SUBBAND-BASED BLIND SIGNAL PROCESSING FOR SOURCE SEPARATION IN CONVOLUTIVE MIXTURES OF SPEECH

Kostas Kokkinakis and Philipos C. Loizou

Center for Robust Speech Systems,
Department of Electrical Engineering, University of Texas at Dallas,
P. O. Box 830688, Richardson, TX 75083–0688, USA
{kokkinak, loizou}@utdallas.edu

ABSTRACT

This paper describes a highly practical blind signal separation (BSS) scheme operating on subband domain data to blindly segregate convolutive mixtures of speech. The proposed method relies on spatio-temporal separation carried out in the time domain by using a multi-channel blind deconvolution (MBD) algorithm that enforces separation by entropy maximization through the popular natural gradient algorithm (NGA). Numerical experiments with binaural impulse responses affirm the validity and illustrate the practical appeal of the presented technique even for difficult speech separation setups.

***Index Terms*—**Subband filtering, blind source separation, multi-channel blind deconvolution, convolutive speech mixtures.

1. INTRODUCTION

Blind source separation (BSS) is a prominent statistical signal processing technique, which seeks to recover the individual contributions of a set of n unobserved but statistically independent (at each time instant t) physical sources $\mathbf{s}(t) = [s_1(t), \ldots, s_n(t)]^T \in \mathbb{R}^n$, while assuming little to almost no *a priori* knowledge about the source-to-sensor geometry or the source signals themselves. To isolate the original or "true" sources in the most practical scenario of *multipath propagation*, one needs to rely solely on information extracted from a set of m *linear* and *convolutive* mixtures of the original signal streams $\mathbf{x}(t) = [x_1(t), \ldots, x_m(t)]^T \in \mathbb{R}^m$ given by

$$\mathbf{x}(t) = \sum_{\ell=0}^{\infty} \mathbf{H}(\ell)\, \mathbf{s}(t-\ell), \quad t = 1, 2, \ldots \tag{1}$$

where $\mathbf{H}(\ell)$ represents an unknown but linear-time invariant (LTI) multiple-input multiple-output (MIMO) mixing system, which can accurately model the acoustic environment (or transmission channel) effects. Even for the most elaborate speech separation tasks, BSS can blindly achieve the recovery of the original sources $\mathbf{s}(t)$, by resorting only to the measurements observed at the sensor input, such that the system outputs $\mathbf{u}(t) = [u_1(t), \ldots, u_n(t)]^T \in \mathbb{R}^n$ read

$$\mathbf{u}(t) = \sum_{\ell=0}^{L-1} \mathbf{W}(\ell)\, \mathbf{x}(t-\ell), \quad t = 1, 2, \ldots \tag{2}$$

where $\mathbf{W}(\ell)$ is the unmixing matrix linking the jth source estimate $u_j(t)$ with the ith sensor observation $x_i(t)$, composed of sufficiently long finite impulse response (FIR) filters with each element given by vector $\boldsymbol{w}_{ji}(\ell) = [w_{ji}(0), w_{ji}(1), \ldots, w_{ji}(L-1)]$ for all coefficients $0 \leq \ell \leq L-1$ with $j = 1, 2, \ldots, n$ and $i = 1, 2, \ldots, m$.

Research supported in part by Grant R01-DC07527 from NIDCD/NIH.

Early BSS approaches focused on treating the problem entirely in the time domain. Still, by having to employ fairly long unmixing filters to reach an adequate level of separation, such techniques are inherently slow and computationally inefficient for long reverberation times. Lured by the potential of substantially reducing excessive computational requirements, many authors have recently suggested to carry out separation in the frequency [13, 15] or the subband domain [4, 5, 6, 12]. The premise is to make efficient use of the discrete Fourier transform (DFT) [4, 5, 13, 15], the discrete cosine transform (DCT) [6], and even the generalized DFT (GDFT) [12] for short-term stationary sources and transform costly convolution operations into straightforward multiplications. By doing so, the overall BSS task can be then elegantly reduced into several independent problems of instantaneous mixtures, one for each frequency subband. Although, moving to the frequency or subband domain is computationally fast, such strategies come with perils of their own, namely *scaling* and *permutation* ambiguities, which quite often have a negative effect on separation performance. Moreover, increasing the FFT blocksize to cater for longer paths can work at the expense of poor algorithm stability and convergence properties as reported in [2].

Recent literature has also seen some novel *multichannel blind deconvolution* (MBD) methods being widely applied in the problem of convolutive BSS [5, 7, 9, 10]. In stark contrast to purely based frequency domain BSS techniques, MBD methods that partially operate in the subband or z-domain are immune to any *permutation* disparities [10]. Furthermore, undesired *whitening* effects due to *scaling* indeterminacies being translated into unknown linear filtering operations, can be completely alleviated to fully retain the original source contributions at the system output [7, 9]. Nonetheless, by relying on blockwise FFT operations to speed up the adaptation of the unmixing weights, MBD approaches may also face performance limitations when longer estimation frames are employed [5].

The present contribution explores the concept of filter bank structures to perform BSS in the time domain by processing decimated convolutive mixtures of speech. Due to the decimation made possible by a reduced bandwidth in each subband, the proposed method can achieve a high computational efficiency equivalent to that of a frequency domain approach. The crucial difference to complex-valued schemes [4, 5, 12] is that here we use modulation only on single sidebands and hence operate on real-valued data, instead. This allows us to promptly adapt FIR filters *independently* for each subband with only a very small number of filter coefficients, and while doing so to still remain unaffected by scaling and permutation arbitrariness. Experimental results substantiate the strong potential of the proposed method even when *long* impulse responses are implicated.

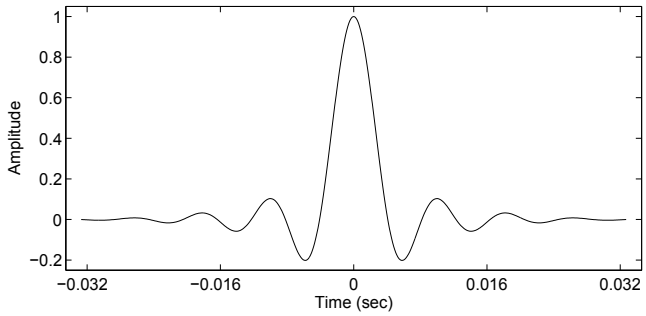

Fig. 1. Impulse response of the low-pass real-valued *prototype* FIR filter $f_0(n)$ with length $N = 512$ (sampled at 8 kHz) used to derive the *analysis* and the corresponding *synthesis* filter banks depicted in Fig. 3.

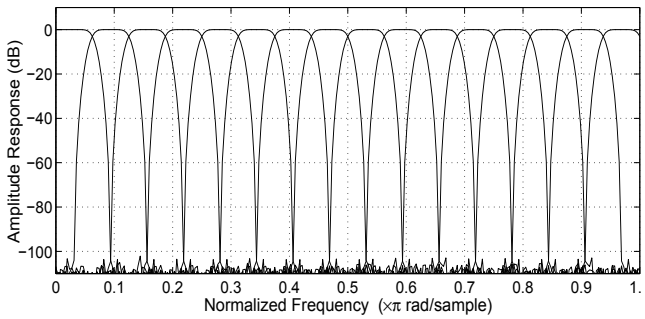

Fig. 2. Frequency response characteristics of the *analysis* filters for a uniform filter bank consisting of $K = 16$ channels. The stopband attenuation is equal to 110 dB.

2. SUBBAND-BASED MULTICHANNEL BLIND DECONVOLUTION

2.1. Subband Decomposition

By definition, subband decomposition encompasses the partitioning of input streams of the observed mixtures $x_i(t)$ $(i = 1, 2, \ldots, m)$ into a finite number of K subbands through filtering with a parallel bank of FIR band-pass filters, the so-called *analysis bank* given by $f_{K-1}(n)$. In practice, the subband analysis filter bank is efficiently implemented as the cosine modulated version of a prototype filter $f_0(n)$ of length N and cutoff frequency $\omega_c = \pi/K$ such that [11]

$$f_k(n) = f_0(n)\,\cos\left[\left(k - \frac{1}{2}\right)\frac{n\pi}{K}\right]. \tag{3}$$

In a similar manner, the filters comprising the so-called *synthesis bank* $g_{K-1}(n)$ used to reconstruct the original signals are given by

$$g_k(n) = g_0(n)\,\cos\left[\left(k - \frac{1}{2}\right)\frac{n\pi}{K}\right] \tag{4}$$

where the baseband synthesis filter $g_0(n)$ is actually a time-reversed copy of the analysis prototype filter $f_0(n)$ equal to

$$g_0(n) = f_0(N - n - 1) \tag{5}$$

with (3)–(5) defined for all $n = 1, 2, \ldots, N$ and $k = 0, 1, \ldots, K - 1$. An important requirement for the design of an ideal linear-phase low-pass prototype filter $f_0(n)$ is the near-perfect reconstruction (PR) property, which stipulates a magnitude response of the form

$$\|F_0(e^{j\omega})\| = \begin{cases} 1, & 0 \le |\omega| \le \omega_c \\ 0, & \omega_c < |\omega| \le \pi. \end{cases} \tag{6}$$

Such a filter would perfectly separate the subbands, as well as yield a flat composite magnitude response. In the time domain this could ultimately be made feasible by resorting to a $\mathrm{sinc}(\cdot)$ function of infinite length. As shown in Fig. 1, for the design of the prototype filter $f_0(n)$, we can employ a truncated $\mathrm{sinc}(\cdot)$ function weighted by a Hamming window $w(n)$ instead, such that

$$f_0(n) = \frac{\sin(n\pi/N)}{n\pi}\, w(n) \tag{7}$$

with $w(n) = 0.54 - 0.46\cos(2n\pi/N)$ and $n = 1, 2, \ldots, N$. The frequency responses of $K = 16$ subband analysis filters derived from cosine modulation of $f_0(n)$ according to (3) are depicted in Fig. 2.

After the subband filtering stage, the effective bandwidth of the decomposed sensor waveforms in each subband is reduced precisely by a factor of $1/K$ compared to the wider bandwidth of the original fullband signals. Therefore, the observed mixtures can be effectively downsampled by an integer decimation factor M such that $M \le K$. Overall, the subband analysis stage yields the real-valued signals

$$x_i(k, \tau) = \sum_{n=1}^{N} f_k(n)\, x_i(\tau - n) \tag{8}$$

with vector $x_i(\cdot)$ defined in (1) for $i = 1, 2, \ldots, m$ and band $k = 0, 1, \ldots K - 1$ where $\tau = rM$ now denotes the time index at the reduced sampling rate for some integer r. Performing critical downsampling for $M = K$ would result in the highest possible computational savings. Yet, to avoid any aliasing distortion effects between adjacent bins when recomposing, it is often common practice to *oversample* the signals by a factor $M < K$, instead (e.g., see [16]).

2.2. Time-Domain Subband Multichannel Blind Deconvolution

Since, in general the decimation factor is much smaller than the length of the mixing (and unmixing) filters, such that $M \ll L$, the signals in each subband are still considered convolutive mixtures of the original sources. In effect, by working under the assumption that spatial independence remains a plausible condition even after subband filtering, the original sources can be recovered *independently* on each subband. The added benefit is that because of decimation, the length of the FIR filters we need to estimate is just L/M. The *isomorphic* mapping between scalar and FIR polynomial matrices (e.g., see [10]), allows several adaptation rules based on the entropy maximization principle [3] to be extended and accommodate multipath effects. Such an efficient update rule is the linear prediction-based natural gradient algorithm (LP-NGA) proposed in [7, 9], which stems from the well-known NGA of [1]. When executed separately for each subband in the two-source and two-sensor convolutive BSS scenario, this reads

$$\underline{\mathbf{W}}_{\ell+1}^{(k)}(z) = \underline{\mathbf{W}}_{\ell}^{(k)}(z) + \mu\, \Delta\underline{\mathbf{W}}_{\ell}^{(k)}(z) \tag{9}$$

where $\underline{\mathbf{W}}(\cdot)$ is the unmixing FIR polynomial matrix, μ denotes the learning parameter and

$$\Delta\underline{\mathbf{W}}_{\ell}^{(k)}(z) = \left(\begin{bmatrix} \bar{1} & \bar{0} \\ \bar{0} & \bar{1} \end{bmatrix} - \mathrm{FFT}[\boldsymbol{\varphi}(\boldsymbol{u})]\, \boldsymbol{u}^H\right)^{(k)} \begin{pmatrix} W_{11}^{(k)} & W_{12}^{(k)} \\ W_{21}^{(k)} & W_{22}^{(k)} \end{pmatrix}. \tag{10}$$

The vectors $W_{ji}^{(k)}(\cdot)$ defined for $i, j = 1, 2$, represent the unmixing FIR filters at each subband $k = 1, 2, \ldots, K - 1$ in the z-domain, $(\cdot)^H$ is the Hermitian operator, the matrix composed of a sequence of all ones $(\bar{1})$ in the main diagonal and all zeros $(\bar{0})$ elsewhere (both

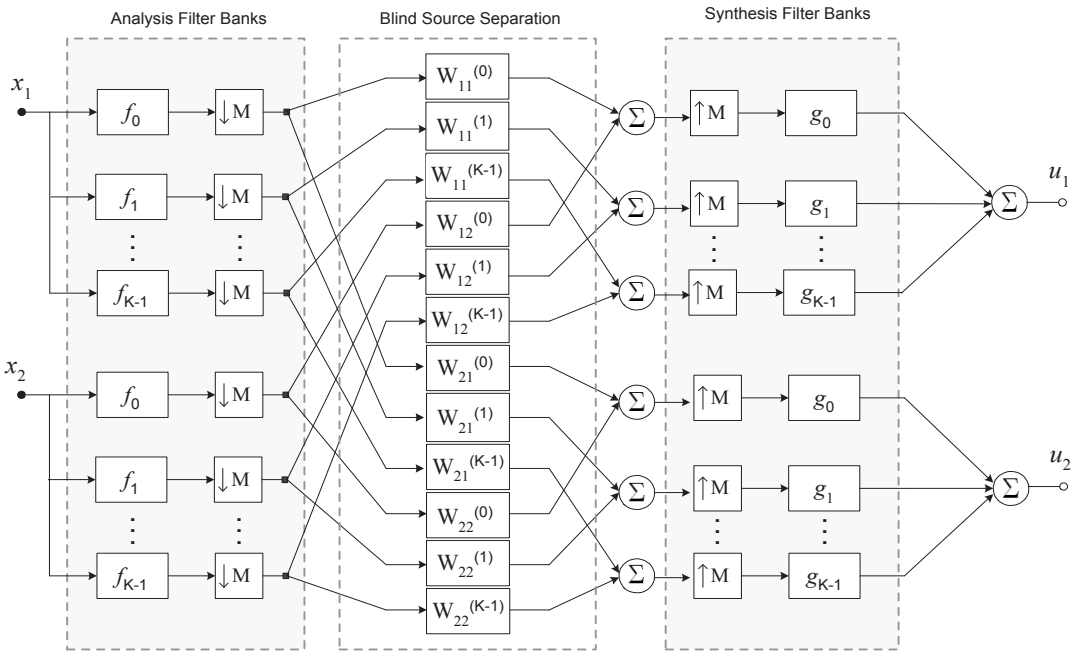

Fig. 3. Proposed subband MBD structure consisting of a subband *analysis*, a separation and a subband *synthesis* stage in the two-source and two-sensor case.

with length L) is the identity (unit) FIR polynomial matrix, whereas the term FFT $[\boldsymbol{\varphi}(\boldsymbol{u})]$ denotes the frequency domain vector of the score function vector $\boldsymbol{\varphi}(\boldsymbol{u})$, operating in the time domain, given by

$$\varphi_i(u_i) = -\frac{d}{du_i}\log p_{u_i}(u_i), \quad i = 1, 2. \tag{11}$$

Although, statistical analysis for the probability density functions (PDFs) of fullband speech samples indicates that a reasonably safe approximation is the Laplacian distribution, a substantially better fit can arise by using the generalized Gaussian distribution (GGD) family (e.g., see [8]) as the hypothesized subband source PDF in (11), which in turn yields the GGD-based score function [7, 9]

$$\varphi_i\big(u_i(k)\big) = \beta_i(k)\,\frac{\operatorname{sign}\big(u_i(k)\big)}{|u_i(k)|}\,|u_i(k)|^{\beta_i(k)} \tag{12}$$

where $\beta_i(k)$ represents the *shape* parameter of the GGD corresponding to each individual source subband taking values between $(0, 1]$. After convergence of (9), the unmixing filters will produce estimates of the original source signals on each subband, which in the 2×2 setting, after dropping the z-domain operator, can be written as

$$\begin{pmatrix} \boldsymbol{u}_1^{(k)} \\ \boldsymbol{u}_2^{(k)} \end{pmatrix} = \begin{pmatrix} W_{11}^{(k)} & W_{12}^{(k)} \\ W_{21}^{(k)} & W_{22}^{(k)} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1^{(k)} \\ \boldsymbol{x}_2^{(k)} \end{pmatrix} \tag{13}$$

for every band index $k = 0, 1, \ldots K-1$. The source estimates extracted at each subband, as shown in (13), are first moved back to the time domain, then upsampled by the interpolation factor M, next filtered by the synthesis filters in (4), and finally added together to ultimately form the fullband recovered signals in the system output

$$u_j(t) = \sum_{k=0}^{K-1}\sum_{n=1}^{N} \mathrm{g}_k(n)\, u_j(k, t-n) \tag{14}$$

for $j = 1, 2$ with t denoting the time index at the restored sampling rate, such that $t = r/M$. The proposed configuration and the order of operations for MBD in the subband level for the 2×2 case, are illustrated in Fig. 3.

3. EXPERIMENTAL RESULTS

To investigate the potential of the proposed subband-based MBD method for achieving speech separation in challenging convolutive environments, two speech signals are convolved with a set of *binaural* room impulse responses (BRIRs) (e.g., see [14]). These exhibit *rapid* variations both in phase and magnitude and are, in general, fairly difficult to invert with FIR filters. The signals used as sources are sentences of one male and one female speaker, 5 seconds in duration, recorded at a sampling rate of 8 kHz, and normalized so that their maximum amplitude is unity. The BRIRs are measured in a $5 \times 9 \times 3.5$ m ordinary classroom using the Knowles Electronic Manikin for Auditory Research (KEMAR), positioned at 1.5 m above the floor and at ear level [14]. By convolving the speech signals with the pre-measured impulse responses, one source is *virtually* placed directly at the front of the listener and the other at an angle of 60° in the azimuth to the right, while both are located at the realistically conversational distance of roughly 1.2 m away from the KEMAR. To further gauge the difficulty of this separation task, the broadband reverberation time of the room is estimated by relying on the time-reversed energy integration procedure to find the energy decay curve that ultimately reveals the time required to reach a 60 dB level of attenuation. As Fig. 4 indicates, for this particular enclosure, $T_R = 150$ ms. To assess the separation ability of the algorithm, we resort to the signal-to-interference-ratio improvement (SIRI) by measuring the *overall* amount of crosstalk reduction achieved by the algorithm *before* (SIR_i) and *after* (SIR_o) the unmixing stage, which in dB is equal to

$$\mathrm{SIRI} = 10\log\left(\sum_{i=1}^{m}\frac{\sum_{\ell=0}^{L-1}|u_{ii}|^2}{\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\ell=0}^{L-1}|u_{ij}|^2}\right) - 10\log\left(\sum_{i=1}^{m}\frac{\sum_{\ell=0}^{L-1}|x_{ii}|^2}{\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\ell=0}^{L-1}|x_{ij}|^2}\right) \tag{15}$$

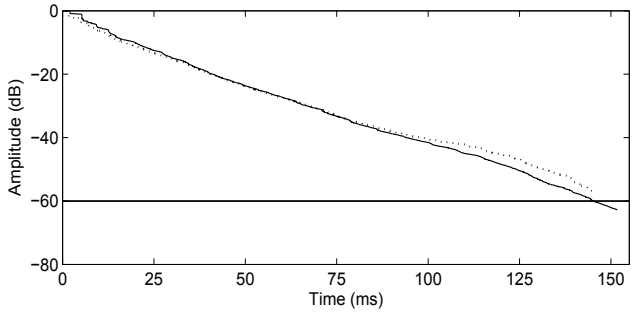

Fig. 4. Energy decay curves of the left-ear (solid line) and right-ear (dash line) impulse responses at 60°. The time taken for the amplitude to drop by 60 dB (thick line) below the original sound energy level is around 150 ms.

with $x_{11}(\cdot)$, $x_{22}(\cdot)$, $u_{11}(\cdot)$, $u_{22}(\cdot)$ representing the *direct-channel* and $x_{12}(\cdot)$, $x_{21}(\cdot)$, $u_{12}(\cdot)$, $u_{21}(\cdot)$ the *cross-channel* individual contributions of the orignal sources $s_1(\cdot)$ and $s_2(\cdot)$ for all $i, j = 1, 2$, realized by creating two set of mixtures and source estimates after assuming that only one of the sources becomes active at each one time. To compare, the standard fullband MBD update (e.g., see [7, 9]) and the new subband-based MBD update procedure shown here in (9)–(10) are put forward to separate the aforementioned convolutive speech mixtures. Filter lengths of $L = 1{,}024$ and $L = 2{,}048$ corresponding to impulse response lengths of 128 ms and 256 ms, respectively, are chosen for the fullband MBD. The original mixtures have an input SIR_i equal to –1.28 dB. Around 10 seconds of data are used for training, while the algorithm performs 20 passes through the data. The SIR_o values measured at the system output are depicted in Table 1, for both filter lengths. The same data are also tackled with the proposed subband domain MBD approach. In this case, the number of subbands is set to $K = 32$ and the downsampling factor is $M = 24$. The prototype filter to generate the subband analysis and synthesis filter banks is shown in Fig. 1. Despite the fairly long length of filter $f_0(n)$ $(N = 512)$, we can still afford to keep the separation filters $W_{ji}(\cdot)$ short. At any event, fast computations during the subband level processing (analysis and synthesis) and separation stages are guaranteed by performing filtering operations in the frequency domain, but reconstructing the final source estimates back in the time domain with the standard overlap-save algorithm. The original mixtures have an SIR_i value of –0.76 dB[1]. The algorithm operates with FIR filters of $L = 32$ and $L = 64$, which are equivalent to just a 4 ms and 8 ms delay, respectively. The amount of training and number of passes remain unchanged. We use the same score function in (12) $(\beta = 0.8)$ with μ tuned for maximum performance, while the overlapping between successive frames (or blocks) of data is set to 50 %. As revealed from the SIR_o values calculated after the subband MBD algorithm converges, the overall SIRI is equal to 8.87 dB for length $L = 32$ and 10.74 dB for $L = 64$, which are almost identical $(L = 1{,}024)$ or substantially higher $(L = 2{,}048)$ than the SIRI values obtained for the fullband MBD approach. The documented performance is indicative of the ability of our technique to adequately cancel out long delay paths, even with relatively short FIR filters while operating with a low computational complexity, overall reduced by a factor of $c = 2M^2/K$.

4. CONCLUSIONS

In this paper, we take on a new subband-based BSS scheme relying on an MBD method, which combines the natural gradient with the entropy maximization criterion to separate convolutive mixtures of speech in the time domain. Experiments in a challenging convolutive setup signify the novelty and potential of our approach by proving that subband MBD can *match* or even *outperform* fullband MBD in terms of performance at a highly reduced computational cost.

[1] SIR_i values in Table 1 corresponding to subband MBD are measured at a subband level and are thus different from the ones cited for fullband MBD.

BSS Method	Filter length	$\text{SIR}_i \mid \text{SIR}_o$
Fullband MBD	1,024	–1.28 | 7.82
	2,048	–1.28 | 5.23
Subband MBD	32	–0.76 | 8.11
	64	–0.76 | 9.98

Table 1. Separation performance for fullband and subband MBD methods.

5. REFERENCES

[1] S.-I. Amari, A. Cichocki and H. H. Yang, "A new learning algorithm for blind signal separation," In *Adv. Neural Informat. Process. Systems.* Cambridge, MA: MIT Press, 1996, Vol. 8, pp. 757–763.

[2] S. Araki, R. Mukai, S. Makino, T. Niskikawa and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech and Audio Process.*, Vol. 11, No. 2, pp. 109–116, Mar. 2003.

[3] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computat.*, Vol. 7, No. 6, pp. 1129–1159, Jul. 1995.

[4] F. D. Beaulieu and B. Champagne, "Fast convolutive blind speech separation via subband adaptation," In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Hong Kong, April 6–10, 2003, Vol. V, pp. 513–516.

[5] N. Grbić, X.-J. Tao, S. E. Nordholm, and I. Claesson, "Blind signal separation using overcomplete subband representation," *IEEE Trans. Speech and Audio Process.*, Vol. 9, No. 5, pp. 524–533, Jul. 2001.

[6] J. Huang, K.-C. Chen and Y. Zhao, "Subband-based adaptive decorrelation filtering for co-channel speech separation," *IEEE Trans. Speech and Audio Process.*, Vol. 8, No. 4, pp. 402–406, Jul. 2000.

[7] K. Kokkinakis and A. K. Nandi, "Optimal blind separation of convolutive audio mixtures without temporal constraints," In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Montréal, Canada, May 17–21, 2004, Vol. I, pp. 217–220.

[8] K. Kokkinakis and A. K. Nandi, "Exponent parameter estimation for generalized Gaussian probability density functions with application to speech modeling," *Signal Process.*, Vol. 85, No. 9, pp. 1852–1858, Sep. 2005.

[9] K. Kokkinakis and A. K. Nandi, "Multichannel blind deconvolution for source separation in convolutive mixtures of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, Vol. 14, No. 1, pp. 200–212, Jan. 2006.

[10] R. H. Lambert, *Multichannel Blind Deconvolution: FIR matrix algebra and separation of multipath mixtures*. Ph.D. Thesis, University of Southern California, Los Angeles, May 1996.

[11] K. Nayebi, T. P. Barnwell, III, and M. J. T. Smith, "Time-domain filter bank analysis: A new design theory," *IEEE Trans. Signal Process.*, Vol. 40, No. 6, pp. 1412–1429, Jun. 1992.

[12] H.-M. Park, S.-H Oh and S.-Y Lee, "A uniform oversampled filter bank approach to independent component analysis," In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Hong Kong, April 6–10, 2003, Vol. V, pp. 249–252.

[13] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Trans. Speech and Audio Process.*, Vol. 8, No. 3, pp. 320–327, May 2000.

[14] B. G. Shinn-Cunningham, N. Kopco and T. J. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Am.*, Vol. 117, No. 5, pp. 3100–3115, May 2005.

[15] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomp.*, Vol. 22, No. 1–3, pp. 21–34, Nov. 1998.

[16] P. P. Vaidyanathan, *Multirate Systems and Filter Banks.* Englewood Cliffs, New Jersey: Prentice-Hall, 1993.