Proceedings of the IEEE International Conference on Acoustics, Speech & Signal Processing, May 7-10 1996, Atlanta, GA, USA

BLIND SEPARATION OF DELAYED SOURCES BASED ON INFORMATION MAXIMIZATION

Kari Torkkola
Motorola, Inc., Phoenix Corporate Research Laboratories,
2100 E. Elliot Rd, MD EL508, Tempe, AZ 85284, USA
tel: (602)413-4129, fax: (602)413-5934, email: A540AA@email.mot.com

ABSTRACT

Recently, Bell and Sejnowski have presented an approach to blind source separation based on the information maximization principle. We extend this approach into more general cases where the sources may have been *delayed* with respect to each other. We present a network architecture capable of coping with such sources, and we derive the adaptation equations for the delays and the weights in the network by maximizing the information transferred through the network. Examples using wideband sources such as speech are presented to illustrate the algorithm.

1 INTRODUCTION

Blind Source Separation (BSS) denotes observing mixtures of *independent* sources, and by making use of these mixture signals only and nothing else, recovering the original signals.

A BSS system that works under real world conditions would have a multitude of applications: antenna array processing, speech signal enhancement in the presence of noise, speaker separation in teleconferencing situations (cocktail party problems), elimination of multipath or reverberating distortion, reducing intersymbol interference, and others. As we point out later, current algorithms have some shortcomings that prevent their successful application to many real world situations, such as to wideband audio recordings.

BSS usually assumes a linear mixture model, $x(t) = As(t)$, where $s(t)$ is the vector of sources at instant t, A is the mixing matrix, and $x(t)$ is the observed vector of mixtures. Presently, we ignore any noise in the mixtures. The sources can be separated by finding a matrix W with the condition $WA = PD$, where P is a permutation matrix and D is a diagonal matrix. Thus the output becomes $u(t) = Wx(t) = PDs(t)$. The sources will be separated but they may be permuted and scaled.

Learning W by observing the x only requires making use of higher order statistics (HOS) [6, 5, 1]. Typically, the weight matrix is computed iteratively by optimizing some sort of a contrast function of the output u. This function usually involves fourth order cumulants. Approaches can be divided into two categories, those that take some aspects of HOS into account implicitly through nonlinear functions of the output [9, 2, 14, 10, 4, 7], and those that do it explicitly [3, 8, 19, 13]. The former category includes many algorithms that are "neural" in nature. Another approach is to use generalized cross-correlations [17, 12, 15].

Bell and Sejnowski have recently showed that another possible criterion for BSS is mutual information [1]. Combined with a nonlinear squashing function normally seen in neural network contexts this criterion seems to provide a theoretical framework for approaches of the first category above.

We will first review the approach of Bell and Sejnowski, and point out shortcomings of this algorithm (and the others!) as applied to real word source separation problems. Especially delays between sources in mixtures will be discussed in more detail.

We then extend the information maximization approach into more general cases where the sources may have been delayed with respect to each other. We present a network architecture capable of coping with such sources, and derive the adaptation equations for the delays and the weights in the network on the basis of Bell's and Sejnowski's information maximization principle. Examples using speech sources are presented to illustrate the performance of the algorithm.

2 INFORMATION MAXIMIZATION

Bell and Sejnowski proposed to learn W by minimizing the mutual information between components of $y(t) = g(u(t))$, where g is a nonlinear function approximating the cumulative density function (cdf) of the sources (see Fig. 1 for a two-source network) [1].

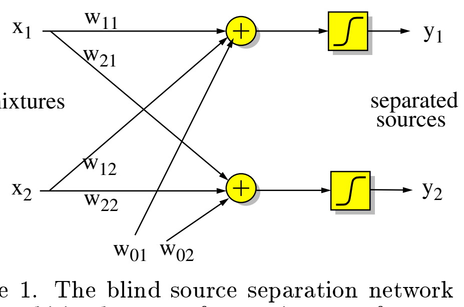

Figure 1. The blind source separation network of Bell and Sejnowski in the case of two mixtures of two sources.

Bell and Sejnowski showed that for positively kurtotic signals (such as speech) minimizing the mutual information between components of y is equal to maximizing the entropy of y, which can be written as $H(y) = -E[ln(f_y(y))]$, where $f_y(y)$ denotes the probability density function of y. Denoting the determinant of the Jacobian of the whole net by $|J|$, $f_y(y)$ can be written as $f_x(x)/|J|$ (Jacobian is a matrix with entries of $\partial y_i/\partial x_j$.) Now, maximizing the entropy of the output leads to maximizing $E[ln|J|]$, which in turn can be developed into the following stochastic gradient ascent rule using instances of $x(t)$ and $y(t)$, instead of using the expectation:

$$\begin{aligned} \Delta W &\propto (1-2y(t))x(t)^T + [W^T]^{-1} \ , \\ \Delta w_0 &\propto 1-2y(t). \end{aligned} \tag{1}$$

To derive the adaptation rule above, $g(y) = 1/(1+e^{-y})$ was used to approximate the cdf. Bell and Sejnowski demonstrated the efficiency of this rule in the separation of as many as ten artificially mixed audio sources.

3 SHORTCOMINGS OF CURRENT ALGORITHMS

As pointed out in [15] and [1], there are several obstacles to applying these algorithms to real world situations. These include the effect of noise to the learning of the separating solution, possibly unknown number of sources (especially noise sources), and the assumption that the source signals are stationary. Most notable is the assumption of the simultaneous mixing of the sources. In any real-world recording, where the propagation of the signals through the medium is not instantaneous (like sound through air or water), there will be delays between the sources in the mixtures. Current BSS algorithms cannot tolerate these delays. A related problem is the multipath arrival of the signals. In general, the sensor is not observing just a clean copy of the source, but a sum of multipath copies distorted by the environment, for example, by the impulse response of a room.

Some of the literature addresses partly the delay and multipath problems using various different methods [14, 8, 19, 13, 11]. No general and rigorous solution exists so far.

Bell and Sejnowski [1] also suggested a learning rule for just a single weight with a delay, but they did not present a general solution to the separation of delayed sources. However, Bell's approach of using information-theoretic principles in the derivation of the adaptation rules is very appealing. In this paper, we will show that the delay problem is indeed solvable using the same approach.

4 AN ARCHITECTURE FOR DELAYS

Let us assume for simplicity two sources and their two mixtures (this can easily be generalized to any number of sources):

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t-D_{12}) \ , \\ x_2(t) &= a_{22}s_2(t) + a_{21}s_1(t-D_{21}). \end{aligned} \tag{2}$$

The delayed sources can be separated by a feedback architecture using adaptive delays. This is the same archtecture which was already proposed in [9], to which Platt and Faggin later added adaptive delays [14]. We propose to use the same feedback principle, but with the information-maximization principle of Bell and Sejnowski. Figure 2. presents the resulting network.

The network thus computes the following, where u_i are the outputs before the nonlinearities, and w_{0i} are the bias weights:

$$\begin{aligned} u_1(t) &= w_1x_1(t) + w_{12}u_2(t-d_{12}) + w_{01}, \\ u_2(t) &= w_2x_2(t) + w_{21}u_1(t-d_{21}) + w_{02}, \\ y_1(t) &= g(u_1(t)), \\ y_2(t) &= g(u_2(t)). \end{aligned} \tag{3}$$

As with Bell's feedforward case, we can maximize the entropy at the output by maximizing $E[ln|J|]$ with respect to the parameters of the network, that is, the weights *and the delays*. Stationarity of the sources is not required.

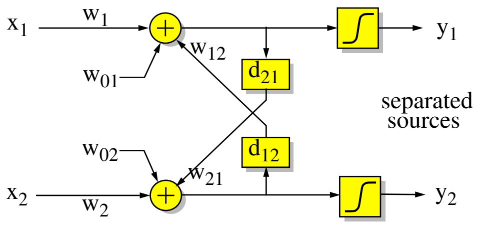

Figure 2. The separation network for delays.

5 ADAPTATION RULES

We will derive adaptation rules in the time domain for the case of two sources. Generalization to multiple sources is trivial as discussed at the end of this section. The starting point will be Fig. 2 and equation 3, where g is the logistic function $g(u) = 1/(1+e^{-u})$. Following Bell and Sejnowski, we can minimize the mutual information between outputs y_1 and y_2 by maximizing the entropy at the outputs, which is equal to maximizing $E[ln|J|]$. The determinant of the Jacobian of the network is now

$$\begin{aligned} |J| &= \frac{\partial y_1}{\partial x_1}\frac{\partial y_2}{\partial x_2} - \frac{\partial y_1}{\partial x_2}\frac{\partial y_2}{\partial x_1} = y_1' y_2' D \ and \\ ln|J| &= ln(y_1') + ln(y_2') + ln(D). \end{aligned} \tag{4}$$

where $D = (\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}) = w_1 w_2$, $y_1' = \frac{\partial y_1}{\partial u_1}$, and $y_2' = \frac{\partial y_2}{\partial u_2}$.

The adaptation rule for each parameter of the network can now be derived by computing the gradient of $ln|J|$ with respect to that parameter. For w_1 we get

$$\Delta w_1 \propto \frac{\partial ln|J|}{\partial w_1} = \frac{1}{y_1'}\frac{\partial y_1'}{\partial w_1} + \frac{1}{y_2'}\frac{\partial y_2'}{\partial w_1} + \frac{1}{D}\frac{\partial D}{\partial w_1} \tag{5}$$

For the logistic function $\frac{\partial y_i'}{\partial y_i} = 1 - 2y_i$. Thus we can write for the partial derivatives:

$$\begin{aligned} \frac{\partial y_1'}{\partial w_1} &= \frac{\partial y_1'}{\partial y_1}\frac{\partial y_1}{\partial u_1}\frac{\partial u_1}{\partial w_1} = (1-2y_1)y_1' x_1 \ , \\ \frac{\partial y_2'}{\partial w_1} &= \frac{\partial y_2'}{\partial y_2}\frac{\partial y_2}{\partial u_2}\frac{\partial u_2}{\partial w_1} = (1-2y_2)y_2' 0 = 0 \ , \\ \frac{\partial D}{\partial w_1} &= \frac{\partial (w_1 w_2)}{\partial w_1} = w_2. \end{aligned} \tag{6}$$

The adaptation rule for w_1 becomes the following from eq. 5 (similarly for w_2):

$$\begin{aligned} \Delta w_1 &\propto (1-2y_1)x_1 + 1/w_1, \\ \Delta w_2 &\propto (1-2y_2)x_2 + 1/w_2. \end{aligned} \tag{7}$$

The bias adaptation will be the same as in [1], namely $\Delta w_{0i} \propto 1 - 2y_i$. The role of these weights and biases is to scale and to shift the data so as to maximize the information passed through the sigmoid g.

For w_{12} the partial derivatives are as follows:

$$\begin{aligned}
\frac{\partial y_1'}{\partial w_{12}} &= \frac{\partial y_1'}{\partial y_1}\frac{\partial y_1}{\partial u_1}\frac{\partial u_1}{\partial w_{12}} = (1-2y_1)y_1' u_2(t-d_{12}),\\
\frac{\partial y_2'}{\partial w_{12}} &= \frac{\partial y_2'}{\partial y_2}\frac{\partial y_2}{\partial u_2}\frac{\partial u_2}{\partial w_{12}} = (1-2y_2)y_2' 0 = 0,\\
\frac{\partial D}{\partial w_{12}} &= \frac{\partial (w_1 w_2)}{\partial w_{12}} = 0. \qquad (8)
\end{aligned}$$

Thus the adaptation for w_{12} will be the following (similarly for w_{21}):

$$\begin{aligned}
\Delta w_{12} &\propto (1-2y_1)u_2(t-d_{12}),\\
\Delta w_{21} &\propto (1-2y_2)u_1(t-d_{21}). \qquad (9)
\end{aligned}$$

These rules decorrelate the present squashed output y_i from the other *source* u_j at delay d_{ij}, which is equivalent to separation. Note that in equations 8 and 9 the time indices of u_1 and u_2 are given in parentheses, whereas for all other variables the time is implicitly assumed to be t. All the partial derivatives starting from eq. 4 are also taken at time instance t. *This is why we do not expand the cross partial derivatives recursively backwards in time!* (Doing this would be fairly straightforward by introducing recursive intermediate quantities in the fashion of LMS-trained recursive adaptive filters [18], but in our experiments the simplified adaptation converged better.)

The partial derivatives for the delay d_{12} will be:

$$\begin{aligned}
\frac{\partial y_1'}{\partial d_{12}} &= \frac{\partial y_1'}{\partial y_1}\frac{\partial y_1}{\partial u_1}\frac{\partial u_1}{\partial d_{12}} = (1-2y_1)y_1' w_{12}(-\dot{u}_2(t-d_{12})),\\
\frac{\partial y_2'}{\partial d_{12}} &= \frac{\partial y_2'}{\partial y_2}\frac{\partial y_2}{\partial u_2}\frac{\partial u_2}{\partial d_{12}} = (1-2y_2)y_2' 0 = 0,\\
\frac{\partial D}{\partial d_{12}} &= \frac{\partial (w_1 w_2)}{\partial d_{12}} = 0, \qquad (10)
\end{aligned}$$

where we have taken advantage of the fact that $\frac{\partial u_2(t-d_{12})}{\partial d_{12}} = \frac{d}{dt}(-u_2(t-d_{12})) = -\dot{u}_2(t-d_{12})$ [1].

The adaptation rules for the delays become the following (Again, we have explicitly written the time indices only for $\dot{u}_i$):

$$\begin{aligned}
\Delta d_{12} &\propto -(1-2y_1)w_{12}\dot{u}_2(t-d_{12}),\\
\Delta d_{21} &\propto -(1-2y_2)w_{21}\dot{u}_1(t-d_{21}). \qquad (11)
\end{aligned}$$

It is remarkable that every adaptation rule is *local*, that is, to adapt a weight or a delay in a branch of the network, only the data coming in or going out of the branch are needed. For example, adaptation of d_{12} and w_{12} will only depend on u_2 and y_1. Generalization to N mixtures can thus be done simply by substituting other indices to 1 and 2 in equations 9 and 11, and by summing such terms.

It must also be noted that the values of the delays need to be positive. The zero delay case would be equal to the instantaneous mixture case with the solution of Bell and Sejnowski in equation 1.

6 EXPERIMENTS WITH SPEECH

We illustrate the behavior of the resulting adaptation rule using speech examples. In the first example, two sources are mixed according to

$$\begin{aligned}
x_1(t) &= s_1(t) + 0.4 s_2(t-10)\ ,\\
x_2(t) &= s_2(t) + 0.8 s_1(t-20). \qquad (12)
\end{aligned}$$

Figure 3. displays the convergence of the delays and the feedback weights. We can see that the delays converge almost immediately, in less than 40 iterations. Only after this the weights start to converge towards the separating solution (-0.4, -0.8). About 200 iterations are needed to get them to converge. The learning rates were kept fixed in this experiment.

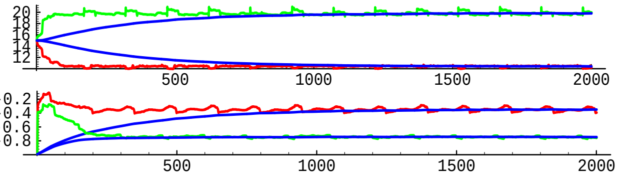

Figure 3. Convergence of the delays (top) and the feedback weights (bottom). The horizontal axis is the number of iterations, where each iteration consists of the presentation of 250 samples of the mixtures. The vertical axis of the top graph is the amount of delay given in samples. In the bottom graph the vertical axis is the value of the weight w_{ij}. d_{12} and w_{12} are plotted with a darker shade, d_{21} and w_{21} with a lighter shade. A (very) slow moving time average of the parameters is also plotted with black.

We do not need to assume that the sources are stationary, and thus we present the input samples consecutively to the adaptation algorithm. This is also computationally lighter than using random time instances of the input because delayed values of the outputs are needed in the adaptation. They can be just buffered as consecutive samples are processed.

A problem is caused by the local nature of the delay adaptation rule. Because speech is very periodic in nature, the rule tends to get trapped in aligning those parts of the waveforms that exhibit strong correlation due to the periodicity of the signals, and not necessarily due to the delay. Figure 4. presents a study of where different initial delay values converge to, in the case of two speech signals. The initial values were delays in a grid of 50×50 samples. Those initial values converging to the correct delays (20, 20) are plotted in a darker color. We can clearly see "attractors", which are caused by the periodicity of the speech signal. An interesting thing to note is that the source signal s_1 is a male voice and s_2 female. The attractor points for d_{21} are thus more densely spaced than those for d_{12} because of the shorter pitch period of the female voice. The sampling frequency in this example was 16kHz.

If the delays converge to a wrong attractor, then the weights obviously are unable to converge to a separating solution. It is thus crucial that the initial delay values are approximately in the correct range. Here, it is possible to use knowledge of the recording situation. Even without this knowledge, cross correlations between the mixtures give good estimates for the initial values of the delays. Another possibility is to remove the disturbing time dependencies within each mixture separately (whitening), learn the delays and weights from the whitened mixtures, and then apply them to the original mixtures.

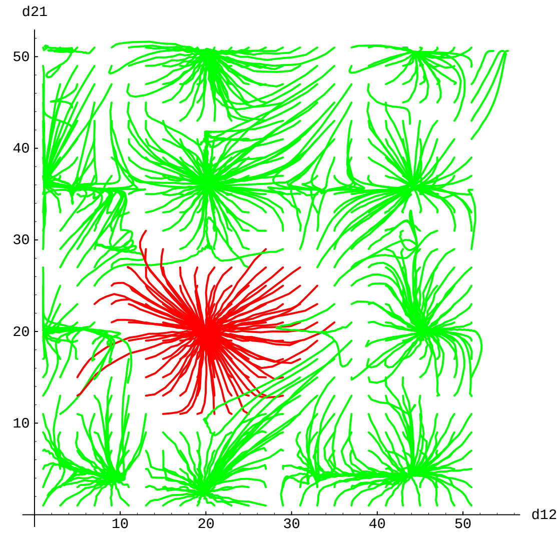

Figure 4. An illustration of the effect of the initial condition to delay convergence (see text). The correct delay values were (20,20).

7 CONCLUSION

We have presented an architecture for blind source separation of delayed sources. We demonstrated that both the weights and the delays in the network can be learned by the information maximization principle first presented by Bell and Sejnowski for the case of instantaneous mixtures [1].

While mere adaptive delays may suffice in some applications, for real audio signals they are not enough. The acoustic environment (like a room) imposes a different impulse response between each source and microphone pair. Moreover, the microphones may have different characteristics, or at least their frequency responses may differ for sources in different directions. This kind of situations can be modeled as convolved mixtures. We have also derived solutions for the blind separation of this kind of mixtures based on the same information maximization principle. This will be presented in a forthcoming paper [16].

This brings the technology one step closer to real world applications, though many items remain to be investigated, such as the effect of independent noise to the adaptation, and how to handle an unknown number of sources (especially noise sources).

REFERENCES

[1] A. Bell and T. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1004–1034, 1995.

[2] G. Burel. Blind separation of sources: A nonlinear neural algorithm. *Neural Networks*, 5(6):937–947, 1992.

[3] J.-F. Cardoso. Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem. In *Proc. ICASSP*, pages 2655–2658, Albuquerque, NM, USA, April 3-6 1990.

[4] J.-F. Cardoso, A. Belouchrani, and B. Laheld. A new composite criterion for adaptive and iterative blind source separation. In *Proc. ICASSP*, volume IV, pages 273–276, Adelaide, Australia, April 19-22 1994.

[5] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[6] P. Comon, C. Jutten, and J. Herault. Blind separation of sources, part II: Problems statement. *Signal Processing*, 24(1):11–20, 1991.

[7] N. Delfosse and P. Loubaton. Adaptive separation of independent sources: A deflation approach. In *Proc. ICASSP*, volume IV, pages 41–44, Adelaide, Australia, April 19-22 1994.

[8] A. M. Engebretson. Acoustic signal separation of statistically independent sources using multiple microphones. In *Proc. ICASSP*, volume II, pages 343–346, Minneapolis, MN, USA, April 27-30 1993.

[9] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[10] J. Karhunen, L. Wang, and R. Vigario. Nonlinear PCA type approaches for source separation and independent component analysis. In *Proc. ICNN*, Perth, Western Australia, November 27 - December 1 1995.

[11] S. Li and T. J. Sejnowski. Adaptive separation of mixed broad-band sound sources with delays by a beamforming Hérault-Jutten network. *IEEE Journal of Oceanic Engineering*, 20(1):73–79, January 1995.

[12] L. Molgedy and H. G. Schuster. Separation of independent signals using time-delayed correlations. *Physical Review Letters*, 72(23):3634–3637, June 6 1994.

[13] M. Nájar, M. A. Lagunas, and I. Bonet. Blind wideband source separation. In *Proc. ICASSP*, volume IV, pages 65–68, Adelaide, Australia, April 19-22 1994.

[14] J. C. Platt and F. Faggin. Networks for the separation of sources that are superimposed and delayed. In J. Moody, S. Hanson, and R. Lippmann, editors, *Advances in Neural Information Processing Systems 4*. Morgan-Kaufmann, 1992.

[15] K. Pope and R. Bogner. Blind separation of speech signals. In *Proceedings of the Australian International Conference on Speech Science and Technology*, Perth, Western Australia, December 6-8 1994.

[16] K. Torkkola. Blind separation of convolved sources based on information maximization. To be published.

[17] E. Weinstein, M. Feder, and A. Oppenheim. Multi-channel signal separation by decorrelation. *IEEE transactions on Speech and Audio Processing*, 1(4):405–413, 1993.

[18] B. Widrow and S. Stearns. *Adaptive Signal Processing*. Prentice-Hall, 1985.

[19] D. Yellin and E. Weinstein. Criteria for multichannel signal separation. *IEEE Transactions on Signal Processing*, 42(8):2158–2168, August 1994.