4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

BLIND SEPARATION OF ACOUSTIC MIXTURES BASED ON LINEAR PREDICTION ANALYSIS

Kostas Kokkinakis, Vicente Zarzoso and Asoke K. Nandi

Signal Processing and Communications Group, Department of Electrical Engineering and Electronics,
The University of Liverpool, Brownlow Hill, Liverpool, L69 3GJ, U.K.
e-mail:{kokkinak,vicente,a.nandi}@liv.ac.uk

ABSTRACT

In this paper we propose a general method for separating mixtures of multiple audio signals observed in a real acoustic environment. The multipath nature of acoustic propagation is addressed by the use of the FIR polynomial matrix algebra, while spatio-temporal separation is achieved by entropy maximization using the natural gradient algorithm. The undesired temporal whiteness of the estimates is overcome with the use of linear prediction (LP) analysis. As opposed to a previous LP-based method, no assumptions on relative strengths of individual sources to specific mixtures are made. Other benefits such as reduced computational complexity and increased convergence speed are also emphasized. Finally, a number of experiments demonstrate the validity and general applicability of the proposed method.

1. INTRODUCTION

Consider the scenario in which multiple sounds emitted from a number of different acoustic sources are perceived by us humans. Despite the fact that all sounds arrive as a single waveform, it is possible to clearly distinguish between them and recognize those of particular interest. This phenomenon, referred to as the cocktail-party effect [3], demonstrates the ability of humans to focus on sounds of interest even in the presence of many competing sounds. In the context of a digital system, the most promising approach to mimic this human skill is blind source separation. BSS first began to receive attention in the early work of [7], where a neuromimetic structure was proposed to extract the independent components that generate a set of observable data.

In this work, we address the blind separation of simultaneous audio sources recorded in a reverberant environment. In this case, each microphone captures a direct copy of the sound sources as well as several reflected and modified copies of the signals at totally different propagation delays. The recorded signals not only suffer from attenuation, caused by loss of energy as the sound propagates, but also from multipath propagation effects due to multiple sound reflections. These delays depend on the relative locations of the sensors and the sound sources and on the speed of the signal [8]. Assuming now that the signals are combined linearly, each microphone captures a weighted sum of time-delayed versions of the original sources, which is a convolution of the original source with the appropriate acoustic transfer function [13].

K. Kokkinakis is supported by the EPSRC, U.K. and V. Zarzoso through a Postdoctoral Research Fellowship awarded by the Royal Academy of Engineering of the U.K.

In [14], the important distinction between multichannel blind deconvolution (MBD) and convolutive BSS methods is stressed. The former aim to render the outputs both spatially and temporally independent making them unsuitable for the blind separation of acoustic mixtures. The latter yield spatially separated estimates only, and are thus better suited to acoustic signal separation. Still, due to their many advantages, the MBD methods have been widely used in the problem of convolutive BSS. In [15]–[16], a feedback network structure with FIR filters is combined with MBD to separate delayed and convolutive mixtures by the information maximization approach. The desired solution is reached under the assumption that the mixing filters are minimum-phase systems, an assumption which may not always hold in practice. More recently in [6], a solution has been given to the problem of single and multichannel blind deconvolution in the case of i.i.d. source signals with the application of the natural gradient algorithm and the minimization of the mutual information criterion. In [10]–[12], the natural gradient algorithm is combined with the FIR polynomial algebra and experiments with real room recordings demonstrate the separation of two speech signals.

While the aforementioned approaches seem to yield good separation results, in the case of self-correlated inputs — speech being a prime example — they exhibit the side effect of whitening. Whitening is defined as the effect of flattening of the signal power spectrum, causing energy at higher frequencies to increase at the expense of energy in the lower frequency bands, generating estimates of a significantly impaired quality. To overcome this problem, [14] suggests a cascaded separating system consisting of separating and linear prediction (LP) filters. Operating on the assumption that each source and hence its spectral characteristics is dominant over the rest in each mixture, it uses a natural gradient algorithm for directly updating the separating filter matrix, while preserving the source colour at its output. However, in realistic recording scenarios where this assumption cannot be guaranteed, the LP filters could introduce reconstruction artifacts. This distortion is mainly due to the fact that the LP coefficients are estimated from a sum of speakers (speech mixtures), instead of from each isolated source and are then applied on the recovered outputs.

In this paper we present a novel method to eliminate the undesired whitening effect in the extracted estimates without availing ourselves of the above assumption. The method is based on the temporal prewhitening of the acoustic mixtures via the use of LP analysis filters. The extracted temporally independent signals are then used to perform BSS in the residual domain. Since BSS filters only perform spatial separation, their application to the sensor output is expected to yield spatially but not temporally independent estimates, thus avoiding the undesired effect of spectral flattening.

2. PROBLEM STATEMENT

Given m measured signals $\mathbf{x}(t) = [x_1(t), \ldots, x_m(t)]^T$ being mixtures of n source signals $\mathbf{s}(t) = [s_1(t), \ldots, s_n(t)]^T$, the aim of blind separation is to produce outputs that recreate the original source signals. Note the term 'blind' that stresses the fact that nothing is known about the source signals themselves nor about their mixing structure. The only assumption made, which in most cases is realistic, is that the source signals are statistically independent. This, also known as the source spatial whiteness or spatial independence assumption, is the core assumption of blind separation. It is also generally assumed that all the signals are zero–mean and that the number of sensors is equal to the number of sources i.e $m = n$. Thus, for $\mathbf{s}(t) \in \mathbb{R}^n$ and $\mathbf{x}(t) \in \mathbb{R}^m$, the ith sensor signal $x_i(t)$ is given by the noiseless linear convolutive mixing model:

$$x_i(t) = \sum_{j=1}^{n} \sum_{k=0}^{l-1} h_{ij}(k)\, s_j(t-k), \quad i = 1, 2, \ldots, m. \tag{1}$$

where t is the discrete-time index, $\{h_{ij}(k)\}$ is the room impulse response characterizing the path from source j to sensor i, and $l-1$ defines the order of the FIR filters used to model the room effects. The z–transform $H_{ij}(z)$ of the acoustic transfer function between the jth source and the ith sensor can be written as:

$$H_{ij}(z) = \sum_{t=0}^{l-1} h_{ij}(t)\, z^{-t} \tag{2}$$

where z^{-t} is the time–shift (delay) operator, yielding the BSS model in the z–domain given by:

$$X_i(z) = \sum_{j=1}^{n} H_{ij}(z)\, S_j(z), \quad i = 1, 2, \ldots, m. \tag{3}$$

with the convolution operation becoming a simple multiplication. The unmixing model can also be rewritten in the z–domain as:

$$U_i(z) = \sum_{j=1}^{n} W_{ij}(z)\, X_j(z), \quad i = 1, 2, \ldots, m. \tag{4}$$

where $W_{ij}(z)$ represents the unmixing or separation system. The design of $W_{ij}(z)$ must allow for noncausal expansion, since the stable inverse filter of a non-minimum phase system–such as the acoustic transfer function–is noncausal.

3. APPROACHES BASED ON NATURAL GRADIENT ALGORITHM

3.1. Instantaneous Mixtures

In [2], the BSS problem is put into an information theoretic framework. This method, also corroborated by [4], uses a feedforward neural network structure to blindly separate the linear and instantaneous mixtures $\mathbf{x} = [x_1, \ldots, x_m]^T$ of the independent sources $\mathbf{s} = [s_1, \ldots, s_n]^T$ using the principle of information maximization. The observations (mixtures) are transformed and processed through a nonlinear function $\mathrm{g}(\cdot)$ that approximates the cumulative density function (cdf) of the sources. The basic idea is that by maximizing the joint signal entropy defined by $H(\mathbf{y})$ it is possible to minimize the mutual information between the outputs of the network. Separation is achieved when the mutual information $I(y_1, y_n) = 0$ making the nonlinear outputs $\mathbf{y} = [y_1, \ldots, y_n]^T = \mathrm{g}(\mathbf{u})$ statistically independent. For maximization of $H(\mathbf{y})$, a blind adaptive algorithm is used to estimate the separation matrix $\mathbf{W}$ so that $\mathbf{u} = [u_1, \ldots, u_n]^T = \mathbf{W}\,\mathbf{x}$. The separation matrix $\mathbf{W}$ can be proportionally updated to its natural gradient yielding the natural gradient algorithm [1]:

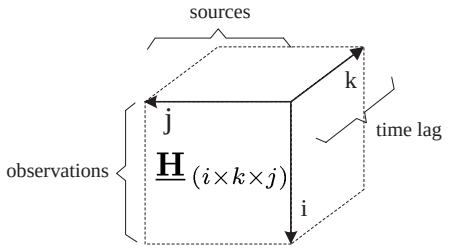

Fig. 1. Geometric representation of the FIR filter matrix.

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \mu \left(\mathbf{I} - \mathrm{g}(\mathbf{u})\, \mathbf{u}^T\right) \mathbf{W}_k \tag{5}$$

where μ is the step size and $\mathbf{I}$ is the identity matrix. The algorithm in (5) avoids matrix inversion and thus proves to be more computationally efficient than the entropy gradient method proposed in [2].

3.2. Convolutive Mixtures

Driven by the problem of representing the multipath effects introduced in the case of convolutive mixtures, [9] extends the instantaneous BSS problem by replacing the scalars in the mixing matrices by FIR filter polynomials. In the FIR polynomial matrix algebra the traditional matrix and vector algebra rules and tools remain intact, while the mixing system elements are represented by FIR filters. The implication is that single channel adaptive filter techniques can easily be extended to the multichannel blind source separation case. Hence, we may define an $\underline{\mathbf{H}}_{(i\times k\times j)}$ matrix of FIR filters in the time–domain, where each element of the matrix is an FIR filter defined as $\mathbf{h}_{ij} = [h_{ij}(0),\, h_{ij}(1),\, \ldots\, , h_{ij}(k)]$. Note the indices, $i = [1, 2, \ldots, m]$, $j = [1, 2, \ldots, n]$ and $k = [0, 1, \ldots, l-1]$, corresponding to the observations, sources and to each filter tap, respectively. For $i, j = 1, 2$, i.e., the two–source and two–sensor mixing case, the FIR matrix in the time–domain may be written as in:

$$\underline{\mathbf{H}}_{(i\times k\times j)} = \begin{bmatrix} \mathbf{h}_{11} & \mathbf{h}_{12} \\ \mathbf{h}_{21} & \mathbf{h}_{22} \end{bmatrix}, \quad i, j = 1, 2. \tag{6}$$

In a similar manner, moving to the frequency domain, we may define the FIR polynomial matrix as:

$$\underline{\mathbf{H}}_{(i\times k\times j)} = \begin{bmatrix} \sum_{\xi=0}^{k} h_{11}(\xi) z^{-\xi} & \sum_{\xi=0}^{k} h_{12}(\xi) z^{-\xi} \\ \sum_{\xi=0}^{k} h_{21}(\xi) z^{-\xi} & \sum_{\xi=0}^{k} h_{22}(\xi) z^{-\xi} \end{bmatrix} \tag{7}$$

where each FIR polynomial is defined as $\sum_{\xi=0}^{k} h_{ij} z^{-\xi}$ for $i, j =$

$1, 2$ i.e a moving–average (MA) process. Any function $f(\cdot)$ acting on an FIR filter with an impulse response $h(n)$, is defined as :

$$f(h) = \text{IFFT}\left[f\left(\text{FFT}\left[\ 0\ \ \ldots\ h(n)\ \ldots\ \ 0\ \right]\right)\right] \qquad (8)$$

where sufficiently many zeros are prepended and postpended to allow the FIR representation to sufficiently approximate the transformed filter and with the function $f(\cdot)$ acting elementwise on the Fourier transform of the filter. With the FIR matrices expressed in the frequency domain, filter convolution operations are reduced to elementwise multiplications, whereas deconvolution operations become simple divisions [9]. In general, the FIR matrix algebra appears to be an effective approach to address the problem of blind separation of convolutive mixtures.

The natural gradient algorithm is combined with the FIR polynomial matrix algebra and the update equation for estimating the FIR separating system $\underline{\mathbf{W}}$ can be written as in [6] and [10]–[12]:

$$\underline{\mathbf{W}}_{k+1} = \underline{\mathbf{W}}_k + \mu\left(\underline{\mathbf{I}} - \mathbf{g}(\underline{\mathbf{u}})\,\underline{\mathbf{u}}^{\text{H}}\right)\underline{\mathbf{W}}_k \qquad (9)$$

where μ is the step size, $\mathbf{g}(\underline{\mathbf{u}})$ the nonlinearity and $(\cdot)^{\text{H}}$ denotes the Hermitian operator, while the unit FIR polynomial matrix $\underline{\mathbf{I}}$ is given by:

$$\underline{\mathbf{I}} = \begin{bmatrix} \bar{0}\,1\,\bar{0} & \bar{0} \\ \bar{0} & \bar{0}\,1\,\bar{0} \end{bmatrix} \qquad (10)$$

where $\bar{0}$ is a sequence of zeros. This adaptive algorithm is the blind multichannel deconvolution natural gradient algorithm. The objective is to blindly extract the original sources from the convolutive mixtures by estimating the separating FIR polynomial matrix $\underline{\mathbf{W}}$ in order to produce the estimates

$$\underline{\mathbf{u}}(z) = \underline{\mathbf{W}}\,\underline{\mathbf{x}}(z) = \underline{\mathbf{W}}\,\underline{\mathbf{H}}\,\underline{\mathbf{s}}(z) = \underline{\mathbf{\Delta}}\,\underline{\mathbf{s}}(z) \qquad (11)$$

where $\underline{\mathbf{\Delta}}$ should approximate a scaled permutation FIR polynomial matrix. Note the column vectors of the estimates, mixtures and sources defined as in $\underline{\mathbf{u}}(z) = [U_1(z), \ldots, U_m(z)]^T$, $\underline{\mathbf{x}}(z) = [X_1(z), \ldots, X_m(z)]^T$ and $\underline{\mathbf{s}}(z) = [S_1(z), \ldots, S_n(z)]^T$, respectively. The recovered estimates are considered equivalent to the original sources if the whole mixing-unmixing system, namely the global FIR polynomial matrix $\underline{\mathbf{G}}$, is equal to $\underline{\mathbf{G}} = \underline{\mathbf{I}}$ with $\mathrm{g}_{11}, \mathrm{g}_{22} = \delta_{ii}$, i.e., the Kronecker delta. Careful consideration must be taken in the choice of a particular nonlinearity to approximate the pdfs of the sources. This solely depends on the nature of the inputs to be separated. The chosen nonlinearity can be of a parametric form, being continuously adapted, or of a fixed (static) form. It is a well-known assumption that audio and speech signals closely follow a Laplacian distribution with a super-Gaussian pdf model given by:

$$p_s(s_i) = \frac{1}{\sqrt{2}\,\sigma_i}\ \exp\left(-\frac{\sqrt{2}\,|s|}{\sigma_i}\right) \qquad (12)$$

where σ_i^2 defines the variance. The ideal form of the nonlinear function is the one that approximates the cdf of the sources and hence for the natural gradient update method [1]:

$$g_u(u_i) = -\frac{\partial \log\, p_s(u_i)}{\partial(u_i)},\ \ i = 1, 2, \ldots, m. \qquad (13)$$

Combining the above with (12) and assuming unit variance source signals, yields the sigmoid function $\mathrm{g}(\cdot) = \text{sign}(\cdot)$ as the typical choice for the nonlinear function to be used in the algorithm.

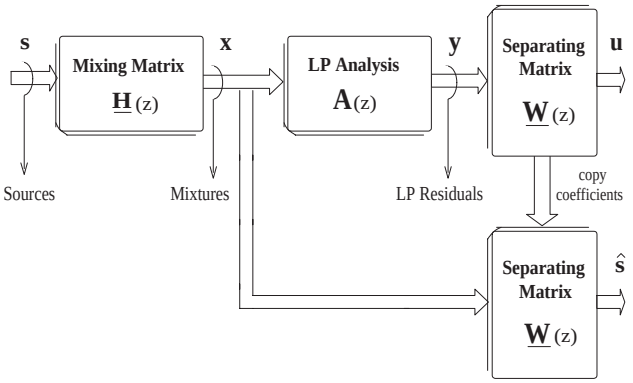

Fig. 2. Proposed system setup for blind separation of convolutive audio mixtures via LP residual analysis.

4. NATURAL GRADIENT ALGORITHM WITH LP ANALYSIS

In this section, we introduce the main concepts behind LP analysis and we further explain the modification proposed on the typical natural gradient algorithm. The system configuration is shown on Fig.2. The rationale behind this approach is to introduce an LP analysis stage, which involves a type of temporal prewhitening of the observed acoustic mixtures. Linear prediction analysis, in general, is used to estimate the LP coefficients that minimize the mean square error between the original signal and the predicted one based on a linear combination of past samples [18]. Hence, we may define a pth order linear predictor filter whose transfer function is given by:

$$A(z) = 1 - \sum_{k=1}^{p}\ \alpha(k)\ z^{-k} \qquad (14)$$

where the vector $\{\alpha(k)\}_{k=1}^{p}$ represents the linear predictor coefficients (LPC's). The LP residual analysis stage is then carried out by filtering the observed speech mixtures with the estimated LP coefficients. This yields the LP residuals, which are temporally independent and are given by:

$$\underline{\mathbf{y}}(z) = \sum_{i=1}^{m}\ A_i(z)\ X_i(z), \quad i = 1, 2, \ldots, m \qquad (15)$$

The extracted LP residuals are then used to blindly adapt the coefficients of the separating FIR polynomial matrix $\underline{\mathbf{W}}$ in the frequency domain, using the standard natural gradient update equation as in (9):

$$\underline{\mathbf{W}}_{k+1} = \underline{\mathbf{W}}_k + \mu\left(\underline{\mathbf{I}} - \mathbf{g}(\underline{\mathbf{u}})\,\underline{\mathbf{u}}^{\text{H}}\right)\underline{\mathbf{W}}_k \qquad (16)$$

with the nonlinearity $\mathbf{g}(\cdot)$ operating in the time domain. The main advantage of this approach is that there can be little to no further entropy increase due to temporal decorrelation. Hence, the entropy maximization criterion is being assisted by the LP analysis and the algorithm follows the directions that spatially separate rather than temporally whiten the signals. The estimated outputs given in the frequency domain by:

$$\underline{\mathbf{u}}(z) = [U_1(z), \ldots, U_m(z)]^T = \underline{\mathbf{W}}\,\underline{\mathbf{y}}(z) \qquad (17)$$

are both spatially and temporally independent. To resolve this we use an extended modification of the method proposed in [5], which was specifically applied in the case of single channel speech dereverberation. The unique spectral characteristics of the original sources are hence restored, by applying the separating FIR matrix to the original mixtures without modification. This is carried out at the extra cost of having to adapt two filters. The outputs of the algorithm, i.e., the extracted estimates being only spatially independent are given by:

$$\underline{\hat{\mathbf{s}}}(z) = \left[\hat{S}_1(z), \ldots, \hat{S}_m(z)\right]^T = \underline{\mathbf{W}}\, \underline{\mathbf{x}}(z) \tag{18}$$

The proposed algorithm operates in the frequency domain using the overlap-save block method. At every iteration, q blocks for each of the input mixtures are processed, with a predefined blocksize L given by:

$$X_i(z) = \text{FFT}\left[x_i(q-1)\,L, \ldots, x_i(q-1)\,(L-1)\right]^T \tag{19}$$

for $i = 1, 2, \ldots, m$. The LP analysis is then performed for every block of the mixture data, yielding the LP residuals in time domain:

$$y_i(q) = x_i(q) - \sum_{k=1}^{p} \alpha_i(k)\, x_i(q-k),\ i = 1, 2, \ldots, m \tag{20}$$

and in the frequency domain:

$$Y_i(z) = X_i(z) - A_i(z)\, X_i(z), \quad i = 1, 2, \ldots, m \tag{21}$$

In a similar manner to the input mixtures, blocks of the LP residuals given by:

$$Y_i(z) = \text{FFT}\left[y_i(q-1)\,L, \ldots, y_i(q-1)\,(L-1)\right]^T \tag{22}$$

are then evaluated at every iteration. The assumptions made are: (**A1**) the system employing the separating FIR polynomial filters is linear; (**A2**) the LP residuals retain enough information to preserve the optimization criterion. Assumption A1 is satisfied if the chosen step sizes of the update equations are kept small [5]. Assumption A2 is an important one as it underlines a basic problem of LP analysis. The solution lies in estimating the set of LP coefficients directly from the mixtures in a way that ensures an accurate estimate of the spectral properties of the signals. Because of the time varying nature of speech the LP coefficients must be calculated from relatively short signal segments. Thus for the LP analysis stage, we propose manipulation of the LP residuals in fairly short, i.e., 20-30 ms segments.

4.1. Connections

Although reminiscent of the approach of [14], the above method presents a number of advantages. The reduction of the correlation achieved by the LP analysis filtering stage proves to be useful in that it removes short-term correlations. By performing filtering on (temporally) prewhitened samples, i) weight updates are rendered independent of one another, ii) the algorithm operation is reduced to spatial separation. All this greatly increases stability and speed of convergence. In terms of algorithm implementation, the LP analysis filtering stage is carried out in a per block basis independently of the general update equation. Avoiding inverse LP filtering for every iteration reduces convergence time as well as computational complexity.

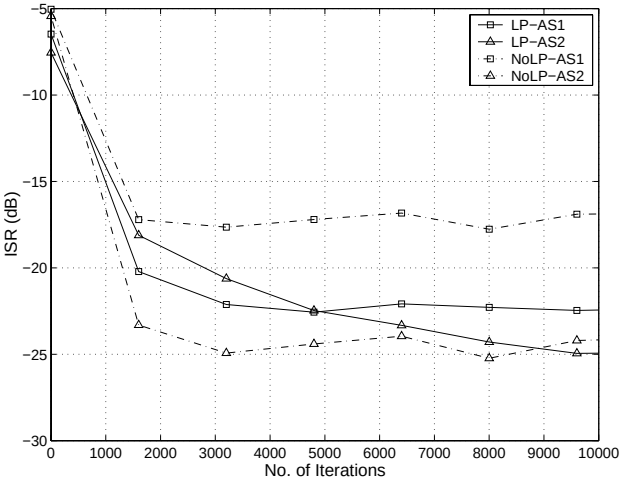

Fig. 3. BSS performance measure index ISR for the convolutive mixtures. LP-AS1: LP-Audio Set 1, LP-AS2: LP-Audio Set 2, NoLP-AS1: No LP-Audio Set 1, NoLP-AS2: No LP-Audio Set 2.

The two-stage methodology adopted in the proposed audio separation scheme is typical of blind space-time equalization methods in multiuser wireless digital communication systems (see, e.g., [17]). In these methods, the first stage aims to remove intersymbol interference, which can essentially be considered as a temporal whitening step. The second stage aims at co-channel interference cancellation through spatial source separation.

In this approach, it is also conjectured that the filters which perform spatial separation on the temporally independent observations, are also capable of spatially separating the coloured observations. The experimental results presented in the following section point to the validity of this assumption.

5. EXPERIMENTAL RESULTS

A number of computer experiments illustrate the above results and evaluate the proposed method in a variety of scenarios. The assessment of the algorithm performance, under synthetic mixing conditions, is based on two different criteria.

5.1. Separation Performance

To assess the capability of the algorithm to separate convolutive mixtures, we employ the interference-to-signal (ISR) performance measure index, expressed in terms of the overall system response as:

$$\text{ISR} = \frac{\left\|\underline{\mathbf{G}}_{ij}\right\|^2}{\left\|\underline{\mathbf{G}}_{ii}\right\|^2}, \quad i \neq j \tag{23}$$

with the global system defined as $\underline{\mathbf{G}} = \underline{\mathbf{W}}\, \underline{\mathbf{H}}$.

Two audio sets of speech signals are convolved with random 3–tap mixing filters to test the algorithm. The first, audio set 1, consists of two recordings of 10 sec each of a male and female anechoic voice. The recording of the second, audio set 2, of a duration of 7 sec has taken place inside a natural environment (noisy room). Both audio sets have been recorded using a sampling frequency of 8 kHz. The mixtures are first tested with the standard

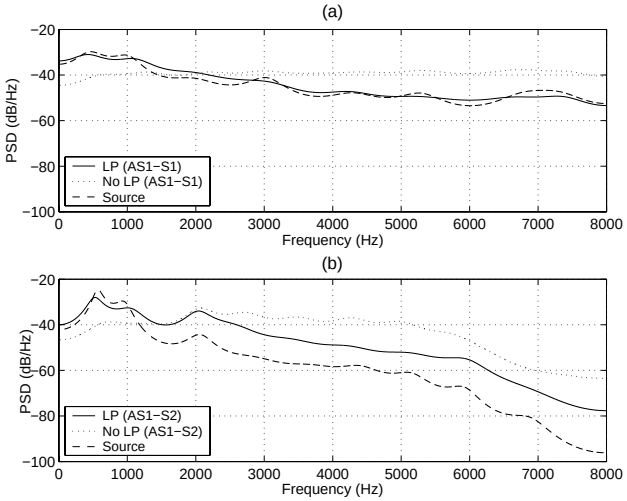

Fig. 4. (a). Power spectral density vs. frequency for Source 1–Audio Set 1 (b). Power spectral density vs. frequency for Source 2–Audio Set 1.

natural gradient algorithm and then with the proposed method employing the LP analysis stage. Both algorithms are executed with step size $\mu = 0.001$ and blocksize $L = 128$, whereas in the proposed method the LP analysis stage is carried out using a 4th-order linear predictor for every block. Fig.3 shows the results of the experiments.

In the case of the first audio set, the proposed algorithm clearly exhibits a better overall performance both in terms of stability and speed of convergence when compared with the natural gradient algorithm when no LP analysis is employed. However, in the case of the second audio set, the algorithm appears to be slower in convergence, yet to some extent, still capable of a lower ISR value.

5.2. Spectral Preservation

In this section, we examine the capability of the algorithm to preserve the spectral characteristics of the original sources, while extracting the estimates. In order to assess this aspect of performance, we introduce a new performance parameter, namely the spectral preservation index (SPI) defined as:

$$\text{SPI} = \text{E}\Big[|S_{x_i}(f) - S_{\hat{x}_i}(f)|^2\Big] \tag{24}$$

where f is a certain frequency, $S_{x_i}(f)$, $S_{\hat{x}_i}(f)$, denote the original and estimated spectral densities, normalized to unit variance, and where $\text{E}\{\cdot\}$ stands for the mathematical expectation operator. For both sets used, the SPI values have been calculated after the algorithms have converged to a separating solution. We calculate the SPI values between the original and estimated signals when only the natural gradient algorithm is used and also when the proposed method with the LP analysis stage, is carried out. These are summarized on Table 1.

Fig.4 shows the power spectral densities plotted against frequency when LP analysis is followed. The similarity of the power spectral densities between the recovered and the original signals validates the spectrum preservation due to the proposed algorithm. On the other hand, note the spectrum of the output from the natural gradient algorithm, which remains nearly flat for the entire range of frequencies, clearly showing that the estimates are whitened. Analogous is the case for both audio sets used in the experiments. Note that audio listening tests have also confirmed the quality of the estimates.

Audio Sets	SPI–LP	SPI–NoLP
AS1–S1	0.077 (−11.13)	1.299 (1.13)
AS1–S2	0.250 (−6.01)	1.646 (2.16)
AS2–S1	0.670 (−1.73)	1.160 (0.64)
AS2–S2	0.465 (−3.32)	1.490 (1.73)

Table 1. SPI values of the power spectral densities of the estimated sources. In parentheses, values given in dB. AS1–S1: Audio Set 1-Source 1, AS1–S2: Audio Set 1-Source2. AS2–S1: Audio Set 2-Source 1, AS2–S2: Audio Set 2-Source2.

5.3. Room Mixtures

We further test the proposed algorithm with real room recordings, corresponding to two male speakers speaking simultaneously in an office. The signals[1] have a duration of 7 sec and were recorded at a sampling rate of 16 kHz. For a more detailed description of the experimental topology, the reader is referred to [12]. The algorithm is executed with a step size of $\mu = 0.002$, whereas the separating FIR polynomial matrix filters are 1024 taps long and the fixed blocksize is 512. To estimate the LP residuals, we use a 17th-order linear predictor corresponding to a processed speech frame size of approximately 30 ms. The mixtures and the recovered sources are shown in Fig.5. Audio listening tests indicate a clean separation. To further assess the performance of the method, we combine the natural gradient algorithm of (16) with the nonlinearity $\phi(u) = \tanh(u) + u$, as derived in [12]. The LP analysis stage is carried out in the same way as before with identical simulation parameters. Experiments using the same audio set, yield clearly separated outputs.

The power spectral densities of the estimates for each of the different nonlinearities, the non-LP natural gradient estimates and the estimates obtained in [12] and [14] are all shown in Fig.5. When LP analysis is used, the extracted estimates are found to preserve the original speech spectral characteristics. By contrast, when the LP analysis stage is omitted, the resulting spectra appear to be flat, indicating the unwanted whitening effect imposed on the extracted signals. In addition, the capability of the method to preserve the spectrum of each output source seems to be invariant with respect to the nonlinearity used. Note also the close similarity of the spectra of our estimates with the post-processed (dewhitened) outputs of [12] and the estimates of [14]. Worth emphasizing is also the fact that the similarity between the spectra of the estimates is not always indicative of the separation level achieved. Yet in this particular example, listening tests indicate that our estimates have the same perceptive quality when compared with the outputs in [12], whereas marginally improve the estimates of [14]. This will be demonstrated in the presentation of this work.

[1]The speech mixtures used in the experiments are available on-line at `http://www.cnl.salk.edu/∼tewon/Blind/blind_audio.html`.

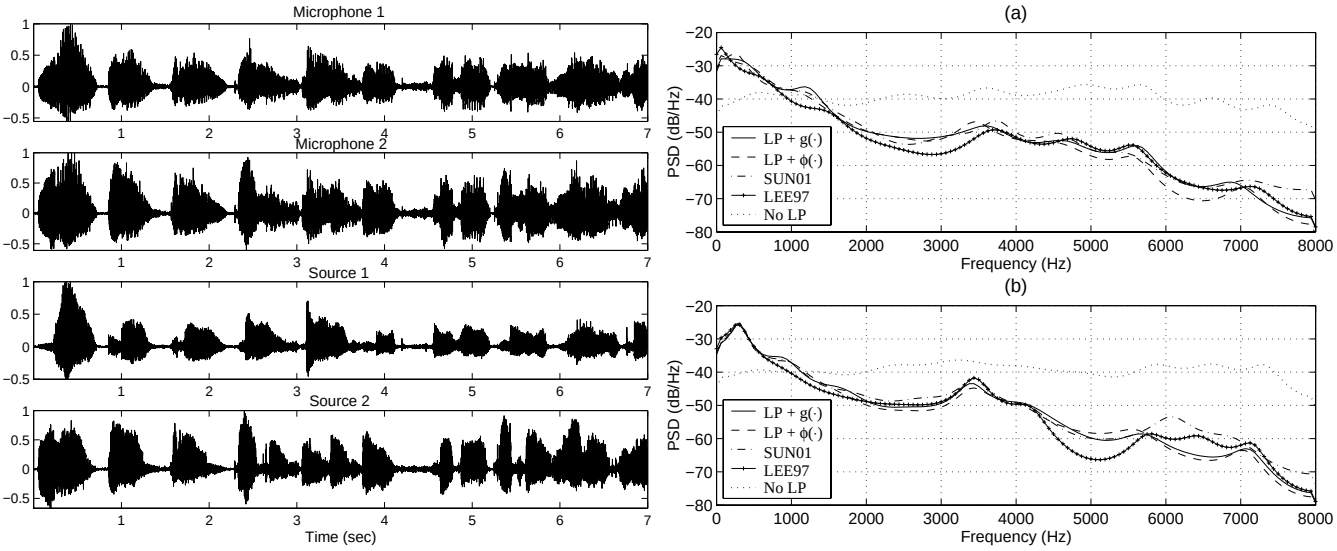

Fig. 5. Speech mixtures and separated sources (left). Power spectral densities of the estimates against frequency (right). Natural gradient with LP analysis and $\mathrm{g(u)} = \mathrm{sign}(u)$: LP + $\mathrm{g}(\cdot)$, natural gradient with LP analysis and $\phi(u) = \tanh(u) + u$: LP + $\phi(\cdot)$, modified natural gradient algorithm of [14]: SUN01, extracted (dewhitened) estimates from [12]: LEE97 and outputs omitting the LP analysis stage: No LP.

6. CONCLUSIONS AND FURTHER WORK

This paper has concentrated on blind separation of speech signals in real acoustic environments. The dynamic nature of the mixing process has been dealt with the use of the FIR polynomial matrix algebra, capable of efficiently representing the multipath and multichannel nature of the problem. We have employed the natural gradient algorithm, which when combined with the entropy maximization criterion yields the successful separation of the audio mixtures. The problem of the temporal whiteness of the estimates, which hinders their listening quality, is overcome by the use of LP analysis in the algorithm. Results have demonstrated the spectrum preservation of the estimates. The fast convergence to a separating solution and the inherent simplicity of this method substantiate it as a significant improvement over previously suggested methods. Further work is focused on extending the proposed method in more realistic scenarios of real room recording situations.

7. REFERENCES

[1] S. Amari, A. Cichocki and H. Yang, "A New Learning Algorithm for Blind Signal Separation" In *Advances in Neural Information Processing Systems 8*, MIT Press, Cambridge, 1996, pp. 757–763.

[2] A. Bell and T. Sejnowski, "An Information Maximization Approach to Blind Separation and Blind Deconvolution" *Neural Computation*, Vol. 7, No. 6, 1995, pp. 1129–1159. 998, pp. 2009–2025.

[3] E. C. Cherry, "Some Experiments on the Recognition of Speech with One and with Two Ears" *Journal of the Acoustical Society of America*, Vol. 25, No. 5, September 1953, pp. 975–979.

[4] P. Comon, "Independent Component Analysis: A new concept?" *Signal Processing*, Vol. 36, No. 3, April 1994, pp. 287–314.

[5] B. W. Gillespie, H. S. Malvar and D. A. F. Florencio "Speech Dereverberation Via Maximum–Kurtosis Subband Adaptive Filtering" In *Proc. ICASSP*, Salk Lake City, May 2001, Vol. 6, pp. 3701–3704.

[6] M. Joho, H. Mathis and G. S. Moschytz "An FFT-Based Algorithm for Multichannel Blind Deconvolution" In *Proc. ISCAS*, Orlando, Florida, May 1999, pp. 203–206.

[7] C. Jutten and J. Herault, "Blind Separation of Sources, Part I: An Adaptive Algorithm based on Neuromimetic Architecture" *Signal Processing*, Vol. 24, No. 1, July 1991, pp. 1–10.

[8] H. Kuttruf, *Room Acoustics*. New York: Elsevier Science Publishers Ltd., 1991.

[9] R. H. Lambert, *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures*. Ph.D. Thesis, University of Southern California, May 1996.

[10] R. H. Lambert and A. J. Bell, "Blind Separation of Multiple Speakers in a Multipath Environment" In *Proc. ICASSP*, Munich, Germany, April 1997, pp. 423–426.

[11] T.-W. Lee, A. J. Bell and R. H. Lambert, "Blind Separation of Delayed and Convolved Sources" In *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge, 1997, pp. 758–764.

[12] T.-W. Lee, A. J. Bell, and R. Orglmeister, "Blind Source Separation of Real World Signals" In *Proc. ICNN*, Houston, Texas, June 1997, pp. 2129–2135.

[13] K. J. Pope and R. E. Bogner, "Blind Signal Separation II. Linear, Convolutive Combinations" *Digital Signal Processing*, Vol. 6, No. 3, 1996, pp. 17–28.

[14] X. Sun and S. C. Douglas, "A Natural Gradient Convolutive Blind Source Separation Algorithm for Speech Mixtures," In *Proc. ICA*, San Diego, CA, December 2001, pp. 59–64.

[15] K. Torkkola, "Blind Separation of Delayed Sources based on Information Maximization" In *Proc. ICASSP*, Atlanta, May 1996, pp. 3510–3513.

[16] K. Torkkola, "Blind Separation of Convolved Sources based on Information Maximization" In *IEEE Workshop on Neural Networks for Signal Processing*, Japan, 1996, pp. 423–432.

[17] A.-J. van der Veen, S. Talwar and A. Paulraj, "A Subspace Approach to Blind Space-Time Signal Processing for Wireless Communication Systems" *IEEE Transactions on Signal Processing*, Vol. 45, No. 1, Jan. 1997, pp. 173–190.

[18] B. Yegnanarayana and P. Satyanarayana, "Enhancement of Reverberant Speech Using LP Residual Signal" *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 3, May 2000, pp. 267–281.