

# Blind Source Separation of Real World Signals

Te-Won Lee  
Max-Planck-Society, GERMANY, and  
Computational Neurobiology Laboratory  
The Salk Institute  
10010 N. Torrey Pines Road  
La Jolla, California 92037, USA  
tewon@salk.edu

Anthony J. Bell  
Computational Neurobiology Laboratory  
The Salk Institute  
10010 N. Torrey Pines Road  
La Jolla, California 92037, USA  
tony@salk.edu

Reinhold Orglmeister  
Department of Electrical Engineering,  
Berlin University of Technology, GERMANY  
orglm@tubif1.ee.tu-berlin.de

## Abstract

*We present a method to separate and deconvolve sources which have been recorded in real environments. The use of noncausal FIR filters allows us to deal with nonminimum mixing systems. The learning rules can be derived from different viewpoints such as information maximization, maximum likelihood and negentropy which result in similar rules for the weight update. We transform the learning rule into the frequency domain where the convolution and deconvolution property becomes a multiplication and division operation. In particular, the FIR polynomial algebra techniques as used by Lambert present an efficient tool to solve true phase inverse systems allowing a simple implementation of noncausal filters. The significance of the methods is shown by the successful separation of two voices and separating a voice that has been recorded with loud music in the background. The recognition rate of an automatic speech recognition system is increased after separating the speech signals.*

## 1 Introduction

In blind source separation the problem is to recover independent sources given sensor outputs in which the sources have been mixed by an unknown channel. The problem has become increasingly important in the signal and speech processing area due to their prospective application in speech

recognition, telecommunications and medical signal processing.

The blind source separation problem has been studied by researchers in the field of neural networks [1, 2, 5, 9, 10, 16, 17] and statistical signal processing [3, 7, 11, 15, 19]. Comon [7] defines the concept of independent component analysis (ICA) which measures the degree of independence among outputs using contrast functions approximated by the Edgeworth expansion of the Kullback-Leibler divergence. The higher order statistics is approximated by cummulants up to 4th order and requires intensive computation. Researchers in neural computation have developed adaptive learning algorithms which are simpler and biologically more plausible [1, 2, 5, 9, 10].

Recently, Bell and Sejnowski [2] have proposed an information theoretic approach to the blind source separation and blind deconvolution problem. This approach has been extended to convolution and time-delays in a feedback architecture [6, 13, 18]. Pearlmutter and Parra have reformulated the ICA in a maximum likelihood (ML) framework [16] where the underlying density is estimated in a context sensitive manner. Although research in blind source separation has been carried out for several years only very few papers have addressed the problem with real acoustic signals recorded in reverberating environments [19, 18, 13, 12].

In this paper, we tackle the problem of separating signals recorded in real environments. The inverting system is approximated by a matrix of finite impulse response (FIR) filters to deconvolve and unmix the mixing system which may

have a nonminimum phase character. The learning rules can be derived from different perspectives such as information maximization, ML and negentropy which result in similar rules for the weight update. Another way of dealing with filters in a multichannel representation is to transform the learning rules into the frequency domain where the convolution and deconvolution property becomes a multiplication and division operation. In particular, the use of FIR polynomial techniques [11] present an efficient tool to solve true phase inverse systems allowing a simple implementation of noncausal filter solutions. The significance of the methods is shown by the successful separation of two voices and separating a voice that has been recorded with loud music in the background. We also show that the recognition rate of an automatic speech recognition system is increased after separating the speech signals.

## 2 Architecture

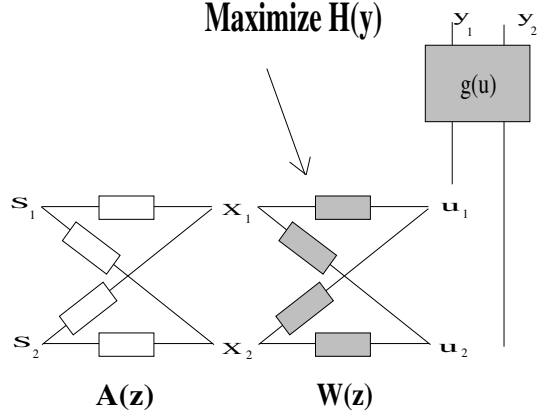
In the linear blind signal processing problem [2, 3, 7],  $N$  signals,  $\mathbf{s}(t) = [s_1(t) \dots s_N(t)]^T$ , are transmitted through a medium so that an array of  $N$  sensors picks up a set of signals  $\mathbf{x}(t) = [x_1(t) \dots x_N(t)]^T$ , each of which has been mixed, delayed and filtered as follows:

$$x_i(t) = \sum_{j=1}^N \sum_{k=0}^{M-1} a_{ijk} s_j(t - D_{ij} - k) \quad (1)$$

(Here  $D_{ij}$  are entries in a matrix of delays and  $a_{ij}$  are the  $M$ -tap filter coefficients between the  $j$ th source and the  $i$ th sensor.) The problem is to invert this environmental scrambling without knowledge of it, thus recovering the original signals,  $\mathbf{s}(t)$ . The type of architecture that we choose for inverting eq.1 is important. An accurate architecture to invert a  $M$ -tap filter is an infinitive impulse response (IIR) filter with  $M$ -taps. However, IIR filters are limited to poles inside the unit circle and therefore, a stable IIR filter exists only for a minimum phase mixing system. FIR filters may be used to approximate the inverse solution. Figure 1 shows the mixing and unmixing system.

$$u_i(t) = \sum_{j=1}^N \sum_{k=0}^{M-1} w_{ijk} x_j(t - k). \quad (2)$$

In this, we have filters,  $w_{ij}$  and the time delays are seen as part of the deconvolution filters. The original uncorrupted source signals,  $s_i$  are reproduced by  $u_i$  when the learned system  $\mathbf{W}(z)$  is the inverse of the mixing system  $\mathbf{A}(z)$ . The design of  $\mathbf{W}(z)$  must allow for noncausal extension since the inverse of a nonminimum-phase system is noncausal.



**Figure 1. (a) The feedforward mixing / convolving system  $\mathbf{A}(z)$  and the inverting system  $\mathbf{W}(z)$  which is used to separate and deconvolve the signals  $\mathbf{x}$ . Each box represents a filter. (b) Entropy maximization at the output of the nonlinear neural processor. The nonlinearity  $g(u)$  can be fixed or can have a parametric form.**

## 3 Algorithm

Recently, several algorithms have been proposed for the blind separation of linear mixtures. Bell and Sejnowski [2] have proposed a simple infomax neural network algorithm where they maximize the joint entropy  $H(\mathbf{y})$  of an observation  $\mathbf{x}$  that has been linearly transformed and processed through a nonlinearity  $\mathbf{y} = g(\mathbf{u})$  with  $\mathbf{u} = \mathbf{W}\mathbf{x}$ . Pearlmutter and Parra [16] derive a similar learning rule from a ML density estimation using the Kullback-Leibler distance measure.

$$\begin{aligned} D(p, \hat{p}) &= \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\hat{p}(\mathbf{x}; \mathbf{w})} d\mathbf{x} \\ &= H(p(\mathbf{x})) - \int p(\mathbf{x}) \log \hat{p}(\mathbf{x}; \mathbf{w}) \end{aligned} \quad (3)$$

$p(\mathbf{x})$  is the probability density function (pdf) of the observation  $\mathbf{x}$  and  $\hat{p}(\mathbf{x}; \mathbf{w})$  is a parametric estimate of the distribution of the independent sources. Girolami and Fyfe [9] start from the negentropy point of view and use a kurtosis measure as projection pursuit.

$$N(p_{\mathbf{u}}) = H(p_G) - H(p_{\mathbf{u}}) \quad (4)$$

Negentropy can also be looked at the ML perspective where we measure the KL-distance of a transformed vector  $\mathbf{u}$  to

normality. Since the observation  $\mathbf{x}$  is close to the Gaussian distribution for a linear mixing of independent variables due to the central limit theorem, the difference between maximizing the distance to the observation or to a Gaussian distribution does not matter in practice. In [4] Cardoso shows that infomax and ML is equivalent because the relation between the KL-distance and the ML differs by the constant Entropy  $H(\mathbf{x})$  which is not dependent on  $\mathbf{W}$ .

$$L = - \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\hat{p}(\mathbf{x}; \mathbf{w})} d\mathbf{x} - H(p_{\mathbf{x}}) \quad (5)$$

In both approaches the output entropy  $H(\mathbf{y})$  of a neural processor is maximized which implies approximating the output density in the sense of minimum KL-distance, by a uniform density. This corresponds to producing a whitened signal with a flat amplitude spectrum at the output of the neural processor and at the same time making the input signals prior to the transfer function  $g(u)$  independent while shaping them according to the derivative  $\partial g(u)/\partial u$  with  $\mathbf{u} = \mathbf{W}\mathbf{x}$  being the estimate of the independent sources. This may be viewed as maximum entropy estimation of the input densities under the parameterization of  $\hat{p}_{\mathbf{x}}(\mathbf{x}; \mathbf{w})$ . We can relate  $\mathbf{x}$  to the nonlinear transfer function  $\frac{\partial g(u_i)}{\partial u_i}$  that gives us the pdf estimate  $\hat{p}_{\mathbf{x}}(\mathbf{x}; \mathbf{w})$ :

$$|\det J(\mathbf{x})| = |\det \mathbf{W}| \prod_{i=1}^n \frac{\partial g(u_i)}{\partial u_i} = \hat{p}_{\mathbf{x}}(\mathbf{x}; \mathbf{w}) \quad (6)$$

where  $J(\mathbf{x})$  is the Jacobian. The logarithmic representation is:

$$\log(\hat{p}_{\mathbf{x}}(\mathbf{x}; \mathbf{w})) = \log |\det \mathbf{W}| + \sum_{i=1}^n \log\left(\frac{\partial g(u_i)}{\partial u_i}\right) \quad (7)$$

Evaluating the expected value for eq.7 gives the output entropy which can be maximized with respect to  $\mathbf{W}$ .

$$\frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} = \mathbf{W}^{-T} + \left( \frac{\frac{\partial g(u_i)}{\partial u_i}}{g(u_i)} \right)_i \mathbf{x}^T \quad (8)$$

Considering the parameter  $\mathbf{W}$ , a better way to maximize entropy in the feedforward and feedback system is not to follow the entropy gradient, as in [2], but to follow its ‘natural’ gradient, as reported by Amari et al [1]:

$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} = \left[ \mathbf{I} + \left( \frac{\frac{\partial g(u_i)}{\partial u_i}}{g(u_i)} \right)_i \mathbf{u}^T \right] \mathbf{W} \quad (9)$$

This is an optimal rescaling of the entropy gradient. It simplifies the learning rule and speeds convergence considerably. The form of the nonlinearity  $g(u)$  plays an essential

role in the success of the algorithm. The ideal form for  $g(u)$  is the cumulative density function (cdf) of the distribution of the independent sources.  $g(u) = \int p_u(u) du$ . This leaves us some *degree of freedom* in choosing a differentiable nonlinear function that fits the unknown true distribution of  $s_i$ . If we choose  $g(u)$  to be a sigmoid function the learning rule reduces to the algorithm proposed in [2] and the algorithm is limited to the separation of super-Gaussian sources. A more accurate, but computational burdensome way is to use the contextual ICA [16] where the pdf is modeled in a parametric form and taking into account the temporal information. Pearlmutter and Parra choose to make  $p_i$  a weighted sum of logistic density functions with variable means and scales, and make these means linear functions of the recent history of source  $i$  as shown in figure 1(b).

$$p_i(u_i(t)|u(t-1), \dots; \mathbf{w}_i) = \sum_{k=1}^K \frac{m_{ik}}{\sigma_{ik}} \frac{\partial g(u_i)}{\partial u_i} \left( \frac{u_i(t) - \bar{u}_{ik}}{\sigma_{ik}} \right) \quad (10)$$

where  $m_{ik}$  are the mixing parameters and  $\sigma_{ik}$  are the scaling parameters.  $\partial g/\partial u = g(1-g)$  denotes the derivative of the logistic density function. The component means  $\bar{u}_{ik}$  are linear functions of the recent time samples of the source. The learning rules for the set of parameters to parameterize the density is given by the gradient ascent of the entropy.

$$\frac{\partial H(\mathbf{y})}{\partial \mathbf{w}_i} = \frac{\frac{\partial p_i(u_i; \mathbf{w}_i)}{\partial \mathbf{w}_i}}{p_i(u_i; \mathbf{w}_i)} \quad (11)$$

This allows the separation of sub/super-Gaussian and to some extent Gaussian distributions. In [14] we discuss several alternatives for the estimation of the underlying density.

An elegant way of generalizing the learning rule to sub/super-Gaussians is to approximate the estimated pdf in form of the Edgeworth approximation up to 4th order. This leads to a simple substitution as shown in [9].

$$\frac{\frac{\partial g(u_i)}{\partial u_i}}{g(u_i)} = -\text{sign}(k_4) \tanh(u_i) - u_i \quad (12)$$

For super Gaussians the kurtosis  $k_4$  is positive and the resulting term  $\tanh(\mathbf{u})\mathbf{u}^T$  in eq.12 corresponds to an anti-Hebbian rule whereas for a negative kurtosis the sources are sub Gaussian and the term becomes a Hebbian term.

It is interesting to note that the direct Bussgang property for blind deconvolution as well as the EASI algorithm by Cardoso [3] lead to a similar learning rule.

$$E\{u_{i+k}u_i\} = E\{g(u_i)u_{i+k}\} \\ \Delta \mathbf{W} \propto \mathbf{I} - \mathbf{u}\mathbf{u}^T - g(\mathbf{u})\mathbf{u}^T \quad (13)$$

## 4 FIR Polynomial Filter Design

The use of IIR filters is restricted to ARMA (autoregressive moving average) systems with minimum phase. Since we cannot obtain this prior knowledge about real recordings we have to assume a nonminimum phase system which may have a noncausal filter system inverse. For example, a non-minimum phase system will occur when a microphone picks up an echo that is stronger than the direct signal. The increase in negative phase is directly related to the amount of temporal delay of a narrowband component at that frequency. Hence, the minimum phase lag property or the minimum group delay property of a nonminimum-phase system is not guaranteed. However, any nonminimum or true phase system can be expressed as  $H(z) = H_{min}(z)H_{AP}(z)$  where  $H_{min}(z)$  is a minimum phase system and  $H_{AP}(z)$  is an *all-pass* system.  $H_{min}(z)$  has all its poles and zeros inside the unit circle and  $H_{AP}(z)$  represents a time delay with a unit frequency magnitude response. Therefore,  $H_{AP}(z)$  preserves the amplitude frequency spectrum and delays  $H(z)$  by reflecting the zeros outside the unit circle to their conjugate reciprocal location inside the unit circle. By time delaying the inverting system up to  $M/2$  taps,  $M$  being the size of the inverting filter, we introduce a  $M/2$  order  $H_{AP}(z)$  which is a technique to realize a noncausal system.

To implement such a system, the mixing and unmixing system in figure 1 can be written in the frequency domain representation where the elements of the matrices are filters and the multiplication operation replaces the convolution property. Lambert [11] has shown that FIR polynomial matrix algebra can be used as an efficient tool to elegantly solve problems for the multichannel source separation. The basic idea of using the FIR polynomial matrix algebra is to extend the algebra of scalar matrices to the algebra of matrices of filters (time-domain) or polynomials (freq. domain). The methods for computing functions of an FIR filter, such as an inverse, involve the formation of a circulant data matrix. Due to this nature we move to the frequency domain representation where eigencolumns of the circulant matrix are the discrete Fourier basis functions of the FFT of corresponding length. The filters now become polynomials of the Laurent series extension (z-transform) and the convolution / deconvolution of filters is reduced to multiplication / division of polynomials. For example, the inverse of a filter  $w(t)$  is such a computation and can be formulated as follows:

$$w(t)^{-1} = \text{FFTSHIFT}(\text{IFFT}(\text{FFT}[000 \cdots w(t) \cdots 000])) \quad (14)$$

The prepending of postpending of zeros is needed to produce a good estimate of the double-sided Laurent series expansion to allow for noncausal expansions of nonminimum

phase roots. The circular reordering in the time domain shifts the zeroth lag to the center of the filter. The complete proof is given in [11]. The learning rule for the two sources / two sensors problem can be reformulated from eq.9 as follows:

$$\begin{aligned} \Delta W(z) = & \left( \begin{bmatrix} \bar{1} & \bar{0} \\ \bar{0} & \bar{1} \end{bmatrix} - \begin{bmatrix} \text{FFT}(\hat{y}_1) \\ \text{FFT}(\hat{y}_1) \end{bmatrix} \begin{bmatrix} \text{FFT}(u_1) & \text{FFT}(u_1) \end{bmatrix}^* \right) \\ & \times \begin{bmatrix} W_{11}(z) & W_{21}(z) \\ W_{12}(z) & W_{22}(z) \end{bmatrix} \quad (15) \end{aligned}$$

Note that the neural processor  $\hat{y}_i = \frac{\partial}{\partial y_i} \frac{\partial y_i}{\partial u_i}$  still operates in the time domain and the FFT is applied at the output. \* denotes the complex conjugate form. Eq.15 is of the form of the least mean squared (LMS) adaptive filters. A fast implementation of the LMS adaptive filters in the frequency domain can be achieved by employing the *overlap and save* block LMS technique [8].

$$X(z) = \text{FFT}[x_{(k-1)n} \cdots x_{kn-1} x_{kn} \cdots x_{kn+n-1}] \quad (16)$$

For a blocksize of 1024 FFT-points the method is 16 times faster than the conventional LMS method [8].

## 5 Experiments with Real Recordings

We have conducted several experiments in a normal office room (3m x 4m) and a conference room (8m x 5.5m). The position of the two distant talking microphones and the location of the sources have been varied for each experiment. In the first set of experiments we have recorded one speaker saying the digits from one to ten while loud music was playing in the background. In this experimental setup the sources and sensors were placed in a rectangular (60cm x 40 cm) order with 60 cm distance between the sources and the sensors. Figure 2 (a) and (b) shows the recorded signals where the speech signal has been heavily corrupted by the music source. The algorithm converged after 30 epochs through a 7 sec. recording with 16kHz (120000 points). The unmixed signals have been obtained using 1024 taps FIR filters which cover a delay of 32ms corresponding to 10m. The separated signals are shown in figure 2 (c) and (d) A listening test shows a clean speech separation. The learned filters are shown in figure 3. In each filter, the leading tap is followed by a strong negative tap which indicates that the infomax algorithm tries to decorrelate adjacent time points. This whitening effect increases the energy in the higher frequency spectrum and reduces the energy of the lower frequency band. Speech signals sound sharper than their original. This effect can be compensated by postprocessing the

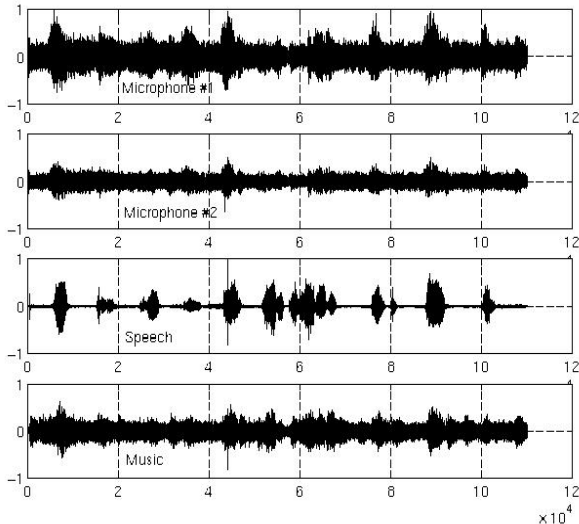


Figure 2. Microphone recordings of two speakers in a normal office room (a) Microphone 1 (b) Microphone 2. The separated signals are (c) speech and (d) music.

unmixed signals with a dewhiting filter. Another set of experiments have been performed with two speakers speaking simultaneously. Figure 4 (a) and (b) show the signals recorded with the same setup but with another speaker saying the digits one to ten in Spanish (*uno dos ... diez*) instead of the music source. The separated signals are shown in figure 4 (c) and (d). A listening test shows an almost clean speech separation. These audio-files are available in <http://www.cnl.salk.edu/~tewon/blind.html>. A prospective application is given in spontaneous speech recognition tasks where the best recognizer may fail completely in the presence of background music or competing speakers as in the teleconferencing problem. We have used an automatic speech recognition system trained on the Wall Street Journal task to test its performance on the recorded and separated signals. The recognition rates are listed in table 1. The results can

Table 1. Speech recognition results

Recog. rate	No. of words	mixtures	separation
Speech-Music	100	14 %	64 %
Speech-Speech	100	42 %	61 %
TOTAL	200	28 %	62.5 %

be further improved by postprocessing the separated signals, e.g., zeroing out the noisy part with a low signalpower de-

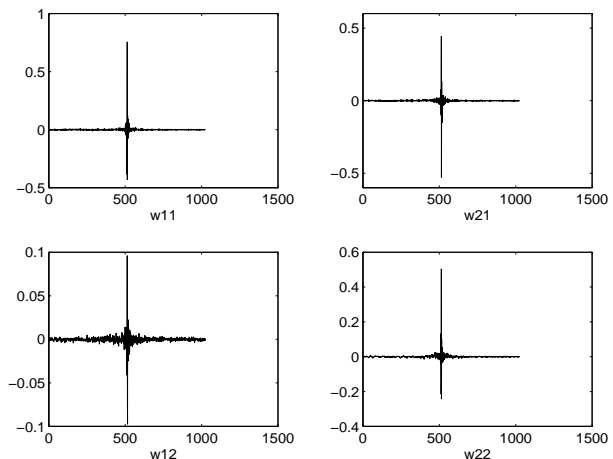


Figure 3. Unmixing and deconvolving FIR 1024-tap filters. Leading weights of the channel filters  $W_{11}$  and  $W_{22}$  are at 512-taps. Cross-channel filters are  $W_{21}$  and  $W_{12}$

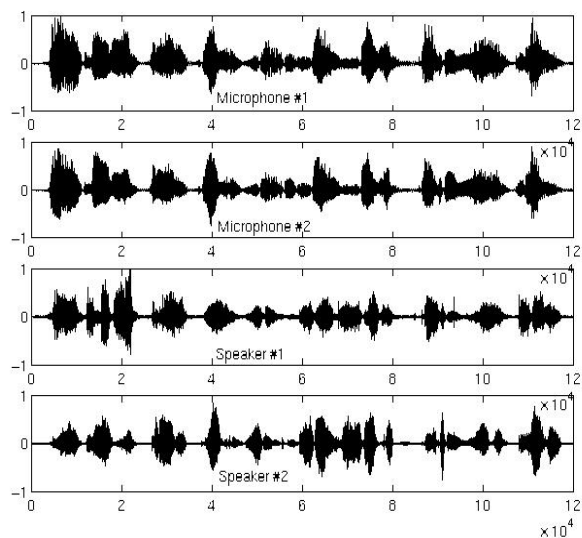
tector and by using a speech recognizer trained on digits.

## 6 Conclusions

We have presented a method to separate and deconvolve sources which have been recorded in real environments. The use of noncausal FIR filters allows us to deal with non-minimum mixing systems. The learning rules can be derived from different perspectives and such as information maximization, maximum likelihood and negentropy which result in similar rules for the weight update. We transform the learning rules into the frequency domain where the convolution and deconvolution property becomes a multiplication and division operation. In particular, the use of FIR polynomial algebra techniques present an efficient tool to solve true phase inverse systems allowing a simple implementation of noncausal filter solutions. The significance of the methods is shown by the successful separation of two voices and separating a voice that has been recorded with loud music in the background. The recognition rate of an automatic speech recognition system is increased after separating the speech signals.

## Acknowledgments

T.W.L. is supported by the Daimler-Benz-Fellowship. A.J.B. is supported by a grant from the Office of Naval Research. We are grateful to Russ Lambert, Juan Huerta, the



**Figure 4. Microphone recordings of two speakers in a normal office room (a) microphone 1 and (b) microphone 2. The separated speakers are shown in (c) Spanish digits *uno dos ... diez* and (d) English digits *one two ... ten*.**

interactive systems group at CMU and Terrence Sejnowski for discussions and comments.

## 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*, 1996.
- [2] A. Bell and T. Sejnowski. An Information Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7:1129–1159, July 1995.
- [3] J-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 45,2:434-444, Dec. 1996.
- [4] J-F. Cardoso. Infomax and maximum likelihood for blind source separation. to appear in *IEEE Signal Processing Letters*.
- [5] A. Cichocki, R. Unbehauen, and E. Rummert. Robust learning algorithm for blind separation of signals. *Electronics Letters*, 30, 17, 1386-1387, 1994
- [6] A. Cichocki, S. Amari and J. Cao. Blind separation of delayed and convolved signals with self-adaptive learning rate. In *Proc. NOLTA'96*, 1996.
- [7] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36(3):287–314, 1994.
- [8] E. Ferrara. Fast implementation of lms adaptive filters. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):474–478, 1980.
- [9] M. Girolami and C. Fyfe. Negentropy and kurtosis as projection pursuit indices provide generalised ica algorithms. In *Advances in Neural Information Processing Systems Workshop 9*, 1996.
- [10] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo. A class of neural networks for independent component analysis. Report A28, Helsinki Univ. of Technology, October 1995. submitted to a journal.
- [11] R. Lambert. Multichannel blind deconvolution: Fir matrix algebra and separation of multipath mixtures. Thesis, University of Southern California, Department of Electrical Engineering, May 1996.
- [12] R. Lambert and A. Bell. Blind separation of multiple speakers in a multipath environment. in *Proc. ICASSP 1997*, Munich.
- [13] T. Lee, A. Bell and R. Lambert. Blind separation of convolved and delayed sources. In *Advances in Neural Information Processing Systems 9*. MIT Press, 1997.
- [14] T. Lee and R. Orglmeister. A contextual blind separation of delayed and convolved sources. In *Proc. ICASSP 1997*, Munich.
- [15] D.T. Pham. Blind separation of instantaneous mixture of sources via an independent component analysis. *IEEE Trans. on Signal Proc.*, 44, 11:2768-2779, 1996.
- [16] B. Pearlmutter and L. Parra. A context-sensitive generalization of ICA. In *ICONIP'96*. In press.
- [17] Z. Roth and Y. Baram. Multidimensional density shaping by sigmoids. *IEEE Trans. on Neural Networks*, 7(5):1291–1298, 1996.
- [18] Kari Torkkola. Blind separation of convolved sources based on information maximization. In *IEEE Workshop on Neural Networks for Signal Processing*, Kyoto, Japan, September 4-6 1996. (in press).
- [19] D. Yellin and E. Weinstein. Multichannel signal separation: Methods and analysis. *IEEE Transactions on Signal Processing*, 44(1):106–118, January 1996.