IEEE
International Conference on
Acoustics, Speech and Signal
Processing

BLIND SEPARATION AND BLIND DECONVOLUTION: AN INFORMATION-THEORETIC APPROACH

Anthony J. Bell and Terrence J. Sejnowski

Computational Neurobiology Laboratory, The Salk Institute,
10010 N. Torrey Pines Road, La Jolla, California 92037

1. INTRODUCTION AND ABSTRACT

Blind separation and blind deconvolution are related problems in *unsupervised learning*. In blind separation [7], illustrated in Fig.1a, a set of sources, $s_1(t), \ldots, s_N(t)$, (different people speaking, music etc) are mixed together linearly by a matrix $\mathbf{a}$. We do not know anything about the sources, or the mixing process. All we receive are the N superpositions of them, $x_1(t), \ldots, x_N(t)$. The task is to recover the original sources by finding a square matrix $\mathbf{W}$ which is a permutation of the inverse of the unknown matrix, $\mathbf{A}$. The problem has also been called the 'cocktail-party' problem.

In blind deconvolution [6], illustrated in Fig.1b, an unknown signal $s(t)$ is convolved with an unknown tapped delay-line filter, $a_1, \ldots, a_L$, giving a corrupted signal $x(t) = a(t) * s(t)$ where $a(t)$ is the impulse response of the filter. The task is to recover $s(t)$ by convolving $x(t)$ with a learnt filter $w_1, \ldots, w_L$ which reverses the effect of the filter a.

Both problems are difficult. In the case of blind separation, the approach is generally to assume that the sources, $\mathbf{s}$, are statistically independent and non-gaussian, then the problem of learning $\mathbf{W}$ becomes the problem of independent component analysis (ICA), [5]. In the case of blind deconvolution, the approach is often to assume that the original signal $s(t)$ consisted of independent symbols (a white process), then deconvolution becomes the problem of removing from $x(t)$ any statistical dependencies across time, introduced by the corrupting filter a. This process is sometimes called the *whitening* of $x(t)$.

Both ICA and whitening require higher-order statistics. Only for gaussian signals is second-order decorrelation sufficient. Such higher-order statistics can come from the explicit estimation of cumulants and polyspectra [5, 6] or from the usage of static non-linearities in the stochastic weight update algorithm [3, 7]. The Taylor series expansions of these non-linearities produce higher-order moments. The particular higher-order moments produced are not rigorously related to those required to exactly calculate statistical dependencies, but nonetheless, these techniques have met with some success.

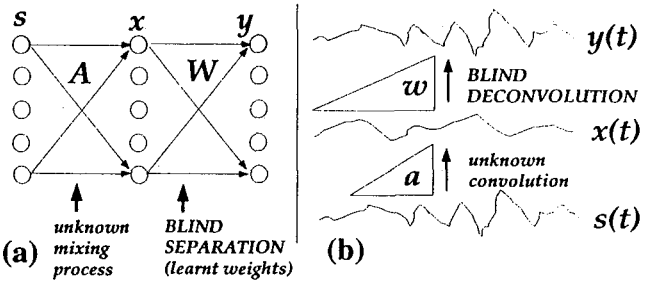

Figure 1: (a) Blind separation of 5 mixed signals. (b) Blind deconvolution of a single signal.

In this contribution, static non-linearities *are* used in combination with an information-theoretic objective function, making the approach more rigorous than previous ones. We derive a new algorithm and with it perform nearly perfect separation of up to 10 digitally mixed human speakers, better performance than any previous algorithms for blind separation. When used for deconvolution, the technique automatically cancels echoes and reverberations and reverses the effects of low-pass filtering.

2. ENTROPY MAXIMISATION.

The problem of reversing the convolution or static mixing of a signal or signals, $\mathbf{s}$, is construed here as a maximisation of the entropy $H(\mathbf{y})$ of a non-linearly transformed signal $\mathbf{y} = g(\mathbf{x})$ where g is some function. Consider the joint entropy of two components of $\mathbf{y}$ (either two output channels in the case of separation, or two time points in the case of deconvolution):

$$H(y_1, y_2) = H(y_1) + H(y_2) - I(y_1, y_2) \qquad (1)$$

Maximising this joint entropy consists of maximising the individual entropies while minimising the mutual

0-7803-2431-5/95 $4.00

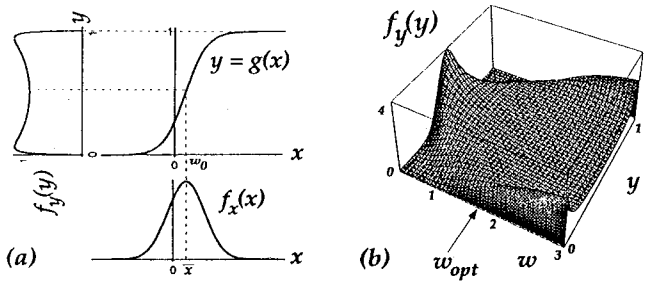

Figure 2: (a) Optimal information flow through a squashing function (Schraudolph et al 1992). Slope and threshold may be adjusted to maximise the entropy of $f_y(y)$. (b) $f_y(y)$ is plotted for different values of the weight w. The optimal weight, w_{opt} transmits most information.

information, $I(y_1, y_2)$, shared between the two. When this latter quantity is zero, the two variables are statistically independent. Both ICA and the 'whitening' approach to deconvolution are examples of minimising $I(y_1, y_2)$ for all pairs y_1 and y_2. This process is also known as *redundancy reduction* (Barlow 1989).

By maximising $H(\mathbf{y})$, we will, in general, reduce $I(\mathbf{y})$. However there are cases where the absolute minimum will not be reached, because of interference from the other terms, $H(y_i)$. When g is the hyperbolic tangent (or logistic) non-linearity, and the sources have super-gaussian (kurtosis > 3) distributions, we have not found this to be a problem. However, on data other than speech, it may well pose problems.

We describe how to maximise $H(\mathbf{y})$, firstly for blind separation.

3. BLIND SEPARATION.

When g is a linear function (ie: $y_i = \sum_{j=1}^{N} w_{ij} x_j$), $H(\mathbf{y})$ can be increased merely by increasing the variances of the outputs without bound. Consider, however, the case when g consists of a linear combination followed by a non-linear squashing function, such as

$$\mathbf{y} = g(\mathbf{x}) = \tanh(\mathbf{W}\mathbf{x} + \mathbf{w}_0) \tag{2}$$

where $\mathbf{w}_0$ is a vector of 'bias' weights. In this case, too large a weight matrix $\mathbf{W}$ will lead to saturation of the outputs at ± 1. Too small a weight matrix will lead to outputs clustered around zero. Both extremes yield a low entropy probability density function (pdf) for $\mathbf{y}$. In other words,

$$H(\mathbf{y}) = -E[\ln f_{\mathbf{y}}(\mathbf{y})] \tag{3}$$

will be small. Fig.2a shows the highest entropy pdf of a single squashed output y for a single input x and a single weight w. In Fig.2b, a family of such pdfs is shown for different values of w. To find the optimal w's (those which maximise $H(\mathbf{y})$), we use the following equation, relating the multivariate density functions of x and $\mathbf{y}$ (Papoulis 1984):

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{f_{\mathbf{x}}(\mathbf{x})}{|J|} \tag{4}$$

where $|J|$ is the absolute value of the Jacobian of the transformation. The Jacobian is the determinant of the matrix of partial derivatives, $[\partial y_i / \partial x_j]_{ij}$. For a linear transformation followed by a non-linear function,

$$J = (\det \mathbf{W}) \prod_{i=1}^{N} y_i' \tag{5}$$

where y_i' is the slope of the ith output function. For (2), $y_i' = 1 - y_i^2$. Substituting (4) into (3) gives:

$$H(\mathbf{y}) = E[\ln |J|] + H(\mathbf{x}). \tag{6}$$

To maximise this by changing $\mathbf{W}$, we need only concentrate on the first term, which can be maximised using a stochastic gradient ascent rule, utilising (5):

$$\Delta \mathbf{W} \propto \frac{\partial}{\partial \mathbf{W}} \ln |J| = \frac{\partial}{\partial \mathbf{W}} \left(\ln |\det \mathbf{W}| + \sum_{i=1}^{N} \ln |y_i'| \right) \tag{7}$$

This yields the simple rules for $\mathbf{W}$ and $\mathbf{w}_0$:

$$\Delta \mathbf{W} \propto [\mathbf{W}^T]^{-1} - 2\mathbf{y}\mathbf{x}^T \tag{8}$$

$$\Delta \mathbf{w}_0 \propto -2\mathbf{y} \tag{9}$$

The $\Delta \mathbf{w}_0$ rule centres the tanh function on the input data $\mathbf{x}$. The first term in the $\Delta \mathbf{W}$ rule forces the outputs $\mathbf{y}$ to represent different components of the inputs (because any degenerate $\mathbf{W}$ will be an unstable point of the learning dynamics). The second term (an *anti-Hebbian* term) stops saturation. These opposing forces are exactly balanced to make the outputs find independent components in the inputs, and adjust their weight vectors to achieve maximum entropy along each component [see (1)]. Because $\mathbf{y}$ is a non-linear function of the inputs, the anti-Hebbian term in (8) supplies the higher-order cross-moments necessary to find the independent components (see discussion in [4, 5, 2]). The difference with the Herault-Jutten rule [7], is that the particular cross-moments used are well-motivated by our information-theoretic objective.

4. BLIND DECONVOLUTION.

The same reasoning as above can be applied to the blind deconvolution problem. If we apply a static non-linear function g to produce $y(t)$ in Figure 1(b), then we

can write the system either as a cascade of convolutions or as a matrix equation:

$$y(t) = g(w(t) * x(t)) = g(w(t) * a(t) * s(t)) \quad (10)$$

$$\mathbf{y} = g(\mathbf{W}\mathbf{x}) = g(\mathbf{W}\mathbf{A}\mathbf{s}) \quad (11)$$

where $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{s}$ are vectors corresponding to time series, and $\mathbf{A}$ and $\mathbf{W}$ are matrices. When the filtering is causal, $\mathbf{A}$ and $\mathbf{W}$ will be lower triangular (Toeplitz). Because $\mathbf{A}^{-1}$ will also be causal, we can invert the effect of $\mathbf{A}$ with $\mathbf{W}$. We can apply the same strategy as with blind separation, namely to maximise the entropy of the output vector $\mathbf{y}$. Analogously to (7), this corresponds to maximising, by adjusting $w(t)$, the log of the absolute value of the Jacobian of the transformation from $\mathbf{x}$ to $\mathbf{y}$:

$$\ln|J| = \ln|\det \mathbf{W}| + \sum_{t=1}^{M} \ln|y'(t)|. \quad (12)$$

Here we sum over a time series of length M, and $y'(t)$ is the slope of the squashing function, g, with respect to its input at time t. The size of M is chosen so as to be larger than the length of the filter $w(t)$ but smaller than the length of the whole time series so that we have an ensemble of time series and may justify the maximisation of the entropy of the random vector $\mathbf{y}$. Because $\mathbf{W}$ is lower-triangular, its determinant is simply the product of its diagonal values, which amounts to w_L^M. This leads to the following simple[1] rules for changing weights, when $g = \tanh$:

$$\Delta w_L \propto \sum_{t=1}^{M} \left(\frac{1}{w_L} - 2x_t\, y_t \right) \quad (13)$$

$$\Delta w_{L-j} \propto \sum_{t=j}^{M} (-2x_{t-j}\, y_t) \quad (14)$$

Here, w_L is the 'leading' weight, and the w_{L-j}, where $j > 0$, are tapp-ed delay lines linking x_{t-j} to y_t. In these rules, only the leading weight adjusts with a weight-normalisation term, while the delay weights attempt to simply decorrelate the past input from the present output. This latter process is what enables this technique to reverse the effect of low-pass filtering, and to remove echoes and reverberations from time series.

5. RESULTS.

Results were obtained using 7 second speech segments from various speakers. All signals were sampled at 8kHz from the output of the auxiliary microphone of a Sparc-10 workstation, and the waveforms were normalised to lie roughly between -3 and 3. Training of the weights was done with a 'small-batch' variant of stochastic gradient ascent: in which weights were adjusted every 100 or so presentations of time points. To ensure stationarity in the training sequence, the time index was permuted before training. An important factor was the choice of learning rates—the proportionality constants in (8)-(9) and (13)-(14). A typical value was 0.001. It was helpful to reduce the learning rate during learning for convergence to good solutions.

[1] The corresponding rules for *non*-causal filters are substantially more complex.

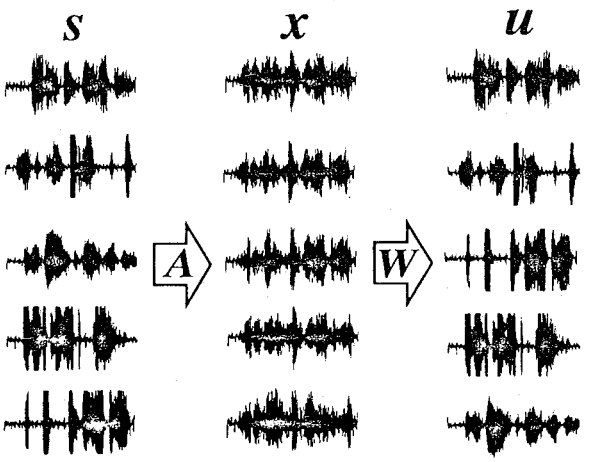

Figure 3: A 5 × 5 information maximisation network performed blind separation, learning the unmixing matrix $\mathbf{w}$. The outputs, $\mathbf{u} = g^{-1}(\mathbf{y})$, are shown here unsquashed by the sigmoid. They can be visually matched to their corresponding sources, $\mathbf{s}$, even though their order was different and some (for example u_1) were recovered as negative (upside down).

5.1. Blind separation results.

The architecture in Fig.1a and the algorithm in (8)-(9) was sufficient to perform blind separation. A random mixing matrix, $\mathbf{A}$, was generated with values usually uniformly distributed between -1 and 1. This was used to make the mixed time series, $\mathbf{x}$ from the original sources, $\mathbf{s}$. The unmixing matrix, $\mathbf{W}$, and the bias vector $\mathbf{w}_0$ were then trained.

For two sources, convergence is normally achieved in less than one pass through the data. An example run with five sources is shown in Fig.3. The mixtures, $\mathbf{x}$, formed an incomprehensible babble. This unmixed solution was reached after around 10^6 time points were presented, equivalent to about 20 passes through the

complete time series.[2] Any residual interference in **Wx** was inaudible, and the matrix **WA** was, on inspection, a permutation and rescaling of the identity matrix. In our most ambitious attempt, ten sources (six speakers, rock music, raucous laughter, a gong and the Hallelujah chorus) were successfully separated, though this took many hours.

In all our attempts at blind separation, the algorithm has only failed when more than one of the sources were gaussian white noise or when the mixing matrix, **A**, was almost singular, both pathological conditions for blind separation algorithms.

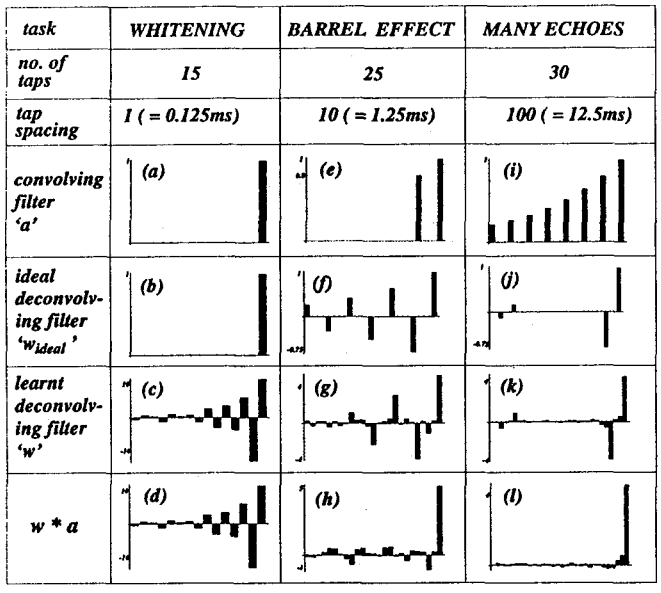

Figure 4: Blind deconvolution results. (a, e, i) Filters used to convolve speech signals, (b, f, j) their inverses, (c, g, k) deconvolving filters learnt by the algorithm, and (d, h, l) convolution of the convolving and deconvolving filters. See text for further explanation.

5.2. Blind deconvolution results.

Figure 4 shows various filters, their ideal inverses and the filters learnt by iteration of (13) and (14) using convolved speech signals. When filter taps were adjacent as in Fig.4a-d, the system learnt a whitening filter, Fig.4c, which flattened the amplitude spectrum up to the Nyquist limit of 4kHz. Otherwise, when the taps were spaced further apart, and the speech was convolved with a finite filter, Fig.4e-h, or a truncated infinite filter, Fig.4i-l, good approximations to the inverse filters were learnt, without interference from 'whitening' effects. The learning was sensitive enough to replicate, in Fig.4k, the correction term on the left of Fig.4j, caused by truncation of the exponentially decaying echoes in Fig.4i.

We have also combined the blind separation and deconvolution techniques, and performed simultaneous unmixing and deconvolution. More details appear in [2].

[2]This took on the order of 5 minutes on a Sparc-10, using efficient vectorised Matlab © code.

6. ACKNOWLEDGEMENTS.

Many thanks to Nici Schraudolph, who initially described the idea in Fig.2a, shared his unpublished calculations [9], and provided detailed criticism. Constructive observations also came from Paul Viola, Barak Pearlmutter, Kenji Doya, Alex Pouget and Simon Haykin.

7. REFERENCES

[1] Barlow H.B. 1989. Unsupervised learning, *Neural Computation*, 1, 295-311

[2] Bell A.J. & Sejnowski T.J. 1995. An information maximisation approach to blind separation and blind deconvolution, *Neural computation*, in press

[3] Bellini S. 1994. Bussgang techniques for blind deconvolution and equalisation, in [6]

[4] Comon P., Jutten C. & Herault J. 1991. Blind separation of sources, part II: problems statement, *Signal Processing*, 24, 11-21

[5] Comon P. 1994. Independent component analysis, a new concept? *Signal Processing*, 36, 287-314

[6] Haykin S. (ed.) 1994. *Blind Deconvolution*, Prentice-Hall, New Jersey.

[7] Jutten C. & Herault J. 1991. Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture, *Signal Processing*, 24, 1-10

[8] Papoulis A. 1984. *Probability, random variables and stochastic processes, 2nd edition*, McGraw-Hill, New York

[9] Schraudolph N.N., Hart W.E. & Belew R.K. 1992. Optimal information flow in sigmoidal neurons, *unpublished manuscript*