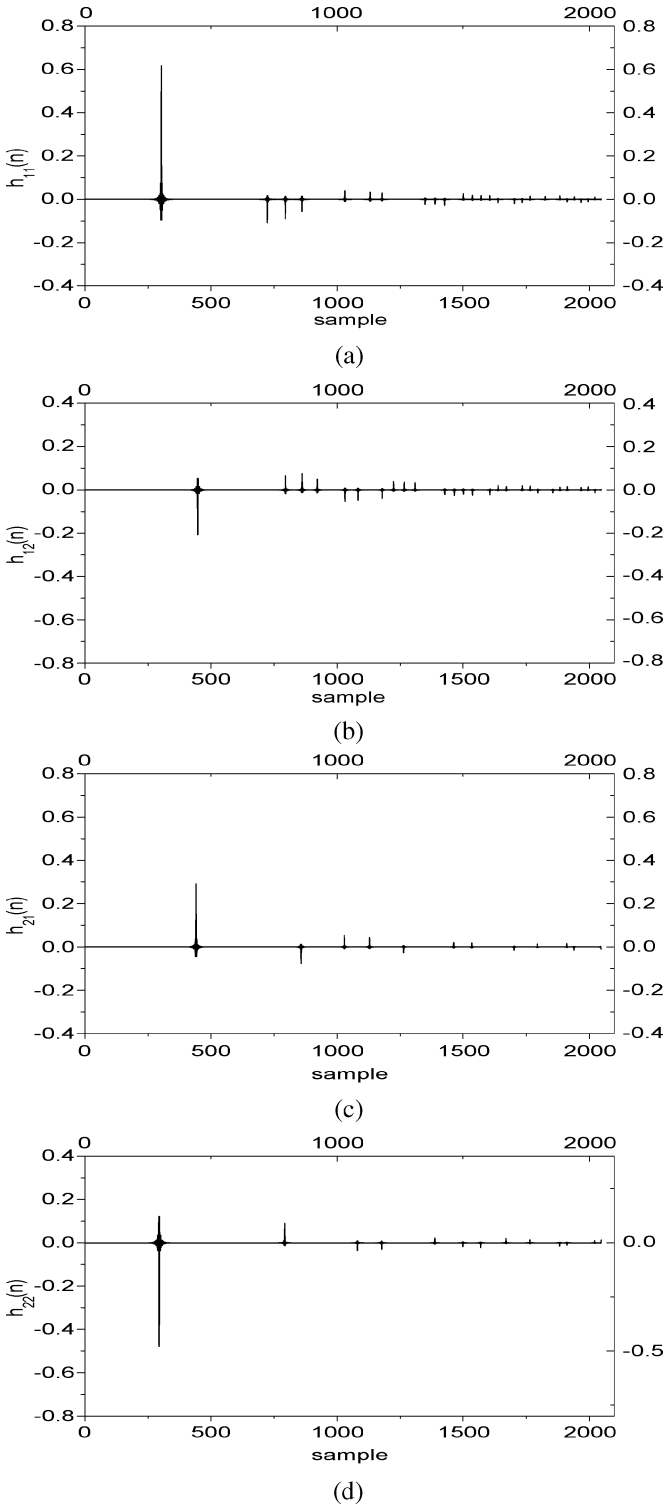


Fig. 1. Simulated impulse responses of the mixing channels in a big hall with size of $10 \times 10 \times 10$ m. The positions of the sources and the microphones are as follows: source 1 at [0, 5, 5]; source 2 at [10, 5, 5]; mic 1 at [4, 6, 5]; mic 2 at [6, 5, 5], respectively. The SIRs of the mixtures are $\text{SIR}_1 = 11.56$ dB and $\text{SIR}_2 = 0.69$ dB, respectively.

Based on the above approach, SIRs for channels 1 and 2 were measured as 23.51 and 20.58 dB, respectively. Note that the two mixtures have almost the same amplitudes during T_1 and T_2, respectively, which means that the SIRs before separation were about 0 dB. Therefore, the two output SIRs show a significant improvement as a result of the proposed algorithm.

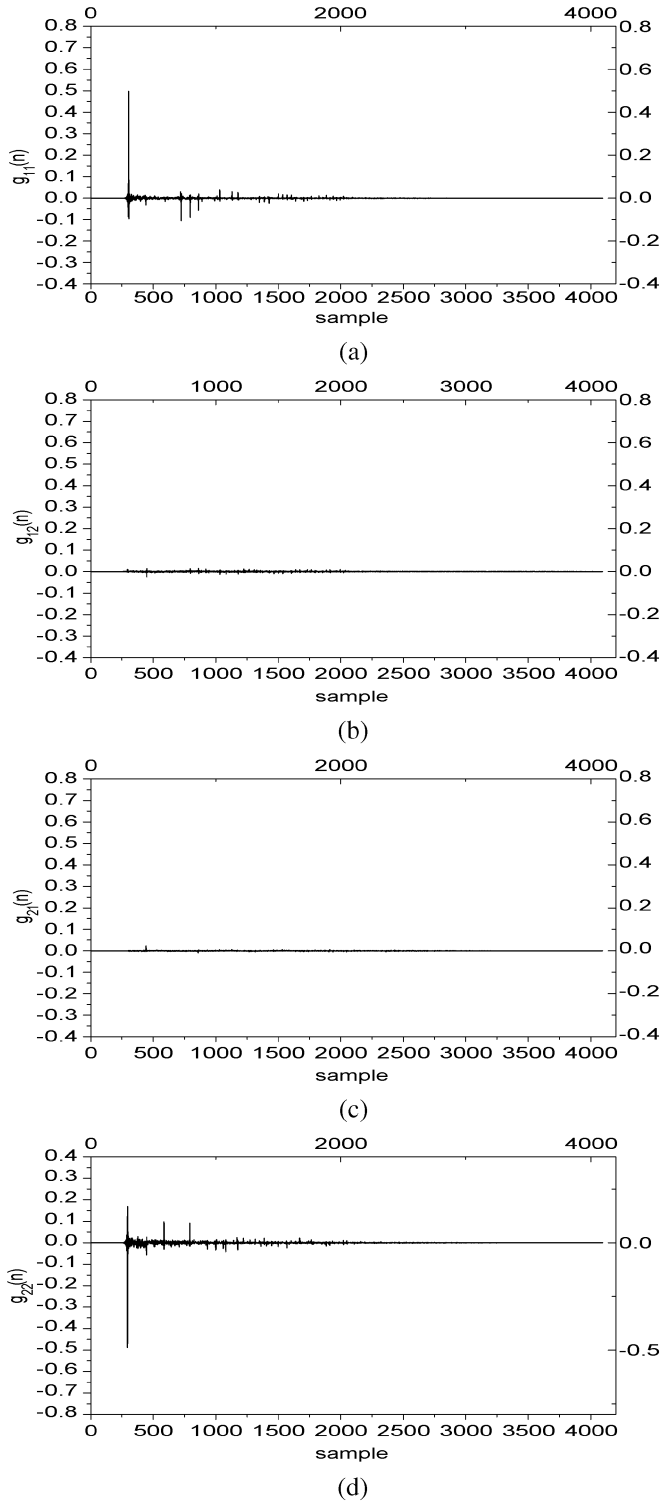


Fig. 2. Global impulse responses of the mixing and separating channels. The SIRs of the separation results are $\mathrm{SIR}_1 = 23.98$ dB and $\mathrm{SIR}_2 = 18.46$ dB, respectively.

2) Separation of Noisy Speech Mixtures: In this second experiment, the data was taken from the website.[3] The four contributions of two sources to two microphones have been recorded separately in a room with dimensions $3.1 \times 4.2 \times 5.5$ m (Height $\times$ Width $\times$ Depth). In these recordings, there is some background noise (random noise, 50-Hz interference, and harmonics). The four contributions were combined into two convolutive mixtures. These mixtures are very difficult to be separated completely because of the existence of background noise.

[3]http://www2.ele.tue.nl/ica99//realworld.html. Case 1.

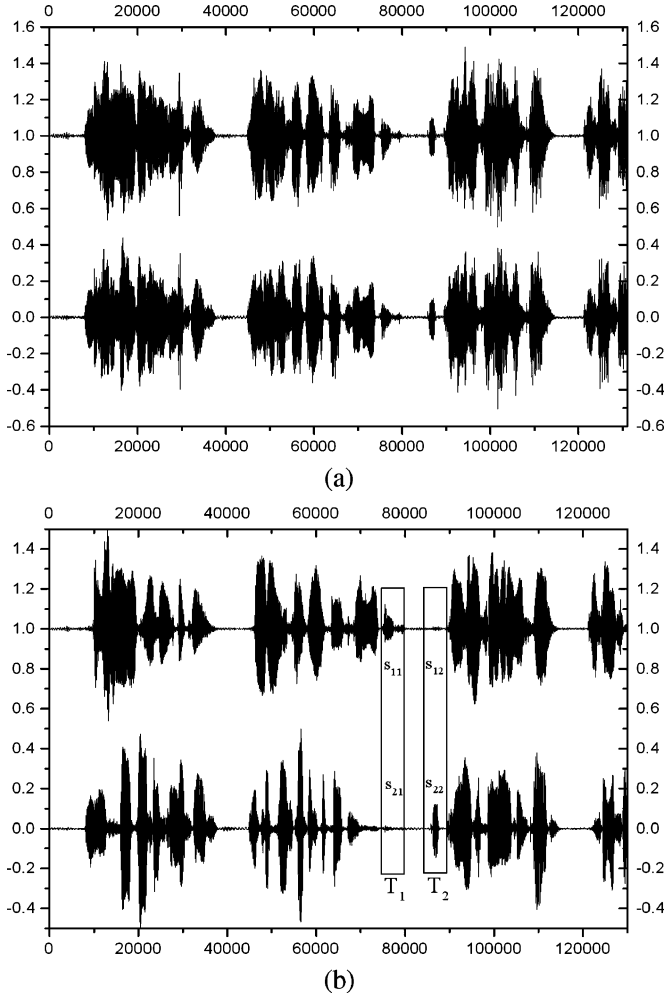


Fig. 3. Real-world recorded speech sequences and the corresponding separation results (clean data). (a) Convolutely mixed speech sequences. (b) The separated speech sequences with our new approach. The two segments of the separated speech sequences T_1 and T_2 ($T_1 = T_2$), which contain 5000 samples, respectively, were used to evaluate the separation performance.

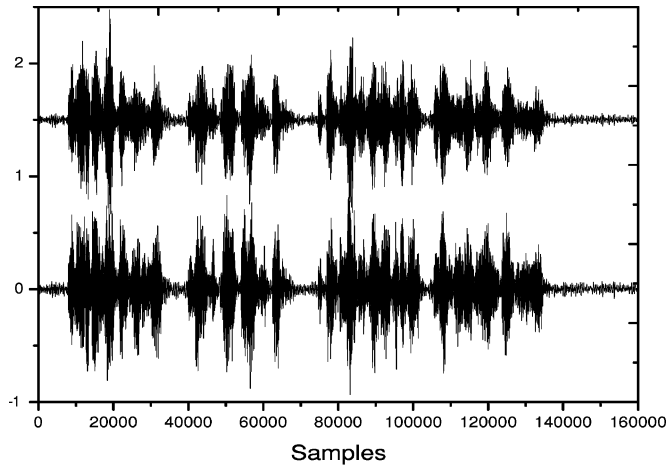
Because the contributions from different sources to different microphones are known in this case, the SIRs can be precisely evaluated before and after separation.

In our experiment, the following parameters were used. The length of the separation filters: $M = 2048$; FFT block size: $K = 4096$; iteration times: 30; $\beta = 0.6$; $\mu = 0.01$; and $r_p = 1.1$ $(p = 1, 2, \ldots, N)$ in (35).

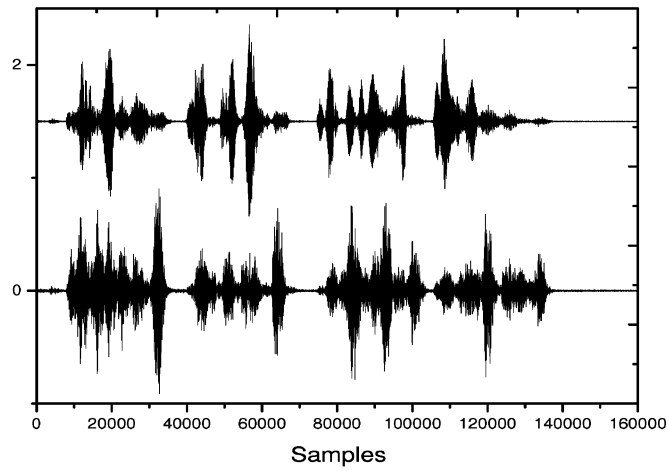
Before separation, the SIRs were 0.71 and 0.17 dB, respectively, and after separation, they became 10.38 and 18.33 dB. The mixtures and the separated sources are shown in Fig. 4.

C. Block Size of FFT and Separation Performance

When the lengths of the separating filter impulse responses are given, the block size of the FFT is a key factor, which will affect the performance of the separation algorithm. Some researchers have argued this issue, thus leading to different conclusions [9], [22]. In this experiment, the real-world recordings used in Section V-B1 were employed for studying the relationship between the separation filter length and the FFT size. For



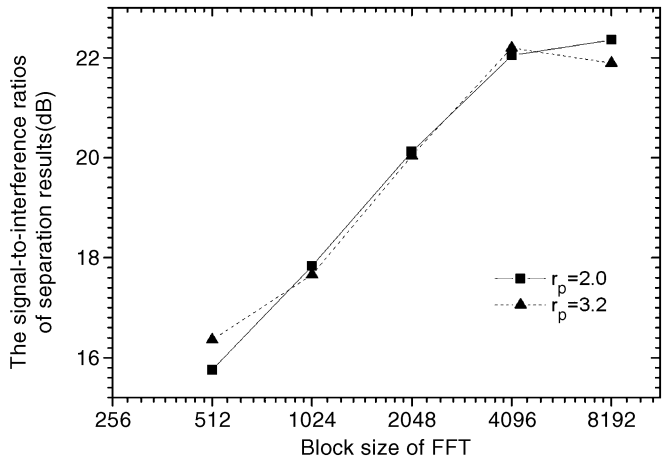
(a)



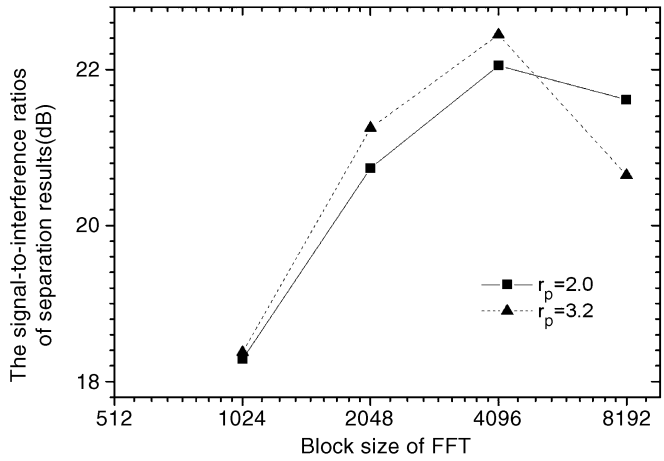
(b)

Fig. 4. Real-world recorded speech sequences and the corresponding separation results (noisy data). (a) Convolutely mixed speeches with background noises: $\mathrm{SIR}_1 = 0.71$ dB, $\mathrm{SIR}_2 = 0.17$ dB. (b) The separation results with our new approach: $\mathrm{SIR}_1 = 10.38$ dB, $\mathrm{SIR}_2 = 18.33$ dB.

our algorithm, simulations showed that the longer the block size of the FFT is, the better the performance will be. However, the improvement becomes small when the block size becomes too large. A reasonable explanation is as follows: First, when the block size of the FFT is small, errors will be introduced into the results because circular instead of linear convolution is performed with the FFT. These kind of errors will be reduced when the block size of the FFT gets large enough. Second, the basis on which the algorithm is established is that source signals are instantaneously mixed in each frequency bin, thus, the block size of the FFT should be large enough so that mixture signals in each frequency bin become much nearer to instantaneous mixtures of source signals at the corresponding frequency bin. So it is more reasonable to exploit much larger FFT block sizes (compared to the length of the impulse responses of the separating filters) when the separating algorithm is implemented in the frequency domain. On the other hand, when the block size is large enough, the above-mentioned two errors cannot be reduced infinitely, and other errors will be introduced, such as limited digits error, so the performance cannot be further improved and may even decrease when the block size is increased further.



(a)



(b)

Fig. 5. Average SIR over the FFT block size. The length of separating filter's impulse responses: (a) $M = 512$. (b) $M = 1024$.

Fig. 5 shows the relationship between the average SIR and the FFT block size. We see that the SIR increases almost linearly with the logarithm of the FFT block size when the block size is not more than 4096. In addition, when the length of the separating filters is increased from 512 taps [Fig. 5(a)] to 1024 taps [Fig. 5(b)], the resulting signal-to-interference ratios remain almost the same. This implies that the length of 512 is sufficient for the separating filters; on the other hand, it also implies that our algorithm is very stable for overestimating the separating filter length. This is very important in practice, because we do not have any knowledge of the lengths of the mixing filters; the length of the separating filters has to be long enough for a correct separation.

We also investigated the effect of the source model parameter r_p on the separation performance, as shown in Fig. 5. It was found that r_p has little influence on the separation performance. This means that the algorithm is not sensitive to the chosen source model.

D. Comparisons With Other Algorithms

In this section, the new algorithm (34) is compared with the algorithms (12) (Sabala: [30]), (15) (Smaragdis: [8]), and the approach proposed by Parra [9].

TABLE II
COMPARISON WITH SABALA'S, SMARAGDIS,' AND PARRA'S ALGORITHMS: CLEAN DATA

SIRs	Alg. (12)	Alg. (15)	L. Parra's	Alg. (34)
SIR_1	1.14dB	2.76dB	9.69dB	23.51dB
SIR_2	8.29dB	16.67dB	12.42dB	20.58dB

TABLE III
COMPARISON WITH SABALA'S, SMARAGDIS,' AND PARRA'S ALGORITHMS: NOISY DATA

SIRs	Alg. (12)	Alg. (15)	L. Parra's	Alg. (34)
SIR_1	0.75dB	0.82dB	5.94dB	10.38dB
SIR_2	1.60dB	5.50dB	4.42dB	18.33dB

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 $f(z) = \tanh(Re\{z\}) + j\tanh(Im\{z\})$.

The recordings used in Section V-B were employed for algorithm comparison.

For the case of clean mixture data in Section V-B-1, the block size of the FFT was chosen as $K = 4096$, and the filter length was set to be $M = 512$ for all the four algorithms. The signal-to-interference ratios of the separated results are listed in Table II.

For the noisy mixture data in Section V-B2, the block size of the FFT was chosen as $K = 4096$, and the filter length was set to be $M = 2048$ for all the four algorithms. The signal-to-interference ratios for the separated signals are listed in Table III.

It can be clearly seen in Tables II and III that our new algorithm has a better performance than the others.

VI. CONCLUSION

In this paper, we proposed a frequency-domain integrated objective function for convolutive BSS on the basis of the Kullback–Leibler divergence. A polar-coordinate activation function was exploited for complex-valued signals. The objective function was minimized with respect to the channel parameters of the separation system, and the corresponding algorithm was developed. The local frequency-domain permutation problem was avoided through the frequency-domain integration and time-domain optimization. Simulation results show that the algorithm does indeed lead to high performance results for the separation of real-world recorded convolutive mixtures.

REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.

[2] S. Amari and A. Cichocki, "Adaptive blind signal processing-neural network approaches," *Proc. IEEE*, vol. 86, no. 10, pp. 2026–2048, Oct. 1998.

[3] A. Belouchrani and K. Abed-Meraim *et al.*, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–443, Feb. 1997.

[4] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *Proc. Ints. Elect. Eng., Radar Signal Process. F*, vol. 140, no. 6, pp. 362–370, Dec. 1993.

[5] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Netw.*, vol. 8, no. 3, pp. 411–419, 1995.

[6] H. Bousbia-Salah, A. Belouchrani, and K. Abed-Meraim, "Blind separation of convolutive mixtures using joint block diagonalization," *6th Int. Symp. Signal Process. and its Applicat.*, vol. 1, pp. 13–16, 2001.

[7] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, "Multichannel blind deconvolution and equalization using the natrual gradient," in *Proc. IEEE Int. Workshop Wireless Commun.*, Paris, France, Apr. 1997, pp. 101–104.

[8] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomput.*, vol. 22, pp. 21–34, 1998.

[9] 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.

[10] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind convolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[11] M. Kawamoto and K. Matsuoka *et al.*, "A method of blind separation for convolved nonstationary signals," *Neurocomputing*, vol. 22, pp. 157–171, 1998.

[12] K. Rahbar and J. Reilly, "Blind source separation of convolved sources by joint approximate diagonalization of cross-spectral density matrices," in *Proc. ICASSP*, 2001, vol. 5, pp. 2745–2748.

[13] J. C. Principe and H.-C. Wu, "Blind separation of convolutive mixtures," in *Proc. IJCNN (Int. Joint Conf. Neural Netw.)*, 1999, vol. 2, pp. 1054–1058.

[14] T. Lee, A. J. Bell, and R. Orglmeister, "Blind source separation of real world signals," *Proc. Int. Conf. Neural Netw.*, vol. 4, pp. 2129–2134, 1997.

[15] M. Kawamoto and Y. Inouye, "Blind deconvolution of MIMO-FIR systems with colored inputs using second-order statistics," *IEICE Trans. Fund.*, vol. E86-A, no. 3, pp. 597–604, 2003.

[16] R. H. Lambert and A. J. Bell, "Blind separation of multiple speakers in a multipath environment," in *Proc. ICASSP*, 1997, vol. 1, pp. 423–426.

[17] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: A unified treatment," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Boston/Dordrecht/London: Kluwer, Feb. 2004, pp. 255–293.

[18] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multichannel signal separation by decorrelation," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 4, pp. 405–413, Jul. 1993.

[19] D. Van Compernolle and S. Van Gerven, "Signal separation in a symmetric adaptive noise canceler by output decorrelation," in *Proc. ICASSP*, 1992, vol. IV, pp. 221–224.

[20] D. Yellin and E. Weinstein, "Criteria for multichannel signal separation," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 2156–2168, Aug. 1994.

[21] H. Saruwatari, T. Kawamura, and K. Shikano, "Fast-convergence algorithm for ICA-based blind source separation using array signal processing," in *Proc. IEEE WASPAA*, New Paltz, NY, Oct. 2001.

[22] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 109–116, Mar. 2003.

[23] N. Mitianoudis and M. E. Davies, "Audio source separation of convolutive mixtures," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 489–497, Sep. 2003.

[24] A. Ciaramella and R. Tagliaferri, "Amplitude and permutation indeterminacies in frequency domain convolved ICA," in *Proc. Int. Joint Conf. Neural Netw.*, 2003, vol. 1, pp. 708–713.

[25] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," *IEICE Trans. Fundam.*, vol. E86, no. 3, pp. 1–7, 2003.

[26] Y. Zhou and B. Xu, "Blind source separation in frequency domain," *Signal Process.*, vol. 83, pp. 2037–2046, 2000.

[27] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 204–215, May 2003.

[28] D.-T. Pham, C. Servière, and H. Boumaraf, "Blind separation of convolutive audio mixtures using nonstationarity," in *Proc. ICA*, Nara, Japan, Apr. 2003, pp. 981–986.

[29] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. Second Int. Workshop Ind. Compon. Anal. Blind Signal Separation*, 2000, pp. 215–220.

[30] I. Sabala, A. Cichocki, and S. Amari, "Relationships between Instantaneous blind source separation and multichannel blind deconvolution," in *Proc. IEEE Int. Conf. Neural Netw.*, 1998, vol. 1, pp. 39–44.

[31] S. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, pp. 251–276, 1998.

[32] A. Cichocki and J. Karhunen *et al.*, "Neural networks for blind separation with unknown number of sources," *Neurocomputing*, vol. 24, pp. 55–93, 1999.
[33] S. C. Douglas, H. Sawada, and S. Makino, "Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 92–104, Jan. 2005.
[34] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.
[35] J. Anemüller and T. Grammss, "On-line Blind Separation of Moving Sound Sources," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA)*, 1999, pp. 331–334.

Tiemin Mei was born in Liaoning, China, on June 29, 1964. He received the B.S. degree in physics from Sun Yat-Sen University, Guangzhou, China, and the M.S. degree in biophysics from the China Medical University, Shenyang, China, in 1986 and 1991, respectively. He is currently pursuing the Ph.D. degree at the School of Electronic and Information Engineering, Dalian University of Technology, Dalian, China.

He was a Visiting Fellow with the School of Electrical Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW, Australia, from 2004 to 2005. He has also been a Member of Academic Staff at Shenyang Institute of Technology, Shenyang, since 1996. His current research interests include stochastic signal processing and speech processing.

Jiangtao Xi (M'95) received the B.S. degree from Beijing Institute of Technology, Beijing, China, in 1982, the M.S. degree from TsingHua University, Beijing, in 1985, and the Ph.D. degree from the University of Wollongong, Wollongong, NSW, Australia, in 1996, all in electrical engineering.

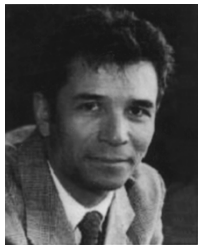
He was a Postdoctoral Fellow at the Communications Research Laboratory, McMaster University, Hamilton, ON, Canada, from 1995 to 1996 and a Member of Technical Staff at Bell Laboratories, Lucent Technologies, Murray Hill, NJ, from 1996 to 1998. He has been a Member of Academic Staff at the University of Wollongong since 1998, where he is currently a Senior Lecturer. His research interests are digital signal processing and its applications.

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

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

Alfred Mertins (M'96–SM'03) received the Diplomingenieur degree from the University of Paderborn, Paderborn, Germany, in 1984, the Dr.-Ing. degree (Ph.D.) in electrical engineering and the Dr.-Ing. habil. degree in telecommunications from the Hamburg University of Technology, Hamburg, Germany, in 1991 and 1994, respectively.

From 1986 to 1991, he was with the Hamburg University of Technology, from 1991 to 1995 with the Microelectronics Applications Center Hamburg, from 1996 to 1997 with the University of Kiel, Kiel, Germany, from 1997 to 1998 with the University of Western Australia, Perth, Australia, and from 1998 to 2003 with the University of Wollongong, Wollongong, Australia. In April 2003, he joined the University of Oldenburg, Oldenburg, Germany, where he is a Professor in the Faculty of Mathematics and Science, Institute of Physics. His research interests include speech, audio, image and video processing, wavelets and filter banks, and digital communications.



Joe F. Chicharo (M'89–SM'94) received the Bachelor's degree with first class honors and the Ph.D. degree from the University of Wollongong, Wollongong, Australia, in 1983 and 1990, respectively, both in electrical engineering.

He has been working at the University of Wollongong since 1985 as a Lecturer (1985–1990), Senior Lecturer (1990–1993), Associate Professor (1994–1997), and Professor (1997-present). He is currently the Dean of The Faculty of Informatics, the University of Wollongong. From 2000 to 2003, he was the Research Director of Australian Collaborative Research Center on Smart Internet Technology. His research interests are in the areas of signal processing, telecommunications, and information technology with over 200 research publications.