A HALF-FREQUENCY DOMAIN APPROACH FOR CONVOLUTIVE BLIND SOURCE SEPARATION BASED ON KULLBACK-LEIBLER DIVERGENCE

*Tiemin Mei, Fuliang Yin**

School of Electronic and Information Eng.
Dalian University of Technology
Dalian 116023 China

Jiangtao Xi, Alfred Mertins, Joe F. Chicharo†

School of Informatics Eng.
University of Wollongong
NSW 2522 Australia

ABSTRACT

This paper studies the problem of blind separation of convolutive mixtures by means of Kullback-Leibler divergence in frequency domain. Unlike exiting approaches, an integrated objective function is defined in frequency-domain with time-domain parameters as the variables. The permutation problem is avoided through the frequency-domain integration and time-domain optimization. Simulation results show that the algorithm is valid and of high performance for the separation of real-world recorded convolutive mixtures.

1. 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, separation of instantaneous mixtures is very successful so far and many approaches have been proposed [1][2][3][4]. However, a more challenging situation is the separation of convolutive mixtures with long mixing channels [5][6][7][8][9].

A general way for solving the convolutive BSS is to extend the approaches for instantaneous mixture to the case of convolutive mixtures, which can be done in either time or frequency domain. An advantage associated with the time-domain approaches is that they usually do not suffer from the so-called unknown permutation problem[5].

Frequency-domain approaches are considered as promising techniques for BSS in the cases of very long mixing channels. It is known that convolutive mixtures in the time domain can be considered as instantaneous mixtures in the frequency domain, so approaches for instantaneous mixture separation can be applied to convolutive mixtures at a specific frequency. However, the permutation ambiguity, which is inherited from instantaneous BSS, makes convolutive BSS very difficult [6][7].

People have done extensive work to remedy the permutation problem. The general way is to identify the permutation based on source signal and/or BSS system properties [6][7]. Despite the extensive efforts so far, the permutation ambiguity problem is still a challenging issue. A better way would be to avoid 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 the optimizing parameters are captured in the time domain.

In this paper, the proposed approach is based on the existing work of using Kullback-Leibler (KL) divergence for instantaneous BSS [2]. An objective function, which is 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 the permutation problem at frequency level is avoided.

2. PROBLEM STATEMENTS

In this paper, we only consider the N-by-N cases. The mixing channels are assumed to be FIR of length L, and the separating channels are also FIR with length $M \geq (N-1)(L-1)+1$ [8]. We assume that the sources are real, of zero mean and independent of each other, and the mixing system is linear and time invariant. We use $\mathbf{s}(n)$, $\mathbf{x}(n)$ and $\mathbf{y}(n)$ to denote the sources, mixtures and the separated outputs, respectively.

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

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

where $\boldsymbol{A}(n) = [a_{ij}(n)]_{N\times N}$ is the FIR filter mixing matrix.

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 at different frequency bins are separable.

The separation system output is given as follows:

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

*This work is supported by the National Natural Science Foundation of China under Grant No.60172073 and No.60372082, and the Trans-Century Training Program Foundation for the Talents by the Ministry of Education of China. This work is also partially supported by Australia Research Council under ARC large Grant No.A00103052.

†A. Mertins is now with the Signal Processing Group, Institute of Physics, the University of Oldenburg, Oldenburg 26111 Germany

where $\boldsymbol{H}(n) = [h_{ij}(n)]_{N\times N}$ is the separation system. From (1) and (2) we have:

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

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

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

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:

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

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

3. A HALF-FREQUENCY DOMAIN APPROACH

For the instantaneous mixing cases (*M*=1 in (2)), Amari *et al.* [2][11] proposed an algorithm based on the KL divergence function, in which the objective function is given as:

$$\phi(\boldsymbol{H}) = -\frac{1}{2}\log\left(\det(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})\right) - \sum_{i=1}^{N}\log p_i(y_i) \quad (6)$$

where $\boldsymbol{H}$ is the separation matrix. Based on this objective function, a very successful natural-gradient based approach for instantaneous mixtures was derived by Amari *et al*[2].

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

$$\mathbf{Y}(l, e^{j\omega}) = \mathbf{H}(e^{j\omega})\mathbf{X}(l, e^{j\omega}) \quad (7)$$

where l is time index and

$$\mathbf{Y}(l, e^{j\omega}) = [y_1(l, e^{j\omega}), ..., y_N(l, e^{j\omega})]^{\mathrm{T}},$$

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

Note that (7) results from applying the STFT to (2). As $\mathbf{H}(e^{j\omega})$ is an instantaneous mixing matrix for any specific ω, (7) implies that instantaneous BSS approaches can be used for all the individual frequency bins. This is the scenario behind the frequency-domain BSS approaches.

As the mixing is instantaneous in nature for each frequency bin, the objective function in (6) can be directly applied to every frequency bin. The work was done by Smaragdis [6] and the resulting objective function is

$$\phi(l, \mathbf{H}(e^{j\omega})) = -\frac{1}{2}\log\left(\det\left(\mathbf{H}^{\mathrm{H}}(e^{j\omega})\mathbf{H}(e^{j\omega})\right)\right) - \sum_{i=1}^{N}\log p_i(y_i(l, e^{j\omega})) \quad (8)$$

where $y_i(l, e^{j\omega})$ is the STFT of $y_i(n)$. The corresponding natural-gradient based algorithm is as follows:

$$\mathbf{H}^{l+1}(e^{j\omega}) = \mathbf{H}^{l}(e^{j\omega}) + \mu \times \left[\mathbf{I} - \mathbf{f}\left(\mathbf{Y}(l, e^{j\omega})\right)\mathbf{Y}^{\mathrm{H}}(l, e^{j\omega})\right]\mathbf{H}^{l}(e^{j\omega}) \quad (9)$$

where

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

is referred to as the activation function.

The frequency-domain algorithm (9) suffers from the permutation ambiguity. Although measures are taken to eliminate the permutation, separation results are not always guaranteed.

In order to overcome the permutation ambiguity, we integrate the frequency-domain objective function (8) with respect to the frequency ω, and replace $p_i(y_i(l, e^{j\omega}))$ with $p_i\left(|y_i(l, e^{j\omega})|\right)$ in (8), which yields an objective function whose variables are just the time-domain parameters of the separation channels. That is,

$$\psi(l, \boldsymbol{H}(n)\,|_{n=0,1,...,M}) = -\frac{1}{2}\int_{-\pi}^{\pi}\log\left(\det\left(\mathbf{H}^{\mathrm{H}}(e^{j\omega})\mathbf{H}(e^{j\omega})\right)\right)d\omega - \sum_{i=1}^{N}\int_{-\pi}^{\pi}\log p_i\left(|y_i(l, e^{j\omega})|\right)d\omega \quad (10)$$

The permutation problem is avoided through the optimization of this new objective function (10) with respect to time-domain parameters of separation system.

The gradient of the objective function (10) is obtained as

$$\begin{aligned}&\frac{\partial\psi(\boldsymbol{H}(n)\,|_{n=0,1,...,M-1})}{\partial\boldsymbol{H}(n)}\\ &= -\frac{1}{2}\left\{\int_{-\pi}^{\pi}\left[\mathbf{I} - \mathbf{R}_{\mathrm{F_Y Y}}(l, \omega)\right]\mathbf{H}^{-\mathrm{H}}(e^{j\omega})e^{j\omega n}d\omega\right.\\ &\quad\left. + \int_{-\pi}^{\pi}\left[\mathbf{I} - \mathbf{R}_{\mathrm{F_Y Y}}(l, \omega)\right]\mathbf{H}^{-\mathrm{T}}(e^{j\omega})e^{-j\omega n}d\omega\right\}\\ &= -\int_{-\pi}^{\pi}\left[\mathbf{I} - \mathbf{R}_{\mathrm{F_Y Y}}(l, \omega)\right]\mathbf{H}^{-\mathrm{H}}(e^{j\omega})e^{j\omega n}d\omega. \quad (11)\end{aligned}$$

where $\mathbf{R}_{\mathrm{F_Y Y}} = \mathbf{F}_{\mathbf{Y}}\left(l, e^{j\omega}\right)\mathbf{Y}^{\mathrm{H}}(l, e^{j\omega})$, and

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

and

$$f_p(y_p(l, e^{j\omega})) = -\frac{\partial\left(\log p_p\left(|y_p(l, e^{j\omega})|\right)\right)}{\partial|y_p(l, e^{j\omega})|}e^{j\theta\left(y_p(l, e^{j\omega})\right)}$$

is the polar-coordinate activation function [10], and where

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

Based on the definition in [9][11], the corresponding natural gradient is as follows:

$$\begin{aligned}&\frac{\partial\psi(\boldsymbol{H}(n)\,|_{n=0,1,\ldots,M-1})}{\partial\boldsymbol{H}(n)}\,|_{\text{Natural}}\\ &=\frac{\partial\psi(\boldsymbol{H}(n)\,|_{n=0,1,\ldots,M-1})}{\partial\boldsymbol{H}(n)}*\boldsymbol{H}^{\mathrm{T}}(-n)*\boldsymbol{H}(n)\\ &=-\int_{-\pi}^{\pi}\left[\mathbf{I}-\mathbf{R}_{\mathrm{F_Y Y}}(l,\omega)\right]\mathbf{H}(e^{j\omega})e^{j\omega n}d\omega. \qquad (12)\end{aligned}$$

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

$$\begin{aligned}\boldsymbol{H}^{l+1}(n)&=\boldsymbol{H}^{l}(n)+\mu\times\\ &\int_{-\pi}^{\pi}\left[\mathbf{I}-\mathbf{R}_{\mathrm{F_Y Y}}(l,\omega)\right]\mathbf{H}^{l}(e^{j\omega})e^{j\omega n}d\omega. \qquad (13)\end{aligned}$$

As the polar-coordinated activation function concerned, we assume that the STFT of the source signals has the generalized Gaussian distribution of the form [12]:

$$p_p\left(\left|y_p(l,e^{j\omega})\right|\right)=\frac{r_p}{2\sigma_p\Gamma(\frac{1}{r_p})}e^{-\frac{1}{r_p}\left(\frac{\left|y_p(l,e^{j\omega})\right|}{\sigma_p}\right)^{r_p}} \qquad (14)$$

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

Based on (14), the activation function $f_p\left(y_p(l,e^{j\omega})\right)$ can be obtained as follows:

$$f_p\left(y_p(l,e^{j\omega})\right)=\frac{\left|y_p(l,e^{j\omega})\right|^{r_p-1}}{\sigma_p^{r_p}(l,\omega)}e^{j\theta(y_p(l,e^{j\omega}))}. \qquad (15)$$

Substituting (15) into (13), we obtain

$$\begin{aligned}\boldsymbol{H}^{l+1}(n)&=\boldsymbol{H}^{l}(n)+\mu\times\\ &\int_{-\pi}^{\pi}\left[\mathbf{I}-\mathbf{D}^{-1}(l,\omega)\mathbf{P}(l,\omega)\right]\mathbf{H}^{l}(e^{j\omega})e^{j\omega n}d\omega \qquad (16)\end{aligned}$$

where

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

and

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

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

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

$$\sigma_p^{r_p}(l,\omega)=\beta\sigma_p^{r_p}(l,\omega)+(1-\beta)\left|y_p(l,e^{j\omega})\right|^{r_p} \qquad (17)$$

where β is the moving average parameter.

4. SIMULATIONS

4.1. Simulation results for the new algorithm

Simulations have been performed using the two speech signals which were provided to the delegates of the ICA'99 conference [13]. The convolutive mixtures were recorded with omni-directional microphones, and the sampling frequency is 16000Hz. We used the first 131072 samples for our simulation. In our simulation, the length of the separation filters is M=512; FFT block size is K=4096; Iteration times: 20; $\beta=0.3$; $\mu=0.01$. We assume that $\left|y_p(l,e^{j\omega})\right|$ are of Gaussian distribution, which implies that $r_p=2,\ p=1,...,N$ in (16). The mixtures and the separated sources are shown in Figure 1, 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 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 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} 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.

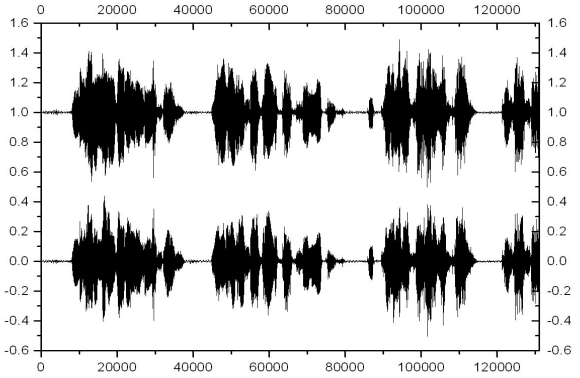
The SIRs for outputs 1 and 2 are then calculated as $10\log_{10}\frac{p_{s_{11}}}{p_{s_{12}}}$ and $10\log_{10}\frac{p_{s_{22}}}{p_{s_{21}}}$, respectively.

Based on the above approach, SIRs for channels 1 and 2 were measured as 23.51dB and 20.58dB, 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 are about 0dB. Therefore the two output SIRs show a significant improvement by the proposed algorithm.

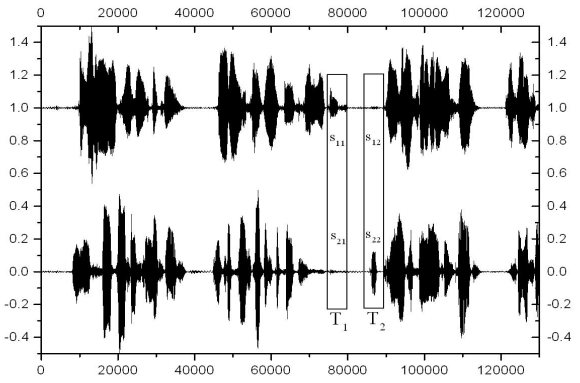
4.2. Comparisons with other algorithms

The new algorithm (16) is compared with the algorithms proposed by I. Sabala [9] and P. Smaragdis [6]. 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 block size of the FFT was chosen as K=4096, and the filter length was set to M=512 for all the three algorithms. The signal-to-interference ratios of the separated results are listed in Table I.

It can be clearly seen in Table I that our new algorithm has a better performance. On the other hand, the signal-to-interference ratios of the two output channels are quite unbalanced as algorithms [9]and [6] are concerned.



(a)



(b)

Figure 1: Real world recorded speech sequences and the corresponding 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, were used to evaluate the separation performance.

5. CONCLUSIONS

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 permutation problem was avoided through the frequency-domain integration and time-domain optimization. Simulation results show that the algorithm is valid and of high performance for the separation of real-world recorded convolutive mixtures.

6. REFERENCES

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

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

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

[4] J. F. Cardoso, A. Souloumiac, "Blind beamforming for non-Gaussian signals," *Radar and Signal Processing, IEE Proc*, F, vol. 140, issue 6, pp.362-370, Dec., 1993.

[5] 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 Communication*, Paris, pp.101-104, Apr. 1997,

[6] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp.21-34, 1998.

[7] L. Parra, C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. On Speech and Audio Processing*, vol. 8, no. 3, pp. 320-327, 2000.

[8] K. Rahbar, J. Reilly, "Blind source separation of convolved sources by joint approximate diagonalization of cross-spectral density matrices," *Proc. of ICASSP*' 01, vol. 5, pp. 2745-2748, 2001.

[9] I. Sabala, A. Cichocki, S. Amari, "Relationships between Instantaneous blind source separation and multichannel blind deconvolution," *IEEE proc. on Neural Networks*, vol.1, pp.39-44, 1998.

[10] H. Sawada, R. Mukai, S. Araki, S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," in *IEEE proc. of ICASSP'02*, vol. I, pp. 1001-1004, 2002

[11] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol.10, pp.251-276, 1998.

[12] A. Cichocki and J. Karhunen, *et al.*, "Neural networks for blind separation with unknown number of sources," *Neurocomputing*, vol. 24, pp. 55-93, 1999.

[13] See: http://www2.ele.tue.nl/ica99/realworld.html.

Table 1: Comparison with other algorithms

	Sabala's	Smaragdis'	Alg. (16)
SIR_1	1.14 dB	2.76 dB	23.51 dB
SIR_2	8.29 dB	16.67 dB	20.58 dB