AN ONLINE ALGORITHM FOR BLIND EXTRACTION OF SOURCES WITH DIFFERENT DYNAMICAL STRUCTURES

Danilo P. Mandic[1] *and Andrzej Cichocki*[2]

1. Department of Electrical Engineering, Imperial College, London, UK.
2. BSP Research Group, RIKEN, Japan.
E–mail: d.mandic@ic.ac.uk, cia@bsp.brain.riken.go.jp

ABSTRACT

A novel online algorithm for Blind Source Extraction (BSE) of instantaneous signal mixtures is proposed. The algorithm is derived for a structure comprising of a demixing stage and an adjacent adaptive predictor, whereby signal extraction is based upon the predictability of an unknown source signal. This way, the coefficients of the demixing matrix and adaptive predictor are estimated simultaneously. To improve the convergence and to be able to cope with the dynamics of the mixture, the algorithm is further normalised based upon the minimisation of the a posteriori prediction error. This makes it also suitable for environments where the mixing process is time varying. No assumptions or constraints on the norm of the coefficients or signals are required. Simulations on mixtures of real world physiological data for both the fixed and time varying mixing matrix support the analysis.

1. INTRODUCTION

Problems related to separating mixed or convolved signals have been a major research topic during the last decade. The recent progress in the area has made it possible to gain insights into very complex problems, such as the ones in biomedical signal processing, where the signals are typically weak and noisy, the channels are heavily correlated and the signal generating mechanism is not known. There are many established Blind Source Separation (BSS) and Blind Source Extraction (BSE) algorithms, for an overview see [1, 2]. In the real world, especially in brain research and medicine, we deal with signals coming from a number of sensors, for instance the number of sensors in electroencephalography (EEG) varies from 56 to 256. If the independent input signals are denoted by $\mathbf{s}(k) = [s_1(k), \ldots, s_M(k)]^T$, and the linear instantaneous mixing process is governed by the mixing matrix $\mathbf{A}_{M\times M}$, then the set of signals obtained from the sensors can be described by

$$\mathbf{x}(k) = \mathbf{A}(k)\mathbf{s}(k) \tag{1}$$

where both the measured signals $\mathbf{x}(k) = [x_1(k), \ldots, x_M(k)]^T$ and sources $\mathbf{s}(k)$ are column vectors and $(\cdot)^T$ denotes the vector transpose. The signal separation process aims at finding an optimisation procedure for the inverse of the matrix $\mathbf{A}(k)$ from (1), that is $\mathbf{W}(k) = \mathbf{A}^{-1}(k)$ such that

$$\mathbf{y}(k) = \mathbf{W}(k)\mathbf{x}(k) \tag{2}$$

where $\mathbf{y}(k) = [y_1(k), \ldots, y_M(k)]^T$ denotes the separated or extracted signals. This procedure is very computationally demanding, and often leads to ill–conditioned problems [1]. In fact, instead of estimating $\mathbf{A}^{-1}$, the set of estimated parameters is such that the product $\mathbf{AW}$ equals a product of a permutation and delay matrix.

In many practical situations we are interested only in a subset of the sensor signals, often in only very few of them, for instance we might want to extract only music from a cocktail party. In such a case there is a need to devise an algorithm that would extract the desired signal(s) from the mixture $\mathbf{x}$ based upon some prescribed criterion. Such algorithms exist and have found practical implementations [3]. A suitable criterion for an online gradient descent algorithm for blind extraction of sources from their linear mixtures might be some fundamental property of a signal, such as smoothness, predictability, or temporal correlation. Such criteria become even more important in noisy measurements, which is typical for practical applications. Recently, one such algorithm, based upon combining linear prediction and BSE was proposed in [4]. The algorithm proposed there, although efficient, was a batch algorithm, which is not suitable for on–line processing, and also possesses considerable computational complexity, since the parameter updates require estimation of the autocorrelation and crosscorrelation matrix. In addition, such an algorithm is not suitable for environments with a time–varying mixing matrix $\mathbf{A}(k)$.

Following the approach from [4, 1] we therefore propose a novel direct gradient online algorithm for blind extraction of linearly mixed sources. The algorithm is based upon the fundamental property of predictability of source signals (for an overview of linear and nonlinear predictability see [5, 6]), and is aimed at minimising the instantaneous squared prediction error of the extracted source. To improve the convergence, the coefficient update is further normalised by a measure of energy of the mixed input and extracted signal. The proposed set of algorithms is shown to be robust to time variations of the mixing matrix. No constraints on the values or norms of the signals and coefficients are required.

2. THE ALGORITHM FOR BSE USING PREDICTABILITY OF SIGNALS

A somewhat simplified BSE scheme, based upon predictability of a source signal as a criterion for extraction, is shown in Figure 1. The Figure illustrates the model for extraction of only one channel (denoted by y_1) from the mixture $\mathbf{x}(k)$. The structure consists of a standard demixing matrix (represented by an appropriate column in Figure 1), followed by an adaptive linear predictor. The output of the adaptive predictor is $\tilde{y}_1$ whereas the associated prediction error

$$e_1(k) = y_1(k) - \tilde{y}_1(k)$$

is a measure upon which the adaptation is based. For the structure shown in Figure 1, we propose a global online learning algorithm for adapting both the adaptive predictor weights $\mathbf{b}_1(k) = [b_{11}(k), \ldots, b_{1N}(k)]^T$ and the corresponding row $\mathbf{w}_1(k) = [w_{11}(k), \ldots, w_{1M}(k)]^T$ of the demixing matrix $\mathbf{W}(k) = [\mathbf{w}_1(k); \cdots; \mathbf{w}_M(k)]$, where the adaptive filter (predictor) is of length N and the demixing matrix $\mathbf{W}$ is of size $M \times M$. There exist online approaches where the cost function consists of both the error term and the term constraining the norm of the extracted signal [1]. However, to devise an algorithm that is robust to perturbations in the mixing matrix and time–varying statistics of source signals, we provide methodology based upon a simple cost function in the form of the squared instantaneous prediction error, with no constraints required on the parameters of the system. Therefore we start from the cost function most frequently used in adaptive filtering, given by

$$J(\mathbf{w}_1(k), \mathbf{b}_1(k)) = \frac{1}{2} e_1^2(k) \tag{3}$$

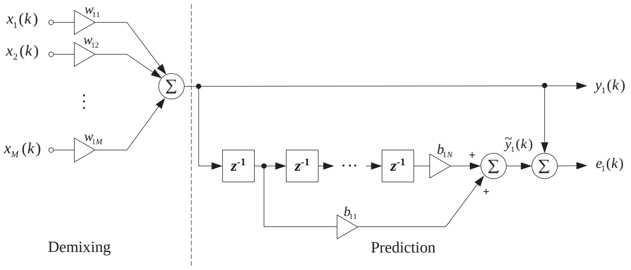

Fig. 1. A BSE structure for extraction of one source component, using a linear predictor

where $e_1(k)$ is the instantaneous output error of the structure (Figure 1) at time instant k. Notice, from Figure 1

$$\mathbf{y}_1(k) = [y_1(k-1), \ldots, y_1(k-N)]^T \tag{4}$$

The output signal and its predicted version are given by

$$\begin{aligned} y_1(k) &= \sum_{i=1}^{M} x_i(k) w_{1i}(k) = \mathbf{x}^T(k)\mathbf{w}_1(k) \\ \tilde{y}_1(k) &= \sum_{j=1}^{N} b_{1j}(k) y_1(k-j) \\ &= \sum_{j=1}^{N} b_{1j}(k) \sum_{i=1}^{M} x_i(k-j) w_{1i}(k-j) \end{aligned} \tag{5}$$

We next derive gradient descent learning rules for the system based upon minimisation of cost function (3). For every element $b_{1j}(k),\ j = 1, \ldots, N$ of the parameter vector $\mathbf{b}_1$ and every element $w_{1i}(k),\ i = 1, \ldots, M$ of $\mathbf{w}_1$, we have

$$b_{1j}(k+1) = b_{1j}(k) - \mu_b \nabla_{b_{1j}} J(\mathbf{w}_1(k), \mathbf{b}_1(k)) \tag{6}$$

$$w_{1i}(k+1) = w_{1i}(k) - \mu_w \nabla_{w_{1i}} J(\mathbf{w}_1(k), \mathbf{b}_1(k)) \tag{7}$$

The error term can be evaluated as

$$\begin{aligned} e_1(k) &= y_1(k) - \tilde{y}_1(k) = \sum_{i=1}^{M} w_{1i}(k) x_i(k) - \\ &- \sum_{j=1}^{N} b_{1j}(k) \sum_{i=1}^{M} x_i(k-j) w_{1i}(k-j) \end{aligned} \tag{8}$$

which gives

$$\begin{aligned} \nabla_{b_{1j}} J(\mathbf{w}_1(k), \mathbf{b}_1(k)) &= e_1(k) \frac{\partial e_1(k)}{\partial b_{1j}(k)} \\ &= -e_1(k) y_1(k-j) \end{aligned} \tag{9}$$

and

$$\nabla_{w_{1i}} J(\mathbf{w}_1(k), \mathbf{b}_1(k)) = e_1(k) x_i(k) \tag{10}$$

The updates for the adaptive filter and demixing coefficients now become

$$\begin{aligned} b_{1j}(k+1) &= b_{1j}(k) + \mu_b e_1(k) y_1(k-j) \\ w_{1i}(k+1) &= w_{1i}(k) - \mu_w e_1(k) x_i(k) \end{aligned} \tag{11}$$

or in the vector form

$$\begin{aligned} \mathbf{b}_1(k+1) &= \mathbf{b}_1(k) + \mu_b e_1(k) \mathbf{y}_1(k) \\ \mathbf{w}_1(k+1) &= \mathbf{w}_1(k) - \mu_w e_1(k) \mathbf{x}(k) \end{aligned} \tag{12}$$

Notice the different learning rates μ_b and μ_w which respectively correspond to the adaption of the coefficients of the adaptive predictor and those of the demixing matrix. They need to be chosen by the user, which is critical to the performance of a gradient descent based learning algorithm. Algorithms with fixed learning rates have difficulties to cope with nonlinear and nonstationary signals and with signals with high variance. Let us therefore next investigate the effect of normalising the learning rates in order to make the convergence faster and make the algorithm better suited for variations in the statistical properties of source signals and perturbations in the mixing matrix.

3. A NORMALISED ONLINE ALGORITHM FOR BLIND EXTRACTION OF SOURCES BASED UPON PREDICTABILITY OF A SIGNAL

Following the approach from [6] let us represent the a posteriori output error of the structure as

$$e(k+1) = e(k) + \sum_{j=1}^{N} \frac{\partial e(k)}{\partial b_{1j}(k)} \Delta b_{1j}(k) + \sum_{i=1}^{M} \frac{\partial e(k)}{\partial w_{1i}(k)} \times$$

$$\times \Delta w_{1i}(k) + \sum_{j=1}^{N}\sum_{i=1}^{M} \frac{\partial^2 e_1(k)}{\partial b_{1j}(k) \partial w_{1i}(k)} + \cdots \quad (13)$$

Assuming, for simplicity, that the second and higher order derivatives in the Taylor series expansion (13) are negligible (which is reasonable for online gradient descent algorithms), and noticing that the terms $\Delta b_{1j}(k)$ and $\Delta w_{1i}(k)$ are readily obtained from (11), we have [7]

$$\begin{aligned} \frac{\partial e(k)}{\partial b_{1j}(k)} &= -y_1(k-j) \\ \frac{\partial e(k)}{\partial w_{1i}(k)} &= x_i(k) \end{aligned} \quad (14)$$

From (13) and (14), we have

$$\begin{aligned} e(k+1) &= e(k) + \sum_{j=1}^{N} (-1) y_1(k-j) \mu_b e_1(k) y_1(k-j) + \\ &\quad + \sum_{i=1}^{M} (-1) x_i(k) \mu_w e_1(k) x_i(k) \\ &= e_1(k) \left[1 - \mu_b \parallel \mathbf{y}_1(k) \parallel_2^2 - \mu_w \parallel \mathbf{x}(k) \parallel_2^2 \right] \end{aligned} \quad (15)$$

Let us assume, for convenience that $\mu_b = \mu_w = \mu_0$. Then, from (15) to achieve $e(k+1) = 0$, the optimal learning rate is obtained by minimizing the term in the square brackets as

$$\mu = \frac{\mu_0}{\parallel \mathbf{y}_1(k) \parallel_2^2 + \parallel \mathbf{x}(k) \parallel_2^2} \quad (16)$$

The constant learning rate μ_0 from (16) is therefore normalised by both the norm of the instantaneous mixed signal $\parallel \mathbf{x}(k) \parallel_2^2$ and the norm $\parallel \mathbf{y}_1(k) \parallel_2^2$ of the extracted version of the signal of interest. From (15) $\mu_0 = 1$, whereas in practice, due to the unknown dynamics and coupling between the adaptation of $\mathbf{w}_1$ and $\mathbf{b}_1$, as well as the approximation in the derivation of the algorithm, the parameter μ_0 needs to be considerably smaller.

3.1. Individual Normalisation of Learning Rates

Intuitively, and from the experience with the normalised LMS (NLMS) algorithm [8], it would be natural to normalise each of the learning rates μ_b and μ_w by an estimate of the tap input power of their respective input signals. There are many combinations of μ_b and μ_w which minimise (15). For convenience, we opt for their convex combination, achieved by introducing a mixing parameter λ into (15), such that $0 \leq \lambda \leq 1$ [9], which gives

$$\begin{aligned} \mu_b &= \frac{\lambda}{\parallel \mathbf{y}_1(k) \parallel_2^2} \\ \mu_w &= \frac{1-\lambda}{\parallel \mathbf{x}(k) \parallel_2^2} \end{aligned} \quad (17)$$

The individual learning rates corresponding to $\mathbf{b}_1$ and $\mathbf{w}_1$ are now normalised by the tap input power of their corresponding inputs. By choosing the value of λ we introduce weighting into the adaptation of $\mathbf{b}$ and $\mathbf{w}$. The algorithms based upon (16) and (17) are expected to exhibit increased convergence and signal extraction ability, as compared to algorithms with a fixed learning rate, as well as to be better suited for processing of signals with large dynamics and in time varying mixing environments.

4. SIMULATION RESULTS

In the simulations, the mixing matrix $\mathbf{A}$ was chosen from the set of random matrices so as to have its condition number less or equal to ten. The source signals were a mixture of physiological recordings, including heart rate, EEG, experimental noise, and respiratory signal, whose mixture is shown in Figure 2. Three of the signals were with long time correlation, one of them decaying in time. The task was to extract only one of the source signals from their mixture, based upon their predictability. For this setting we investigated the performances of the proposed online direct gradient descent algorithm given in Section 2 and its normalised versions described in Sections 3 and 3.1. The demixing matrix $\mathbf{W}$ was randomly initialised, whereas the coefficients of an adaptive predictor were initialised either randomly or with zero values. In most cases, the random initialisation provided better results. The simultaneous adaptation of coefficient vectors $\mathbf{b}_1$ and $\mathbf{w}_1$ is a complex task, since the coefficient vectors are, strictly speaking, not independent.

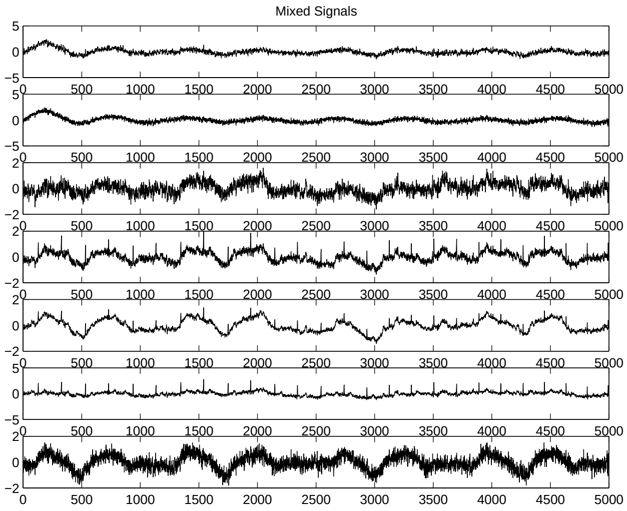

Fig. 2. Linear instantaneous mixture of independent source signals

Therefore the learning rate in the global gradient descent algorithm for the structure had to be chosen fairly small, and varied between 10^{-3} to 10^{-6}. Consequently, the learning rates of the normalised algorithms were also chosen to be small, but typically one to two orders of magnitude larger than those of the algorithm with fixed learning rates. The length of the adaptive filter was varied between $N = 1$ and $N = 400$.

Figure 3 illustrates source extraction using a fixed learning rate parameter. The signal extracted changed with the length of the adaptive predictor, as highlighted in the Figure caption. The smoother sources $S3$ and $S4$ required longer adaptive predictors than the sources with faster transitions $S1$ and $S2$. Single source extraction was not sufficiently good for large stepsizes, whereas for relatively small stepsizes, the extraction was possible, albeit with a significant noise.

Figure 4 illustrates the extraction using the algorithm with joint normalisation of stepsizes, given in (16). Source $S4$, which is the smoothest, was best extracted using an adaptive predictor of length $N = 200$, whereas source $S1$ (the most rapidly varying heart rate variability signal) was best extracted using a predictor of $N = 2$. Generally speaking, smoother sources required longer predictors. There was much less noise in extracted sources than in the case of the algorithm with the constant learning rate.

Figure 5 illustrates extraction of a single source using individual normalisation for μ_b and μ_w (17). The extraction results were better than those using joint normalisation, and the sources were again able to be extracted based upon the length of the predictor required. The quality of extraction was better than in the previous two cases.

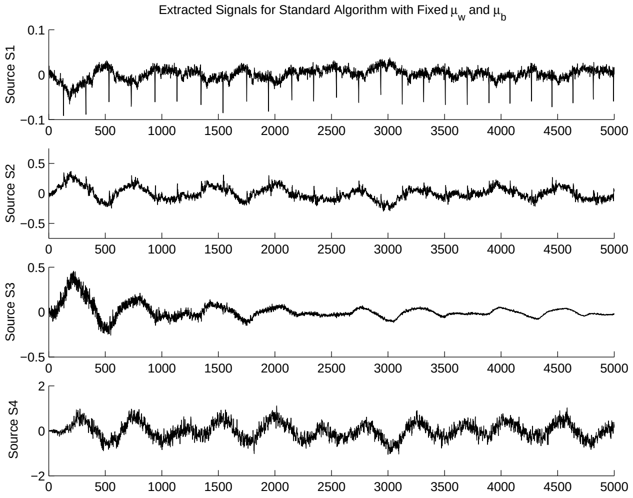

Fig. 3. Signals extracted using fixed learning rates with respectively $N = 1$, $N = 40$, $N = 50$, and $N = 400$

The proposed structure and algorithm were able to extract sources based upon their predictability and clearly distinguish between the sources by varying the length of the adaptive predictor from Figure 1. In all the cases, the source extracted and the order of the predictor followed the same pattern, requiring typically $N = 1 - 4$ for source $S1$, $N = 10 - 20$ for source $S2$, $N = 30 - 80$ for source $S3$, and $N > 100$ for source $S4$. The algorithm performed better on signals with structure (underlying nonlinear dynamics), than on signals without much structure and with sharp tran-

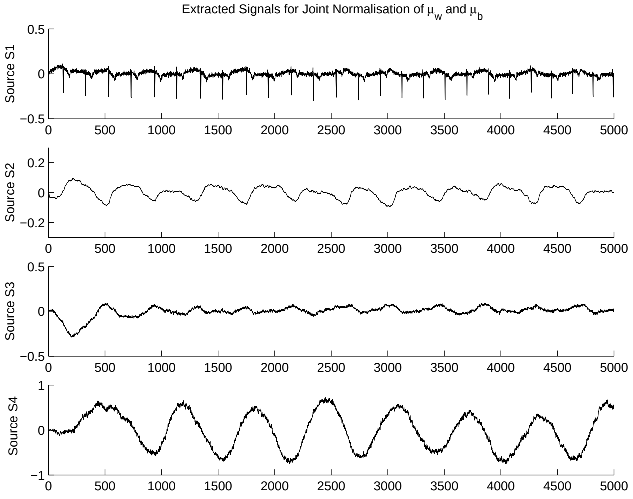

Fig. 4. Signals extracted using joint learning rate normalisation (equation (16)) with respectively $N = 1$, $N = 12$, $N = 40$, and $N = 200$

sition, such as the HRV signal.

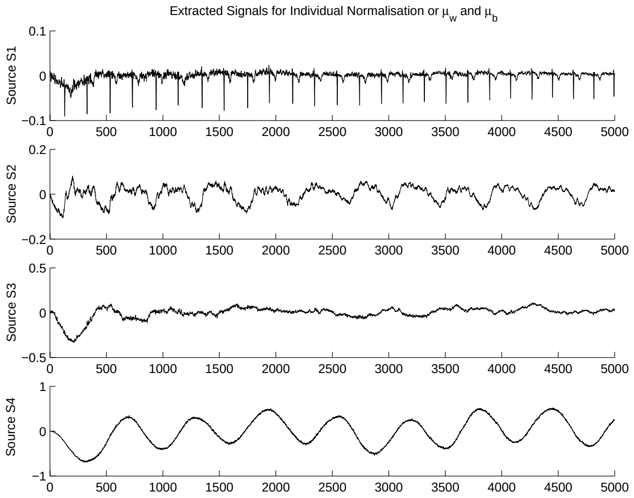

Fig. 5. Signals extracted using individual learning rate normalisation (equation (17)) with respectively $N = 2$, $N = 20$, $N = 30$, and $N = 400$

4.1. Experiments with a Time–Varying Mixing Matrix

In the next set of experiments, behaviour of the proposed algorithms was investigated for a time–varying mixing matrix. For that purpose, the mixing matrix $\mathbf{A}$ was modelled as random walk, given by (using the MATLAB notation)

$$\mathbf{A}(k) = \mathbf{A} + \text{rand}(\text{size}(A))/A_0 \tag{18}$$

where factor A_0 was used to scale the additive white noise. Matrix $\mathbf{A}$ was first scaled so that the sum of the absolute values of every row was equal to unity. For the experiment, A_0 was chosen such that the noise matrix had its norm approximately equal to that of $\mathbf{A}$, which provided significant perturbation of $\mathbf{A}$ in time.

The extraction results for the case of the algorithm with the fixed learning rate are shown in Figure 6. Due to time variation of $\mathbf{A}$, the standard algorithm showed poor performance. The results for joint normalisation of μ_w and μ_b for the time varying mixing matrix $\mathbf{A}$ are shown in Figure 7. The results were much better than for the standard algorithm and the sources could be clearly distinguished. For the algorithm with individual normalisation of μ_w and μ_b the results of simulations are shown in Figure 8, and are of better quality than the former two.

In comparison with the results with fixed $\mathbf{A}$, shown in Figures 3, 4 and 5, the respective extracted waveforms for time varying mixing environment were markedly noisier. In addition, for the time varying case, the magnitudes of the extracted signals tend to be considerably smaller. As for the lengths of the adaptive filter which provide best extraction of individual signals, similar conclusions as for the case of the fixed mixing matrix hold.

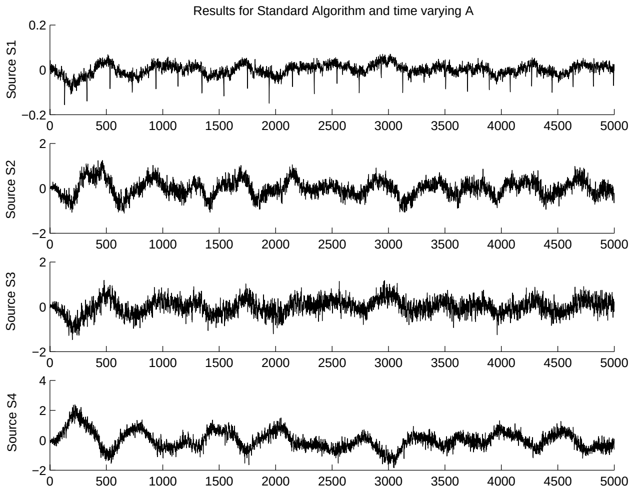

Fig. 6. Signals extracted using fixed learning rates with time–varying mixing matrix $\mathbf{A}$

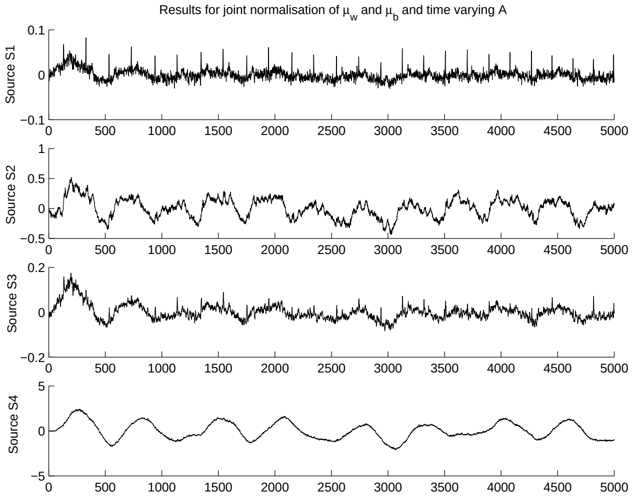

Fig. 7. Signals extracted using jointly normalised learning rates (equation (16)) with time–varying mixing matrix $\mathbf{A}$

5. DISCUSSION AND CONCLUSIONS

An online learning algorithm for blind extraction of single sources, which combines the concepts of blind extraction and adaptive prediction, has been proposed. It has been initially shown that by varying the length of the adaptive predictor it has been possible to extract particular single sources from their mixture. To improve the convergence and

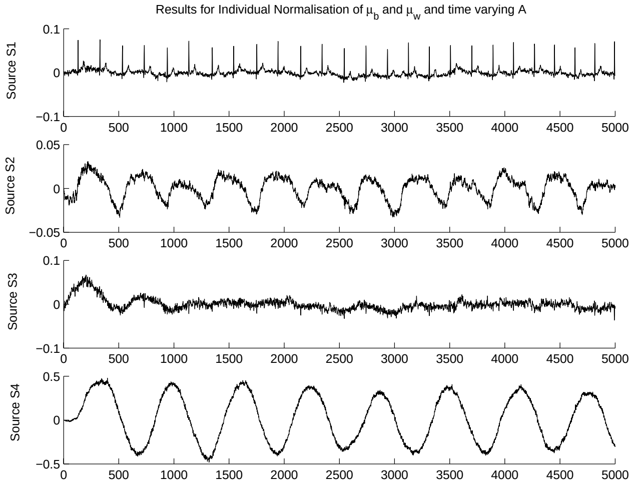

Fig. 8. Signals extracted using individual learning rate normalisation (equation (17)) with time varying mixing matrix **A**

quality of extraction, two normalised algorithms have been derived for the update of the demixing matrix and adaptive predictor coefficients, one with the single normalised learning rate and the other with individually normalised learning rates. The normalised algorithms comprehensively outperformed the standard one, exhibiting faster convergence and less noise in the extracted signals, and showed excellent performance for both the fixed and time varying mixing environment. The waveforms of extracted signals in the case of a time varying mixing environment were noisier and had smaller amplitudes.

To build a single signal extractor based upon predictability of source is a challenging task. The length of the predictor plays a crucial role here, especially for signals that exhibit some kind of structure in the phase space, as is the case with many physiological signals. If the tap input delay line of an adaptive predictor is too short then the tap input is dominated by noise. On the other hand, if the tap input delay line is too long, so as to be longer than the embedding dimension of the signal augmented by the prediction horizon of a signal, then the beginning and end of the tap delay line are not correlated, which results in poor prediction. Therefore the length of the tap delay input line can indeed serve to select which signal to extract provided the signal has some kind of structure and is predictable. This, in turn, means that the algorithm is not very well suited for extraction of noise signals. However, most of the physiological signals of interest possess structure in the phase space, making the architecture and proposed algorithms well suited for such a task. The set of algorithms proposed here has been therefore derived with the aim to cater for the time variation of the mixing matrix and the varying dynamics of the source signals.

Extracting sources based upon their predictability fully complies with the results from psychology (where the cocktail party effect was actually first introduced in 1953 [10]), where it has been shown that humans are able to better separate sounds if the predictability of the messages was higher. In addition, the performance of listeners varied with the amount of information in messages, rather than with the amount of physical stimulation [11], which is very much the case with the results presented here.

Acknowledgements: We would like to thank Prof. Jonathon Chambers of Bath University for his critical comments.

6. REFERENCES

[1] A. Cichocki and S.-I. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. Wiley, 2002.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001.

[3] “www.bsp.brain.riken.go.jp/ICALAB.”

[4] A. Cichocki, T. Rutkowski, and K. Siwek, “Blind signal extraction of signals with specified frequency band,” in *Proceedings of NNSP 2002*, pp. 515–524, 2002.

[5] J. Makhoul, “Linear prediction: A tutorial overview,” *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[6] D. P. Mandic and J. A. Chambers, *Recurrent Neural Networks for Prediction: Architectures, Learning Algorithms and Stability*. Wiley, 2001.

[7] D. P. Mandic, “The NNGD algorithm for neural adaptive filters,” *Electronics Letters*, vol. 36, no. 9, pp. 845–846, 2000.

[8] D. T. M. Slock, “On the convergence properties of the LMS and the normalized LMS algorithms,” *IEEE Transactions on Signal Processing*, vol. 41, no. 9, pp. 2811–2825, 1993.

[9] C. Boukis, *Adaptive Signal Processing Structures and Identification Algorithms for Feedback Control*. PhD Thesis, Imperial College, London, UK, 2003.

[10] E. C. Cherry, “Some experiments on the recognition of speech, with one and two ears,” *Journal of the Acoustic Society of America*, vol. 25, pp. 975–979, 1953.

[11] D. E. Broadbent, *Perception and Communication*. Pergammon Press, 1958.