NOISE REDUCTION AND SPEECH ENHANCEMENT VIA TEMPORAL ANTI-HEBBIAN LEARNING

M. Girolami

Department of Computing and Information Systems[1]
The University of Paisley
Paisley, PA1 2BE, Scotland

ABSTRACT

Temporal extensions of both linear and nonlinear anti-Hebbian learning have been shown to be suited to the problem of blind separation of sources from their convolved mixtures. This paper presents a generalized form of anti-Hebbian learning for a partially connected recurrent network based on the maximum likelihood estimation principle. Inspired by features of the binaural unmasking effect the network and associated online adaptation are applied to the enhancement of speech, which is corrupted by interfering noise, competing speech and reverberation. Graded simulations based on speech corrupted with increasingly complex levels of reverberation are reported. It is shown that for high levels of reverberation the proposed method compares favorably with classical adaptive filter approaches to speech enhancement in real acoustic environments.

1. INTRODUCTION

The signal-processing problem of noise reduction and speech enhancement has received considerable attention within the adaptive filter community [1]. One of the most challenging areas of this research is the development of adaptive algorithms for hearing aids [2, 3]. A closely linked problem, which has been the focus of research in recent years from the artificial neural network community (ANN), is that of blind separation of sources and in particular the convolutive blind source separation problem [4; 9]. These approaches have been based primarily on the recurrent network originally proposed in [7] and employed as a model of anti-Hebbian learning in [12].

The general model of cross-channel interference for a recorded signal is defined as

$$x_i(t) = \sum_j a_{ij} * s_j(t) \qquad (1)$$

Where $x_i(t)$ denotes the finite sample recorded at the i^{th} sensor at time *t*. This recording will be a summation of all the sources (noise and desired signals) convolved by the channel transfer function a_{ij}. This can be written in vector notation for the case where there are *N* sources and *N* recordings $\mathbf{x}(t) = \mathbf{A}(z)\mathbf{s}(z)$. To recover the corrupted sources it is then necessary to at best invert the channel transfer functions. An inverting transformation **W**(z) (in *z* notation) is sought such that $\mathbf{y}(t) = \mathbf{W}(z)\mathbf{x}(z)$. In the case of acoustic transmission media the channel transfer functions can be extremely complex and time varying. Many simplifying assumptions are made on the transfer functions to constrain this highly complex situation to one which is more tractable. Such assumptions are detailed in [10] where the transfer functions can be modeled as either stationary finite or infinite causal and minimum-phase impulse response filters. In some situations these assumptions may be valid, i.e. anechoic or very light reverberation levels which are difficult to enumerate [10]. The classical adaptive noise cancellation (ANC) algorithm utilizing the least mean square (LMS) algorithm has been extensively employed in noise reduction for speech enhancement [1,2,3].

Binaural hearing has been shown superior [11] to monaural at maintaining the intelligibility of speech in the presence of reverberation, continuous speech shaped noise, or competing connected speech. That is, binaural processing does not simply perform coherent addition of the signals at the two ears. The "binaural unmasking" effect [11] lowers the hearing threshold, may operate in frequency sub-bands, and appears to utilize binaural correlation properties to de-emphasize an undesired signal. Now linear anti-Hebbian learning is driven solely by signal correlation in providing decorrelated network output [12]. It is then feasible to consider anti-Hebbian learning as a potential model for performing a form of binaural masking. A brief review of the adaptive methods, which have been applied to source separation of convolved mixtures, is now given.

In [7] the original network structure proposed by Jutten and Herault is extended to one possessing Finite Impulse Response (FIR) filter weights (Figure 1). Algorithms are developed for weight update in attempting to cancel out fourth order cross cumulants at all finite time delays within the filter structure. Decorrelation-based algorithms for multi-channel signal separation are considered in [4]. The efficacy of these algorithms is demonstrated on recordings of mixtures of up to four speakers and excerpts of music. Recordings for the reported simulations are made in an anechoic chamber which exhibits little acoustic dispersion, and so the inverting filters are only required to identify the cross microphone delays. In [13] extensions of Bussgang type algorithms for source separation are proposed. Simulations reported are based on separation of speakers in real room environments using the Equivariant [13] Bussgang type algorithms developed, however no quantitative results are made available. Feed-forward and recurrent network algorithms are extended in [5] to take into account delays and convolutions. A novel learning rate adaptation algorithm is proposed for signals, which may be mixed in time varying environments. The standard

[1] Currently on secondment to the ABS Laboratory, Frontier Research Program, RIKEN, 2-1 Hirosawa, Wako-shi, Saitama, 351-01, Japan

cubic nonlinearity is used within the reported simulations, which are run on artificial data.

The Infomax algorithm [9] is extended to consider convolved mixtures of sources. A feedback network similar to that proposed in [7] is considered. Results reported include separation of a mixture of two speakers recorded in a small conference room with cross channel interference reduction of 12dB reported. Source separation of speakers and music recorded in real acoustic environments are presented in [8]. Comparisons of separation performance are made based on recognition rates of Automatic Speech Recognition Systems (ASR). For speech recorded with music, increases of 50% in recognition rates are reported after separation processing. Speech corrupted with competing speech yielded a 19% increase in recognition rate after processing. The use of filters with acausal extensions is employed which ensures that the separating filters will be stable even if the inverse mixing filters are non-minimum phase [1]. Very recently time domain convolutive source separation algorithms have been considered [6]. By explicitly considering the minimization of the asymptotic error variance separating functions are developed which are related to the probability density functions of the underlying sources. In the following section we will observe that the maximum likelihood framework for parameter estimation will naturally asymptotically minimize the parameter estimation error variance.

Researchers have found for convolutive mixtures of signals (speech is the most often considered source signal) second order statistics appear to suffice [10]. No benefit had been found by explicit use of either higher order statistics or sigmoidal nonlinearities. In the case of the Infomax temporal extension [9] maximizing the information through a hyperbolic tangent is shown analytically to yield independent components, second order temporal anti-Hebbian learning yields substantially similar separation performance [14].

Based on the equivalence of mutual information minimization and maximum likelihood estimation (MLE) for instantaneous BSS [16] we can consider extending the MLE to include temporal context. In the following section the maximum likelihood framework for developing anti-Hebbian rules which can then be applied to source separation is presented.

2. TEMPORAL ANTI-HEBB LEARNING

Let us consider N independent observations of a random variable vector $\{\mathbf{x}_i\}: i=1\ldots\mathrm{N}$, which will take values in $\Re^{L\times N}$ where L is the vector dimension. The variable is distributed according to the probability density function (pdf) $p(\mathbf{x};\theta)$ where $\theta\in\Theta$ is the set of parameters describing the form of the density function. Let $p(\mathbf{x};\hat{\theta}).\hat{\theta}\in\Theta$ be a parametric estimation of the pdf $p(\mathbf{x};\theta)$ after observing $\{\mathbf{x}_i\}: i=1\ldots\mathrm{N}$. The log-likelihood that the observations are drawn from a parametric pdf $p(\mathbf{x};\hat{\theta})$ is given as $N^{-1}\sum_{i=1}^{N}\log p(\mathbf{x}_i;\hat{\theta})$. We wish to parameterize the density estimate such that the likelihood of the observed data being drawn from the parametric density is maximized. The MLE is defined as

$$\hat{\theta}_{ML}=\arg\max\left(N^{-1}\sum_{i=1}^{N}\log p(\mathbf{x}_i;\hat{\theta})\right) \quad (1)$$

If the observations of $\{\mathbf{x}_i\}: i=1\ldots\mathrm{N}$ are independent and identically distributed as $p(\mathbf{x};\theta)$ then $\hat{\theta}_{ML}\xrightarrow{P=1}\theta$; $\mathrm{N}\to\infty$. So the ML estimate of the pdf parameters tends to the true parameter values with probability one as the number of observations tends to infinity. We must also note that the MLE in the limit will reach the Cramer-Rao bound, which indicates the parameter estimates will be unbiased and the asymptotic error variance will be minimized [1]. So we have that the parameter estimation error tends in distribution, to the zero mean normal distribution with covariance equal to the inverse of the Fisher information matrix $\mathbf{M}$, $\sqrt{\mathrm{N}}(\hat{\theta}-\theta)\to N(0,\mathbf{M}^{-1})$. Now as we have the following: -

$$\begin{aligned}\mathrm{N}\to\infty\Rightarrow\hat{\theta}_{ML}&=\arg\max_{\hat{\theta}}\left(\int p(\mathbf{x};\theta)\log p(\mathbf{x};\hat{\theta})d\mathbf{x}\right)\\&=\arg\max_{\hat{\theta}}\left(\int p(\mathbf{x};\theta)\log\left(\frac{p(\mathbf{x};\hat{\theta})}{p(\mathbf{x};\theta)}p(\mathbf{x};\theta)\right)d\mathbf{x}\right)\end{aligned}$$

$$\Rightarrow\arg\max_{\hat{\theta}}\left(\int p(\mathbf{x};\theta)\log\left(\frac{p(\mathbf{x};\hat{\theta})}{p(\mathbf{x};\theta)}\right)d\mathbf{x}+\int p(\mathbf{x};\theta)\log(p(\mathbf{x};\theta))d\mathbf{x}\right)$$

$$\Rightarrow\arg\max_{\hat{\theta}}\left(-K_L\left(p(\mathbf{x};\theta)\middle\|p(\mathbf{x};\hat{\theta})\right)-H(p(\mathbf{x};\theta))\right) \quad (2)$$

Where $K_L(a\|b)$ is the Kullback-Leibler divergence between the estimated and the true pdf, [16]. H(a) is the standard Shannon differential entropy of the given pdf. In [16] Cardoso links (2) to source separation for instantaneous mixing matrices; this is then applicable to mixing filters. As the parametric model is given by the unmixing / deconvolving filters and noting that the rightmost term in (2) can be neglected as it is not a function of the estimates of the parametric variables, the generic MLE is then $\arg\max_{\hat{\theta}}\left(-K_L\left(p(\mathbf{x};\theta)\middle\|p(\mathbf{x};\hat{\theta})\right)\right)$. This is equivalent to

$$\arg\max_{\mathbf{W}(z)}\left(-K_L\left(p(\mathbf{A}(z)\mathbf{s};\mathbf{A}(z))\middle\|p\left(\mathbf{W}^{-1}(z)\hat{\mathbf{s}};\mathbf{W}(z)\right)\right)\right) \quad (3)$$

Where $p(\mathbf{A}(z)\mathbf{s};\mathbf{A}(z))$ is the true distribution of the observed data $\mathbf{x}$ and $p\left(\mathbf{W}^{-1}(z)\hat{\mathbf{s}};\mathbf{W}(z)\right)$ is the parametric estimate where the network filter weights are drawn from the parametric set $\hat{\theta}=\mathbf{W}(z)\in\Theta$. As the unobserved source data will have independent components, the form of the parametric model to be used $p\left(\mathbf{W}^{-1}(z)\hat{\mathbf{s}};\mathbf{W}(z)\right)$ must then be factorable. The maximization of the MLE (3) using the parametric models discussed will then minimize the Kullback mismatch between the parametric pdf of the estimated latent generators $\hat{\mathbf{s}}$ and the actual factorable pdf of the independent source variables. We have an asymptotically efficient estimator for the inverting filters in the MLE. In this section we develop a network model and learning algorithm to separate, in a blind manner, signals which have been convolved together.

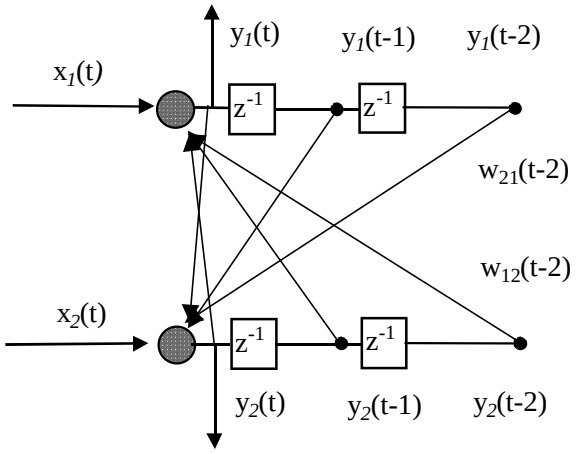

Figure 1. Diagram of a two input two output realization of the temporal anti-Hebbian network.

From Figure 1, the output of each node at time t for an $N \times N$ network with memory based synaptic weights of length L is given as

$$y_i(t) = x_i(t) + \sum_{j \neq i} \sum_{k=0}^{L} w_{ij}(t-k) y_j(t-k) \tag{4}$$

where the subscripts denote spatial relations between the nodes within the network and the (t-k) terms denote delays of k samples from time t. The value of the MLE is given by the solution of $\frac{\partial}{\partial \hat{\theta}}\left(N^{-1} \sum_{i=1}^{N} \log p\left(\mathbf{x}_i; \hat{\theta}\right)\right) = 0$. We note that in the limit where the number of sample observations tends to infinity, we have an expectation. This can now be solved iteratively using the Robbins-Monro algorithm such that parameter updates will be

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \mu_n \frac{\partial}{\partial \hat{\theta}} \log\left(p\left(\mathbf{x}_{n+1}; \hat{\theta}_n\right)\right) \tag{5}$$

The parametric form of the transformed observation (4) has only one set of parameters in this case, that is $w_{ij}(t-k) : \forall \left(i \neq j \in 1 \ldots N\right) \wedge \forall k \in 1 \ldots L$. Note that the parametric model will require to be factorable, with this constraint we can then write

$$w_{ij}^{n+1}(t-k) = w_{ij}^{n}(t-k) + \mu_n \frac{\partial}{\partial w_{ij}^{n}(t-k)} \log\left(\prod_{l=1}^{N} p_l\left(y_l^{n+1}; w_{ij}^{n}(t-k)\right)\right)$$

$$w_{ij}^{n+1}(t-k) = w_{ij}^{n}(t-k) + \mu_n \frac{p_i'\left(y_i^{n+1}; w_{ij}^{n}(t-k)\right)}{p_i\left(y_i^{n+1}; w_{ij}^{n}(t-k)\right)} y_j(t-k) \tag{6}$$

The generic sequential parameter update algorithm of (6) has a number of important points. Firstly as the sample size tends to infinity then (6) will converge with probability one to an unbiased estimate of the true parameters of the underlying parametric model. This pre-supposes that the parametric model chosen is correct for the underlying latent data. MacDonald [17] has shown that the gross statistics of natural speech can be approximated by a gamma distribution variant or the Laplacian density. The Laplacian density is a special form of the Generalised Gaussian and in this case it is a simple matter to see that the weight update equation will be given by

$$w_{ij}^{n+1}(t-k) = w_{ij}^{n}(t-k)_i - \mu_n \frac{sign\left(y_i(t)\right) y_j(t-k)}{E\left\{\left|y_i(t)\right|\right\}} \tag{7}$$

This update has been proposed in [15] and compared with standard linear decorrelation adaptation and temporal infomax. It has been found to give significant SNR improvement for blind source separation of natural speech.

3. EXPERIMENTS

The focus of this paper is primarily on the application of the neural adaptation to realistic situations which are characteristic of a real acoustic situation. To compare the performance of this algorithm with what is effectively a toy situation anechoic conditions are used as the baseline performance measure against which all other performance is measured.

The anechoic condition allows a test of the ability of the network to reduce the interfering noise level without the complication of having to consider reverberation. For the anechoic condition transfer functions will be simply the delay due to the distance between the speaker and the microphone, whereas for the real-room data the transfer functions will be much more complex due to the multi-path acoustic reflections. The recording situation was modeled as a speech source at a distance of 0.5m directly in front (0 degrees azimuth) of the input microphones (omnidirectional and placed at opposite points of a spherical simulated head of diameter 18cm), and a masking source of speech shaped noise at 135 degrees azimuth, and a distance of 4m [3, 18].

Two sets of real room recordings were made. The first set of recordings were made in a living room of dimensions $8.5 \times 6.0 \times 2.5$m. The measure of reverberation in an acoustic environment is the standard T_{60} measure. The T_{60} time for the room was measured to be approximately 0.34 seconds. Omni-directional microphones were placed 0.4m apart and at a distance of 0.5m from the loudspeakers, in a square format. This simple symmetric structure of stationary point sources and microphones dramatically reduces the complicating effects of reverberation and unequal lengths of cross coupling transfer functions. Nevertheless this is a challenging simulation and provides an indication to the more realistic problem of unequal transfer paths as explored in the final and most realistic acoustic situation.

The second set of real room recordings were made in a room of dimensions 6.5m x 4.5m x 2.5m. The room was carpeted and included soft furnishings. The T_{60} time in this case was measured as 0.30s. The geometry of the microphones and signal sources within the real room were as described for the anechoic case except that a Kemar manikin replaced the sphere used to create the head shadow and diffraction effects. The microphones were fixed within the ear canal of the manikin using standard acoustic couplers [3, 18]. The same phrases (of duration 2.25 second and sampled at 20kHz) were uttered in each simulation and the microphone output was presented only once to the network with

adaptation taking place at each sample presentation of the microphone output.

The results, in terms of SNR are detailed in Figure 2. For the anechoic case the results are most impressive with an SNR of 34dB being reached after only approximately 2000 samples. However, once moving from this situation to a real room where the effect of reverberation is minimized by constraining the recording setup to a symmetric source and receiver geometry, a significant drop of 14dB in noise reduction performance is noticed. This is of particular importance as it indicates that a finite impulse response filter perhaps may not faithfully model the multipath effects induced by significant levels of reverberation. This effect is most marked in the second real room recordings where the symmetric geometric constraints are removed and the levels of SNR improvement are of the order of 5dB.

It should be noted that the LMS algorithm applied in an adaptive noise cancellation (ANC) configuration achieved a noise reduction of 3dB for the final experimental recordings [18]. So it is clear that the temporal anti-Hebbian learning algorithm (7) provides superior performance to wideband LMS ANC.

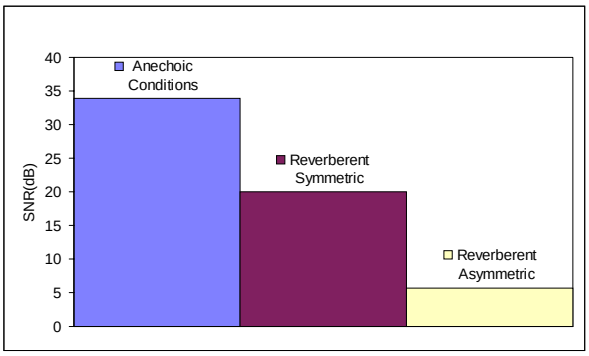

Figure 2. Chart of the average signal to noise ratio achieved for each experimental situation.

4. SUMMARY

The toy problem of anechoic conditions has been considered by a number of authors however, the more realistic situation of complicating reverberant acoustic environments has received little attention from the ANN community. The problem complexity increases substantially in the real world situation, however the technique proposed here has been found to compare more favorably than standard wideband LMS ANC filtering techniques. Further work requires to identify more realistic models for reverberation.

5. REFERENCES

[1] Haykin, S. Adaptive filter theory, 2nd ed., Prentice Hall, Englewood Cliffs, NJ, 1991.

[2] Toner,E., Campbell,D.R.**,** 'Speech Enhancement using Sub-Band intermittent adaptation' , *Speech Communication*, **12**, 253-259, 1993.

[3]Shields, P.W., Campbell D.R.,', Multi-Microphone Sub-Band Adaptive Signal Processing For Improvement of Hearing Aid Performance: Preliminary Results Using Normal Hearing Volunteers', Proc. ICASSP-97, *I.E.E.E Conference on Acoustics, Speech and Signal Processing*, **1**, 415-418, 1997.

[4] Chan, D, C, B., Godshill, S, J. and Rayner, P, J, W. , Multi-channel Multi-tap Signal Separation By Output Decorrelation, Cambridge University, CUED/F-INFENG/TR 250, ISSN 0951-9211, 1996.

[5] Cichocki, A., Amari, S, I., and Cao, J. Blind Separation of Delayed and Convolved Signals with Self-Adaptive Learning Rate, *International Symposium on Nonlinear Theory and Applications*, pp. 229-232, 1996.

[6] Charkani, N., and Deville, Y., Optimisation of the Asymptotic Performance of Time-Domain Convolutive Source Separation Algorithms., *Proc European Symposium on Artificial Neural Networks*, pp 273 – 278, ISBN 2-9600049-7-3, 1997.

[7] Nguyen Thi H L., Jutten C., Blind Source Separation for Convolutive Mixtures, *Signal Processing*, **45** (2), pp 209 – 229. 1995.

[8] Lee, T,W., Bell, A.J., and Orgmeister, R. Blind Source Separation of Real World Signals. In Proc. *I.E.E.E / I.C.N.N, International Conference on Neural Networks*, Vol 4, pp 2129 – 2134, 1997.

[9] Torkkola, K., Blind Separation of Convolved Sources Based on Information Maximisation, *IEEE Workshop on Neural Networks for Signal Processing*, NNSP'96, 1996.

[10] Van Gerven, S. Adaptive Noise Cancellation and Signal Separation with Applications to Speech Enhancement., PhD Thesis, Katholieke Universiteit Leuven, 1996.

[11] Durlach, N.I., Gabriel, K. J., Colbum, H. S., and Trahiotis, C. Interaural correlation discrimination: II. Relation to binaural unmasking, J. *Acoust. Soc. Am.* **79** (5), 1548-1557, 1986.

[12] Foldiak, P. Models of Sensory Coding, PhD Thesis, Physiological Laboratory, University of Cambridge, 1990.

[13]Lambert, R. Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures. PhD Thesis, University of Southern California, 1996.

[14] Girolami, M and Fyfe, C. A Temporal Model of Linear Anti-Hebbian Learning. Neural Processing Letters Journal, Vol 4, Issue 3, pp 1-10, 1997.

[15] Girolami, M. Symmetric Adaptive Maximum Likelihood Estimation for Noise Cancellation and Signal Separation. Electronics Letters, Vol, 33, No.17, pp 1437 – 1438, 1997.

[16] Cardoso, J, F. Infomax and Maximum Likelihood for Blind Source Separation, *I.E.E.E Signal Processing Letters*, **4**, pp 109 – 111, 1997.

[17] McDonald R. Signal to noise and idle channel performance of differential pulse code modulation systems- Particular applications to voice signals. *BSTJ*, **45** (1123-1151), 1966.

[18] Shields, P., Girolami, M., Campbell, D., Fyfe, C. Adaptive Processing Schemes Inspired by Binaural Unmasking for Enhancement of Speech Corrupted with Noise and Reverberation To Appear in Proc of 1'st European Workshop on Neuromorphic Systems, Stirling, Scotland, 1997.