BLIND SOURCE SEPARATION OF NONSTATIONARY CONVOLUTIVELY MIXED SIGNALS IN THE SUBBAND DOMAIN

Iain Russell [*], *Jiangtao Xi* [*], *Alfred Mertins* [**], *and Joe Chicharo* [*]

[*] School of Elec., Comp., and Tele. Eng., University of Wollongong, Wollongong, N.S.W. 2522, Australia, Email: {iainr,jiangtao,chicharo}@uow.edu.au

[**] Signal Processing Group, Institute of Physics, University of Oldenburg, 26111 Oldenburg, Germany, Email: alfred.mertins@uni-oldenburg.de

ABSTRACT

This paper proposes a new technique for blind source separation (BSS) in the subband domain using an extended lapped transform (ELT) decomposition for nonstationary, convolutively mixed signals. As identified in [1] the motivation for subband-based BSS is the drawback of frequency domain BSS when dealing with separating mixed speech signals over a few seconds resulting with few samples in individual frequency bins leading to poor separation performance. In the proposed approach mixed signals are decomposed into subband components by an ELT and within each subband a time domain Newton BSS algorithm is employed based on the nonstationarity property of the input signals and the joint diagonalization of output correlation matrices with time varying second order statistics (SOS). This subband version is compared to a fullband version using the same BSS algorithm.

1. INTRODUCTION

Blind source separation (BSS) is a problem that estimates unobserved source signals using only information contained in mixtures of these source signals. Neither signal sources nor the mixing system are known *a priori*. With the advent of more powerful DSP chipsets, BSS has found useful purpose in speech enhancement applications including speech recognition, hearing aids and hands free telephony.

Where signal sources are speech and the observed signals are the mixture of those sources in a reverberant environment, to estimate the underlying sources from the resulting observed signals one needs to estimate unmixing FIR filters of several thousand taps. A commonly used approach to solve such problems with a high number of dimensions is to transform the problem to the frequency domain using a discrete Fourier block transform [2, 3]. However the problem with this is that when a long frame is used to estimate a long unmixing filter to cover realistic reverberation, after transforming to the frequency domain the number of samples in each frequency bin becomes small, and separation performance is degraded [1]. In addition, with simple blockbased DFTs the outputs of the system may exhibit click artifacts at block boundaries. This may be alleviated by using overlapping blocks with the subband synthesis filters based on lapped transform basis functions [4].

By performing subband decomposition on the mixed signals before applying the time domain BSS algorithm, we are reducing a problem that in the fullband has a high number of parameters, to a set of BSS problems in the respective subbands with fewer unknowns in each subband. We use a uniform FIR filterbank and also utilize oversampling to avoid aliasing influence caused by separation processing within subbands.

The subband-based BSS approach, using the fast Newton-type algorithm from [5] within each subband, is compared with the fullband-based BSS approach. The algorithm in [5] is a modification of the method in [6] and is applicable to convolutive mixing. In this paper, we restrict ourselves to real-valued signals and systems. With all closed form expressions of first and second order information, this fast method converges much better than the otherwise used gradient-type methods.

The paper is organized as follows. Section 2 gives a brief description of modelling BSS in a convolutive environment and also defines the basic BSS algorithm used to perform separation in the various subbands. In Section 3 the general framework for subband decomposition is investigated with a direct form implementation. Factors such as the design of analysis and synthesis stages of the filterbank as well as the oversampling factor will be discussed. Section 4 looks at the integration of the time domain Newton BSS algorithm from [5] performed within all respective subbands of the mixed signals. Section 5 provides the simulation results focusing on a comparison between the convergence behavior in the fullband of the BSS method and the proposed subband method in this paper. Real filter mixing responses are used and these are measured from a typical reverberant office environment with speech segments taken from the TIMIT corpus of speech used as the input sources. Finally, a conclusion is provided in Section 6.

The following notations are used in this paper. Vectors and matrices are printed in boldface. Matrix and vector transpose are denoted by $(\cdot)^T$. $E(\cdot)$ means the expectation operation, and vec$(\cdot)$ stacks the columns of a matrix to a column vector. $\|\cdot\|_F$ is the Frobenius norm of a matrix. With $\mathbf{a} = diag(\mathbf{A})$ we obtain a vector whose elements are the diagonal elements of $\mathbf{A}$ and $diag(\mathbf{a})$ is a square diagonal matrix which contains the elements of $\mathbf{a}$. $ddiag(\mathbf{A})$ is a diagonal matrix where its diagonal elements are the same as the diagonal elements of $\mathbf{A}$ and $off(\mathbf{A}) \triangleq \mathbf{A} - ddiag(\mathbf{A})$. $\mathbf{1}_{N\times N}$ is an $N \times N$ matrix of ones and $\mathbf{I}_N$ is the $N \times N$ identity matrix.

2. CONVOLUTIVE BSS IN THE TIME DOMAIN

Suppose we have N discrete time sources

$$\mathbf{s}(t) = [s_1(t), ..., s_N(t)]^T, \tag{1}$$

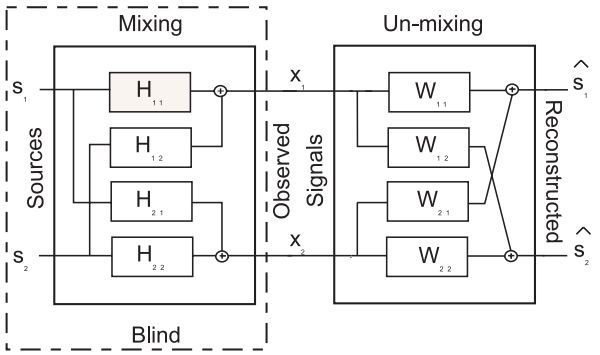

Fig. 1. Fullband TITO BSS system

where we assume that the individual sources are independent of each other. These sources are mixed in a reverberant environment using a convolutive model providing M sensor or observed signals:

$$\mathbf{x}(t) = [x_1(t), ..., x_M(t)]^T. \tag{2}$$

If $\mathbf{H}(t)$ is an $M \times N$ mixing matrix with its element $\mathbf{h}_{ij}(t)$ being the impulse response from jth source signal to ith measurement then assuming the multiple input multiple output (MIMO) mixing channels can be modelled as FIR filters with length P, the mixed signals can be written as:

$$\mathbf{x}(t) = \sum_{\tau=0}^{P-1} \mathbf{H}(\tau)\mathbf{s}(t-\tau). \tag{3}$$

The M observed signals $\mathbf{x}(t)$ are coupled to the N reconstructed signals $\hat{\mathbf{s}}(t)$ via the demixing system. The demixing system has a similar structure to the mixing system. It contains $N \times M$ FIR filters of length Q,where $Q \geq P$. The demixing system can also be expressed as an $N \times M$ matrix $\mathbf{W}(t)$, with its element $\mathbf{w}_{ij}(t)$ being the impulse response from jth measurement to ith output. The reconstructed signal can be obtained as:

$$\hat{\mathbf{s}}(t) = \sum_{\tau=0}^{Q-1} \mathbf{W}(\tau)\mathbf{x}(t-\tau) \tag{4}$$

where $\hat{\mathbf{s}}(t) = [\hat{s}_1(t), ..., \hat{s}_N(t)]^T$. For the case of $N = M = 2$ (the two input two output (TITO) case) the mixing and unmixing system are shown in Fig. 1.

Equation (4) can be written as the following matrix form

$$\hat{\mathbf{s}}(n) = \mathcal{W}\mathcal{X}(n) \tag{5}$$

where $\mathcal{W}$ is a $(N \times QM)$ matrix given by

$$\mathcal{W} = [\mathbf{W}(0), \mathbf{W}(1), ..., \mathbf{W}(Q-1)] \tag{6}$$

and $\mathcal{X}(n)$ is a $(QM \times 1)$ vector defined as

$$\mathcal{X}(n) = \begin{bmatrix} \mathbf{x}(n) \\ \mathbf{x}(n-1) \\ \vdots \\ \mathbf{x}(n-(Q-1)) \end{bmatrix}. \tag{7}$$

Correlation matrices for the recovered sources, at time frame k, for all necessary time lags τ can be obtained as

$$\begin{aligned} \mathbf{R}_{\hat{s}\hat{s},k}(\tau) &= \mathcal{W}E\{\mathcal{X}(k)\mathcal{X}^T(k+\tau)\}\mathcal{W}^T \\ &= \mathcal{W}\mathbf{R}_{\mathcal{X}\mathcal{X},k}(\tau)\mathcal{W}^T. \end{aligned} \tag{8}$$

Considering the correlation matrices with all different time lags we obtain the following cost function:

$$\mathcal{J}_1 \triangleq \sum_{\tau=-\tau_{min}}^{\tau_{max}} \sum_{k=1}^{K} \beta_k \|off(\mathcal{W}\mathbf{R}_{\mathcal{X}\mathcal{X},k}(\tau)\mathcal{W}^T)\|_F^2. \tag{9}$$

The values β_k are positive weighting *normalization* factors defined as:

$$\beta_k = (\sum_{\tau=-\tau_{min}}^{\tau_{max}} \sum_{k=1}^{K} \|\mathbf{R}_{\mathcal{X}\mathcal{X},k}(\tau)\|_F^2)^{-1}. \tag{10}$$

Each value of k represents a different time window frame where the SOS are considered stationary over that particular time frame. In adjacent non-overlapping time frames k and $k+1$, the SOS are changing due to the nonstationarity assumption. To solve for the unknown demixing system we must solve a nonlinear constrained optimization problem with NQM unknowns:

$$\begin{aligned} \mathcal{W}_{opt} \; = \quad & \arg\min_{\mathcal{W}} \; \mathcal{J}_1(\mathcal{W}) \\ & s/t \;\; \|ddiag(\mathcal{W}\mathcal{W}^T - \mathbf{I}_N)\|_F^2 = 0. \end{aligned} \tag{11}$$

Here we use a constraint implemented as a penalty term in the cost function with a fixed α to prevent the trivial solution and we employ the Newton method using the closed form analytical first and second order expressions given in Table 1. $\mathbf{R}_{\mathcal{X}\mathcal{X},k}(\tau)$ is denoted as $\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}$. The Newton method from [5] using weighted penalty terms is summarized in Table 2. The closed form expression $\mathbf{H}_2$ is a new addition to [5]. The matrices $\mathbf{P}_{off}$, $\mathbf{P}_{diag}$, and $\mathbf{P}_{vec}^{(N,L)}$ in Table 1 are mainly defined in accordance with [6]. $\mathbf{P}_{off}$ and $\mathbf{P}_{diag}$ are given by

$$\mathbf{P}_{off} = \text{diag}(\text{vec}(\text{off}(\mathbf{1}_{N\times N}))),$$

$$\mathbf{P}_{diag} = \text{diag}(\text{vec}(\mathbf{I}_N)).$$

The matrix $\mathbf{P}_{vec}^{(N,L)}$ is the permutation matrix defined by

$$\mathbf{P}_{vec}^{(N,L)} \, \text{vec}(\mathbf{W}^T) = \text{vec}(\mathbf{W})$$

for $N \times L$ matrices $\mathbf{W}$. Note that for $N \neq L$ the matrix $\mathbf{P}_{vec}^{(N,L)}$ is, in general, not self-inverse like the one that occurs is [6].

3. SUBBAND MODEL

There are three stages to the model. A subband analysis stage consisting of a uniform bank of FIR filters, the subband processing stage which performs the separation in the respective subband, and a synthesis stage which is used to reconstruct the separated subband signals back into their fullband versions. An $\bar{M}$ channel uniform oversampled analysis ELT filter bank is employed for decomposition of the M observed mixed signals into $\bar{M}$ subbands. An ELT is a filter bank in which the impulse responses of the synthesis filters are the ELT basis functions, and the impulse responses

Table 1. Closed form analytical expressions for the gradient and Hessian of the cost function and constraints.

Cost function - $\mathcal{J}_{\mathcal{W}}$
$\mathcal{J}_{\mathcal{W}} \triangleq \sum_{\tau=-\tau_{min}}^{\tau_{max}} \sum_{k=1}^{K} \|off(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}\mathcal{W}^T)\|_F^2$
Gradient - $\mathbf{G}_{\mathcal{W}}$
$\mathbf{G}_{\mathcal{W}} = 2\sum_{\tau=-\tau_{min}}^{\tau_{max}} \sum_{k=1}^{K} \{off(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}\mathcal{W}^T)\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T + off(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T\mathcal{W}^T)\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}\}$
Hessian - $\mathbf{H}_{\mathcal{W}}$
$\mathbf{H}_{\mathcal{W}} = 2\sum_{\tau=-\tau_{min}}^{\tau_{max}} \sum_{k=1}^{K} \{(\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k} \otimes off(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}\mathcal{W}^T)) + (\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T \otimes off(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T\mathcal{W}^T)) + (\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}_{off}(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k} \otimes \mathbf{I}_N) + (\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}_{off}(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T \otimes \mathbf{I}_N) + (\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}^{(N,N)}_{vec}\mathbf{P}_{off}(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k} \otimes \mathbf{I}_N) + (\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}_{off}\mathbf{P}^{(N,N)}_{vec}(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T \otimes \mathbf{I}_N)\}$
Row-normalized Constraint
$\mathcal{J}_2 = \|ddiag(\mathcal{W}\mathcal{W}^T - \mathbf{I}_N)\|_F^2$
Constraint Gradient
$\mathbf{G}_2 = 4ddiag(\mathcal{W}\mathcal{W}^T - \mathbf{I}_N)\mathcal{W}$
Constraint Hessian
$\mathbf{H}_2 = 4\,(\mathbf{I}_{MQ} \otimes ddiag(\mathcal{W}\mathcal{W}^T - \mathbf{I}_N)) + 4\,(\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}_{diag}(\mathcal{W} \otimes \mathbf{I}_N) + 2\,\mathbf{P}^{(N,MQ)}_{vec}(\mathbf{I}_N \otimes \mathcal{W}^T)\mathbf{P}_{diag}(\mathcal{W} \otimes \mathbf{I}_N) + 2\,(\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}_{diag}(\mathbf{I}_N \otimes \mathcal{W})[\mathbf{P}^{(N,MQ)}_{vec}]^T$

of the analysis filters are the time-reversed basis functions. A subsampling factor of $R = \bar{M}/4$ was used. The purpose of oversampling as opposed to critical sampling is to reduce the aliasing effects introduced via subband processing with the BSS algorithm described in Section 2.

The direct form uniform FIR filterbank provides a general and flexible framework to use. This flexibility comes with how we choose to design the impulse responses of the analysis and synthesis filters. Obviously, perfect reconstruction (PR) and minimal aliasing are fundamental criteria. The traditional spectral decomposition used for convolutive BSS is designing the impulse responses of the synthesis FIR filters to be the basis functions that define the DFT. Being a block transform the length L of each of the analysis and synthesis filters is equal to the number of subbands $\bar{M}$. The benefit of using an ELT to design the filter bank is that not only do we alleviate the "boundary problems" [4] associated with block transforms, the length of the FIR filters is not restricted thus allowing a better frequency selectivity for each subband filter. The impulse responses of the synthesis FIR filters $f_k(n)$ based on the ELT are defined by using the cosine modulation function:

$$f_k(n) = h(n)\sqrt{\frac{2}{\bar{M}}}\cos\left[(n + \frac{\bar{M}+1}{2})(k+\frac{1}{2})\frac{\pi}{\bar{M}}\right] \qquad (12)$$

Table 2. Newton-type algorithm for the joint-diagonalization task with a weighted constraint. The operator $\text{mat}_{N,MQ}(\mathbf{x})$ reshapes a vector $\mathbf{x}$ of length NMQ to an $N \times MQ$ matrix, where the vector elements are entered column-wise into the matrix.

Initialization $(r = 0) : \mathcal{W}_0$

For $r = 1, 2, ...$

$\mathbf{w}_r = \mu(\mathbf{H}_{\mathcal{W}} + \alpha\mathbf{H}_2)^{-1}\text{vec}(\mathbf{G}_{\mathcal{W}} + \alpha\mathbf{G}_2)$
$\triangle\mathcal{W}_r = \text{mat}_{N,MQ}(\mathbf{w}_r)$
$\mathcal{W}_{r+1} = \mathcal{W}_r - \triangle\mathcal{W}_r$

where $k = 0, 1, ..., \bar{M}-1$, and $n = 0, 1, ..., L-1$. For PR of the filterbank a scalar of $\sqrt{\frac{R}{\bar{M}}}$ must be multiplied with each $f_k(n)$. For our ELT, $L = 4\bar{M}$. The impulse responses of the analysis filters $h_k(n)$ are simply the time-reversed versions. The prototype filter $h(n)$ is defined as:

$$h(n) = -\frac{1}{2\sqrt{2}} + \frac{1}{2}\cos\left[(n+\frac{1}{2})\frac{\pi}{2\bar{M}}\right] \qquad (13)$$

4. TIME DOMAIN BSS ALGORITHM IN THE SUBBAND DOMAIN

After decomposing the mixed signals $\mathbf{x}(t)$ into $\bar{M}$ subbands, we get the subband signals $\mathbf{X}_{ELT}(k,m)$ where m is the time index and $k = 0, 1, ...\bar{M}-1$ is the subband index. The BSS algorithm described in Section 2 can be used on each of the corresponding subbands of the M mixed signals. As opposed to trying to solve an optimization problem for the fullband unmixing system where there are NMQ free parameters, shorter FIR filter systems can be solved for each subband where the number of parameters NMQ/R is smaller due to the down sampling factor R. To integrate the algorithm described in Section 2 we simply substitute the subband versions of the mixed signals $\mathbf{X}_{ELT}(k,m)$ and the unknown system $\mathcal{W}_{ELT}(k,m)$, for the fullband versions of the mixed signals $\mathbf{x}(t)$ and unknown system $\mathcal{W}$, and solve the separation problem for $k = 0, 1, ..., \bar{M}-1$. This will provide the respective unmixed signals for each subband $\hat{\mathbf{S}}_{ELT}(k,m)$. These unmixed signals in each subband can then be passed through the synthesis stage for reconstruction of the fullband version of the unmixed recovered signals $\hat{\mathbf{s}}(t)$.

5. SIMULATION RESULTS

To obtain the mixing system of a real reverberant office room, two loudspeakers and two microphones were set up. A maximum length sequence (MLS) of pseudo-random numbers was produced through each speaker and recorded through each microphone simultaneously. The cross-correlation of the generated and observed signals produces the required impulse responses of the mixing channels. The impulse responses had a reverberation time of $T_R = 200ms$ corresponding to $P = 1600$ FIR filter coefficients. Solving the Wiener-Hopf equations and using optimal filtering theory an unmixing system with $Q = 2048 > P$ was found. The

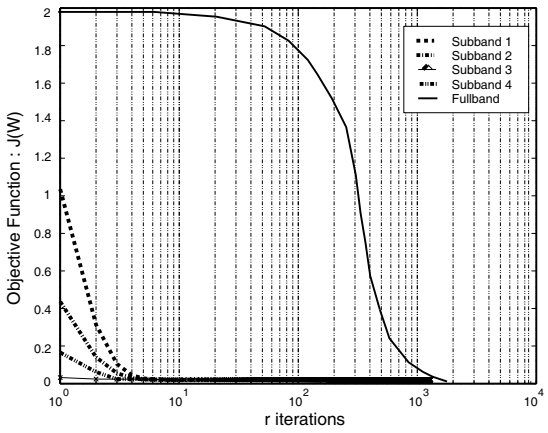

Fig. 2. Comparison of convergence graphs for subbands vs. fullband BSS algorithm.

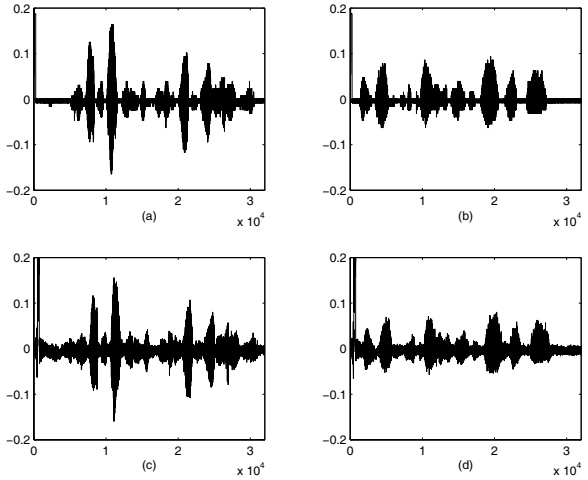

Fig. 3. (a) and (b) are the two original signals, (c) and (d) are the separated results after synthesis of subband solutions.

reverberation time of these unmixing responses is $T_R = 250ms$. In the fullband BSS problem, we would need to solve a nonlinear constrained optimization problem with 8196 unknown variables. By decomposing the unknown system $\mathcal{W}$ into $\bar{M} = 64$ subbands with a subsampling factor $R = 16$ we have in each subband 512 variables to solve. Time for convergence is relatively long in the fullband however by effectively solving more problems with fewer dimensions, the problem can be solved with a lower convergence time. Two four-second segments of 8kHz speech taken from the TIMIT corpus of speech were mixed together using the mixing system of the office room to produce two mixed signals. These fullband mixed signals were passed through the analysis bank to provide the $\mathbf{X}_{ELT}(k,m)$ subband mixed signals. The Newton time domain BSS for convolutive mixtures algorithm was used to solve for the subband unknown demixing system. Initial values of each subband unknown demixing system were randomly generated by adding Gaussian random variables with standard deviation $\sigma = 0.1$ to the coefficients of the true subband system $\mathcal{W}_{ELT_{ideal}}(k,m)$. This is derived by passing the fullband unmixing system responses through the analysis bank. In realistic scenarios where the true system is unknown, but some prior information on the location of sources is available, beamforming techniques can be utilized for initialization and prevention of the permutation problem, similar to the method in [1]. The weighting factor for the penalty term through empirical analysis was set to $\alpha = 0.2$ while the learning coefficient was set to $\mu = 0.8$ and the number of time frames was set to $K = 128$ thus taking into account the quasi-stationary nature of speech over approximately $20 - 30$ ms. If the scalar weighting factor is chosen too small then the trivial solution is not prevented where a value too large would lead to a non-optimal solution. Fig. 2 shows the convergence of the objective function of the first four subbands in comparison with the objective function in the fullband version. The summation of iterations till convergence over all subbands is less than the iteration time for convergence in the fullband version. Iterations r is defined as passing through the entire set of data. Fig. 3 shows good reconstruction of TIMIT speech signals up to a global permutation after synthesizing the separated mixed signals for all subbands $\bar{M}$.

6. CONCLUSION

This paper has provided a new method of solving a convolutively mixed nonstationary BSS problem with real data in a reverberant environment by employing subband decomposition based on an ELT FIR filterbank. Computational overhead is reduced by solving many problems in the subband domain with fewer dimensions as opposed to one problem in the fullband domain.

7. REFERENCES

[1] S. Araki, S. Makino, R. Aichner, T. Nishikawa, and H. Sarawatari, "Subband based blind source separation with appropriate processing for each frequency band," in *Proc. 4th Int. Sym. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Kobe, Japan, April 2003, pp. 499–504.

[2] K. Pope and R. Bogner, "Blind signal separation II: Linear, convolutive combinations," *Digital Signal Processing*, vol. 6, no. 1, pp. 17–28, Jan. 1996.

[3] K. Pope and R. Bogner, "Blind signal separation I: Linear, instantaneous combinations," *Digital Signal Processing*, vol. 6, no. 1, pp. 5–16, Jan. 1996.

[4] H. S. Malvar, *Signal Processing with Lapped Tranforms*, Artech House, Norwood, MA, 1992.

[5] I. Russell, A. Mertins, and J. Xi, "Time domain optimization techniques for blind separation of non-stationary convolutive mixed signals," in *Proc. 5th IASTED Int. Conf. on Signal and Image Processing*, Honolulu, Hawaii, USA, Aug. 2003, pp. 440–445.

[6] M. Joho and K. Rahbar, "Joint diagonalization of correlation matrices by using Newton methods with application to blind signal separation," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings*, Rosslyn, VA, USA, Aug. 2002, pp. 403–407.