A CONTEXTUAL BLIND SEPARATION OF DELAYED AND CONVOLVED SOURCES

Te-Won Lee

Max-Planck-Society, Fault-Tolerant Computing Group
in Potsdam, Germany, AND
the Interactive Systems Group at Carnegie Mellon
University, Pittsburgh, PA 15213, USA
tewon@cs.cmu.edu

Reinhold Orglmeister

Electronics Institute,
Department of Electrical Engineering,
Berlin University of Technology, Germany,
orglm@tubife1.ee.tu-berlin.de

ABSTRACT

We present a new method to tackle the problem of separating mixtures of real sources which have been convolved and time-delayed under real world conditions. To this end, we learn two sets of parameters to unmix the mixtures and to estimate the true density function. The solutions are discussed for feedback and feedforward architectures. Since the quality of separation depends on the modeling of the underlying density we propose different methods to closer approximate the density function using some *context*. The proposed density estimation achieves separation of a wider class of sources. Furthermore, we employ the FIR polynomial matrix techniques in the frequency domain to invert a true-phase mixing system. The significance of the new method is demonstrated with the successful separation of two speakers and separation of music and speech recorded with two microphones in a reverberating room.

1. INTRODUCTION

In blind source separation the problem is how to recover independent sources given the sensor outputs in which the sources have been mixed in 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 and statistical signal processing. Comon [6] has defined 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 the 4th order and requires intensive computation. Researchers in neural computation have developed adaptive learning algorithms which are simpler and biologically more plausible [1, 2, 7].

Recently, Bell and Sejnowski [2] have proposed an information theoretic approach to the blind source separation problem. Torkkola has applied this approach to convolution and time-delays [12]. In [3], we have extended Torkkola's architecture to a full recurrent filter system that deals with convolved and time-delayed sources. Pearlmutter and Parra have reformulated the ICA in a Maximum Likelihood framework [10] 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 real reverberating environments [14, 12, 9].

In this paper, we present a new method that combines the algorithms for convolved and delayed sources in [3] with the context sensitive generalization of ICA [10]. Two unmixing architectures, feedback and feedforward, are presented. Although the feedback solution is elegant its structure is restricted to non minimum phase inverse systems. Therefore, we make use of FIR polynomial matrix techniques in the frequency domain [8] to approximate a true phase inverse solution with non causal extensions. Furthermore, we propose alternative ways to model the underlying density: The logistic function can be extended by some flexible sigmoids and the generalized Gaussian provides good approximation with less parameters. In case of a contextual modeling the generalized Gaussian can be extended to a sum of Gaussian densities using the Parzen window density estimation [13]. Although density estimation is computationally burdensome it has the advantage of separating a wider class of sources (including super-Gaussian and sub-Gaussian). The algorithm has been applied to the difficult problem of separating two speakers and one speaker with music in the background recorded in a conference room and a normal office environment.

2. ARCHITECTURE

Figure 6 shows the mixing $A(z)$ and unmixing system $W(z)$. In real world situation the mixing of sources $s(t)$ involves convolution and time-delays as follows: $x_i(t) = \sum_{j=1}^{N} \sum_{k=0}^{M-1} a_{ijk} s_j(t - D_{ij} - k)$ where a_{ijk} are the filter coefficients, D_{ij} denotes time delays and $x_i(t)$ are the observations. The obvious solution to invert the mixing system $A(z)$ is a full feedback matrix of M-taps IIR filters with appropriate time delay compensations before the cross filters. An inverting system that separates the mixtures without deconvolving the sources has been presented by Torkkola in [12]. This architecture has been extended to deal with deconvolved and time delayed sources in [3]. Such a system can be written as follows: $\mathbf{u}(t) = \mathbf{x}(t) - \mathbf{W}_0\mathbf{u}(t) - \sum_{k=1}^{M-1} \mathbf{W}_k\mathbf{u}(t-k)$ $\mathbf{W}_0$ denotes the leading weights and w_{ijk} denotes the unmixing and deconvolving filters. The use of IIR filters is restricted to ARMA (autoregressive moving average) systems with minimum

0-8186-7919-0/97 $10.00 © 1997 IEEE

phase. A non-minimum phase system leads to an instable inverse system with poles outside the unit circle. However, we can approximate an inverse system using a FIR representation that is capable of a non-causal filter expansion. The feedforward architecture can be written as follows: $u_i(t) = \sum_{j=1}^{N} \sum_{k=-M/2}^{M/2} w_{ijk} x_j(t-k)$ where the leading weights are placed at half of the filter size $M/2$ to allow for non-causal expansions.

3. ALGORITHMS

3.1. Source Separation Based on Information Maximization

The separation of independent components from a linear mixture can be described by a general measure of independence between the pdf of a random variable $p_{\mathbf{u}}(\mathbf{u})$ and the pdf of its components $\prod_{i=1}^{n} p_{u_i}(u_i)$. The Kullback-Leibler divergence measures the degree of distance defined by:

$$\delta(p_{\mathbf{u}}(\mathbf{u}), \prod_{i=1}^{n} p_{u_i}(u_i)) = \int p_{\mathbf{u}}(\mathbf{u}) \log \frac{p_{\mathbf{u}}(\mathbf{u})}{\prod_{i=1}^{n} p_{u_i}(u_i)} d\mathbf{u} \quad (1)$$

and vanishes if and only if $p_{\mathbf{u}}(\mathbf{u})$ factories which leads to $\delta(p_{\mathbf{u}}(\mathbf{u}), \prod_{i=1}^{n} p_{u_i}(u_i)) = 0$. The observation $\hat{p}_{\mathbf{u}}(\mathbf{u};\mathbf{C}) = \prod_{i=1}^{n} p_{u_i}(u_i)$ can be obtained using a density estimator with parameters $\mathbf{c}$. The Kullback-Leibler divergence has also the form of the mutual information of $\mathbf{u}$ and this can be rewritten in terms of entropies as follows: $\delta(p_{\mathbf{u}}(\mathbf{u}), \hat{p}_{\mathbf{u}}(\mathbf{u};\mathbf{C})) = H(p_{\mathbf{u}}(\mathbf{u})) - H(p_{\mathbf{u}}(\mathbf{u}) \mid \hat{p}_{\mathbf{u}}(\mathbf{u};\mathbf{C}))$ Bell and Sejnowski [2] have proposed an information-theoretic approach where they maximize the mutual information that an output $y = g(u)$ of a neural processor contains about its input u. They have shown that for invertible and continuous deterministic mappings $g(u)$, the mutual information between inputs and outputs can be maximized by maximizing the entropy of the outputs alone where the output pdf satisfies: $p_{\mathbf{y}}(\mathbf{y}) = \frac{p_{\mathbf{u}}(\mathbf{u})}{|\det J(\mathbf{u})|}$ with $\det J(\mathbf{u})$ being the determinant of the Jacobian $J_{i,j} = \frac{\partial g_i}{\partial u_j}$. Maximizing the output entropy $H(\mathbf{y})$ then implies approximating the output density in the sense of minimum KL-divergence, by a uniform density. This corresponds to producing white noise at the output of the neural processor and at the same time making the input signals prior the transfer function $g(u)$ independent while shaping them according to the derivative $\partial g(u)/\partial u$ that has higher kurtosis than the pdf of the sources. This may be viewed as maximum entropy estimation of the input densities under the parameterization of $\hat{p}_{\mathbf{u}}(\mathbf{u};\mathbf{c})$. We can relate $p_{\mathbf{y}}(\mathbf{y})$ to the nonlinear transfer function that gives us the pdf estimate $\hat{p}_{\mathbf{u}}(\mathbf{u};\mathbf{C})$: $|\det J(\mathbf{u})| = |\det W| \prod_{i=1}^{n} \frac{\partial g(u_i)}{\partial u_i} = \hat{p}_{\mathbf{u}}(\mathbf{u};\mathbf{C})$ The logarithmic representation is:

$$\log(\hat{p}_{\mathbf{u}}(\mathbf{u};\mathbf{c})) = log|\det W| + \sum_{i=1}^{n} \log(\frac{\partial g(u_i)}{\partial u_i}) \quad (2)$$

Evaluating the expected value for eq.2 gives the output entropy. We can now maximize the output entropy to derive two different set of parameters namely $\mathbf{W}$ in charge of unmixing the signals $\mathbf{x}$ and $\mathbf{C}$ parameterizing the shape of the pdf estimation $p_i(u_i;\mathbf{c}_i)$:

$$\frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} = \mathbf{W}^{-T} + \left(\frac{\frac{\partial p_i(u_i;\mathbf{c}_i)}{\partial \mathbf{c}_i}}{p_i(u_i;\mathbf{c}_i)} \right)_i \mathbf{x}^T \quad (3)$$

$$\frac{\partial H(\mathbf{y})}{\partial \mathbf{c}_i} = \frac{\frac{\partial p_i(u_i;\mathbf{c}_i)}{\partial \mathbf{c}_i}}{p_i(u_i;\mathbf{c}_i)} \quad (4)$$

Considering the set of parameters $\mathbf{W}$, 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]:

$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} \quad (5)$$

This is an optimal rescaling of the entropy gradient. It simplifies the learning rule and speeds convergence considerably.

3.2. Learning Rules for Feedback and Feedforward Systems

The learning rules for the feedback system has been derived in [3] and are as follows: $\Delta \mathbf{W}_0 \propto -(\mathbf{I} + \mathbf{W}_0)(\mathbf{I} + \hat{\mathbf{y}}\mathbf{u}^T)$, $\Delta \mathbf{W}_k \propto -(\mathbf{I} + \mathbf{W}_k)\hat{\mathbf{y}}\mathbf{u}_{t-k}^T$, $\Delta d_{ij} \propto -\hat{y}_i \sum_{k=1}^{M-1} \frac{\partial}{\partial t} w_{ijk} u(t - d_{ij} - k)$. Independently, Cichocki et al. [5] have presented similar results with a full feedback system.

The learning rules for a feedforward inverse system can be formulated using the FIR polynomial matrix algebra as described by Lambert [8]. 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. Therefore, by using FIR polynomial matrices we reduce the convolution and deconvolution problem to element wise multiplication and element wise division of polynomials. Although the application of the Fourier transform translates the entire FIR filter matrix into the FIR polynomial matrix and vice versa we need to satisfy a well approximation of the double-sided Laurent series expansion which can be achieved by pre pending and post pending zeros in the time domain to allow for non-causal expansions of non-minimum phase roots. The rules in eq.3 and eq.5 can be reformulated in the frequency domain:

$$\underline{\mathbf{W}} = \underline{\mathbf{W}} + (\underline{\mathbf{W}}^{-1} + FFT(\hat{\mathbf{y}}(\mathbf{x}))\underline{\mathbf{X}}^{\mathrm{H}}) \quad (6)$$

$$\underline{\mathbf{W}} = \underline{\mathbf{W}} + (\underline{\mathbf{W}} + FFT(\hat{\mathbf{y}}(\mathbf{x}))\underline{\mathbf{U}}^{\mathrm{H}}\underline{\mathbf{W}}) \quad (7)$$

where the H superscript denotes the complex transpose. 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. The implementation of the FIR polynomials in the frequency domain can be done using block LMS techniques.

4. CONTEXTUAL DENSITY ESTIMATION

Eq.2 and eq.4 leave us some *degree of freedom* as long as they are differentiable in choosing a nonlinear function that fits the unknown true distribution of u_i. Although good results have been presented in [2, 3, 12] with a single neuron the question is how well a single sigmoid can approximate the pdf of the multidimensional vector. The idea of density shape matching with sigmoids has been proposed by Roth and Baram [11]. Pearlmutter and Parra [10] have extended the *infomax* ICA algorithm to a context-sensitive generalization of ICA where they 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 6(b).

$$p_i(u_i(t)|u(t-1),...;\mathbf{c}_i) = \sum_{k=1}^{K} \frac{m_{ik}}{\sigma_{ik}} \frac{\partial g(u_i)}{\partial u_i} \left(\frac{u_i(t) - \bar{u}_{ik}}{\sigma_{ik}} \right) \tag{8}$$

where m_{ik} are the mixing parameters and σ_{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: $\bar{u}_{ik} = \sum_{\tau=1} a_{ik}(\tau)u_i(t-\tau) + b_{ik}$ where the linear filter prediction coefficients a_{ik} and the bias b_{ik} are also elements of the weight set $\mathbf{c}_i$. Using the density distribution p_i and its derivative $\partial p_i(u_i;\mathbf{c}_i)/\partial \mathbf{c}_i$ we can apply a stochastic gradient descent rule to learn the prediction coefficients $a_{ik}(\tau)$, the bias b_{ik} and the logistic distribution parameters: means m_{ik} and scales σ_{ik}. The exact learning rules are derived in [10]. The contextual ICA requires a large number of data points to provide a fairly good pdf estimate. A more suitable but computationly burdensome way of modeling the underlying true density is using a weighted sum of *flexible sigmoid* functions instead of the pure logistic function. The *flexible sigmoid* is given by the differential equation: $\frac{\partial g(u)}{\partial u} = g(u)^p(1-g(u))^r$ For example, if $p, r > 1$ the shape of the pdf turns spikier with longer tails and gives a better approximate of speech signals. The disadvantage is that we need to build a look-up table for each of the given parameters p, r. However, for many examples a single flexible sigmoid function may be sufficient to estimate the density function. Another class of nonlinearity that can be used as a density shaper is the generalized Gaussian nonlinearity. We have experienced that the initial state of the pdf can be modeled better by a generalized Gaussian function.

$$p_i(u_i;\mathbf{c}_i) \propto \exp(-\frac{|u|^s}{sE|u|^s}) \tag{9}$$

Eq.9 describes pdf's ranging from impulsive $s < 1$, Gaussian $s = 2$ to more bounded pdf with $s >> 1$. Our observation showed that for real recordings the output entropy increased with a decrease in gradient norm when using a non-Gaussian nonlinearity beyond initialization. A plausible reason for a generalized Gaussian nonlinearity for the initial state is that the linear mixing of independent sources *gaussianize* the sources due to the central limit theorem. In the optimization process we use a flexible logistic density function with variable means and scales. Since the parameters s, r, p cannot be described analytically we learn these parameters by gradient ascent of the entropy surface. We have also performed experiments with a non-parametric density estimation method such as the Parzen window density estimation [13]. The basic form of the density is as follows:

$$p_i(u_i;\mathbf{c}_i) = \frac{1}{N_k} \sum_{k=1}^{K} R(u - u_k;\mathbf{c}_i) \tag{10}$$

where N_k is the width of the window and $R(.)$ is the smoothing function. If $R(.)$ is chosen to be a generalized Gaussian density function the pdf can be modeled as a mixture of generalized Gaussian. Here, the pdf can be as well conditioned on the recent history as in the cICA case. The nonlinearity is given by the sum of generalized Gaussian nonlinearities.

5. EXPERIMENTAL RESULTS

5.1. Simulation Results with the Feedback Architecture

We performed several simulations with the feedback architecture where we chose the mixing system to be min-phase. The density shaping method is applied to separate mixtures of super-Gaussian and sub-Gaussian sources. In case of two sources one flexible nonlinearity such as the generalized Gaussian or the flexible sigmoid is sufficient whereas for a larger number of source the contextual modeling with a sum of sigmoids or sum of generalized Gaussian nonlinearity is necessary. In general, the feedback structure is not able to invert real room recordings and limited success is observed for recordings where the sources are placed close to the microphone. Due to limited space we show figures only for the feedforward system.

5.2. Results on Real Recordings

In these experiments, we tackle the difficult problem of separating two speakers recorded with two microphones in a real room. To this end, we use the feedforward architecture and employ the algorithms in eq.6 and eq.4. The density shape estimator eq.8 is first approximated by a generalized Gaussian nonlinearity in which the s-parameters are learned according to eq.4. In the process of optimization we switch to the nonlinearity approximated by a flexible sigmoids for each u_i. Figure 6 (a) and (b) show the recordings of one person saying the digits one to ten while loud music plays in the background. In this experimental setup the sources and sensors are placed in a square order with 60cm distance between the sources and the microphones. The algorithm converges after 30 epochs through a 10sec recording with 16kHz. The unmixed signals is obtained using 256 taps FIR filters which cover a delay of 16ms. The separated signals are shown in figure 6 (c) (d) and a listening test shows an almost clean separation. In another example, we could recover two speakers recorded in a normal room. 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 perform experiments with 10 sentences recorded with loud music in the background and 10 sentences recorded with a competing speaker. After separation, the recognition rate increases considerably for both cases. Speech recognition results are listed in detail in [9].

6. CONCLUSIONS

We have presented a new method that combines the learning rules to blindly recover convolved and delayed sources from their mixtures with a contextual density shaping algorithm which allows a more realistic modeling of the underlying density. To this end, we have proposed different methods to estimate the density function. Since real recordings require a true phase system inverse we employ the FIR polynomial matrix technique in the frequency domain with extension to non causal filter solutions. The new method has been successfully applied to the separation problem of two speakers and speaker separation with music recorded in a real reverberating room. These techniques bring us several steps closer to success on real-world data.

ACKNOWLEDGMENTS

T.W.L. is supported by the Daimler-Benz-Fellowship, We are grateful to Tony Bell, Russ Lambert and Alex Waibel 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] A. Bell, T. Lee and R. Lambert Blind separation of convolved and delayed sources. In *Advances in Neural Information Processing Systems 9.* MIT Press, 1997.

[4] J-F. Cardoso and B. Laheld. Equivariant adaptive source separation. IEEE Trans. on S.P., Dec. 1996.

[5] J. Cao A. Cichocki, S. Amari. Blind separation of delayed and convolved signals with self-adaptive learning rate. In *Proc. NOLTA'96*, 1996.

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

[7] 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.

[8] Russ Lambert. Multichannel blind deconvolution: Fir matrix algebra and separation of multipath mixtures. Thesis, University of Southern California, Department of Electrical Engineering, May 1996.

[9] T. Lee and R. Orglmeister. Blind source separation of real-world signals. submitted to *Proc. ICNN*, Houston, USA, 1997.

[10] B. Pearlmutter and L. Parra. A context-sensitive generalization of ICA. In ICONIP'96 . In press.

[11] Z. Roth and Y. Baram. Multidimensional density shaping by sigmoids. *IEEE Trans. on Neural Networks*, 7(5):1291–1298, 1996.

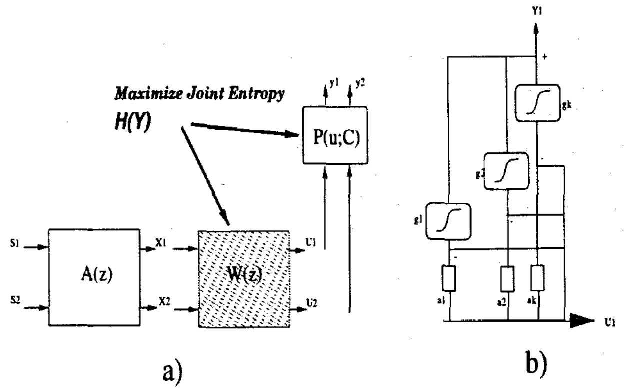

Figure 1. a) A(z) denotes the mixing system and W(z) denotes the unmixing system. W(z) can be a feedback or a feedforward architecture. b) Contexual density estimation as a sum of nonlinearities

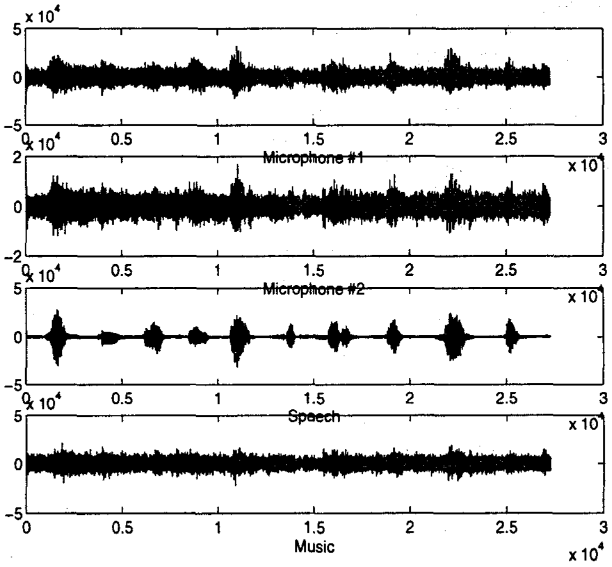

Figure 2. a) Mic. #1 b) Mic. #2 c) separated speech signal d) separated music signal

[12] 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).

[13] Paul Viola. Alignment by maximization of mutual information. Thesis, MIT, AI Lab, June 1995.

[14] D. Yellin and E. Weinstein. Multichannel signal separation: Methods and analysis. *IEEE Transactions on Signal Processing*, 44(1):106–118, January 1996.