Blind Source Separation for Adaptive Speech Control

Jonathan Harris[1], Tom Moir[2], and Fakhrul Alam[1]

[1]School of Engineering and Advanced Technology, Massey University, Auckland, New Zealand
[2]School of Engineering, Auckland University of Technology, Auckland, New Zealand

Abstract

Crosstalk resistant adaptive noise cancellation (CTRANC) is a method of separating convolutively mixed sources where little *a priori* information is known about the system. Possible areas of application for such an algorithm include speech signal processing, in telecommunications, and in the biomedical industry. In this paper we propose a novel adaptation to the traditional CTRANC which increases computational efficiency when the number of sources fits the requirement of $L = 2^n$ where $n \in \mathbb{Z}$ and $n > 1$. Preliminary results also show a modest improvement in separation performance when comparing it to the multiple-input multiple-output method proposed by Mei and Yin (2004).

1 Introduction

The blind source separation problem is the problem of trying to identify the individual sources with no *a priori* knowledge of the sources or the mixing system. Normally, all that can actually be acquired is different mixtures of the sources using multiple sensors. If only interested in the unmixed signals, this crosstalk has a detrimental effect on the usefulness of the acquired signal, and if significant enough, may render the raw signal totally unusable.

A real-life example of blind source separation is what is known as the “cocktail party problem”. Consider the case where there is a room of people, all of whom are talking simultaneously; the human brain is able to adequately extract one person’s speech from the rest. If an algorithm can be found to replicate these results, it would provide a very useful tool in the area of automatic speech recognition, which in turn could be used for speech control.

The application of such an algorithm need not be restricted solely to audio applications. In the medical world it could be used to isolate signals for electrocardiograms (ECGs) or electromyograms (EMGs) (Zhang and Cichocki 2000). It could also be used in telecommunications to reduce the crosstalk created by multiple transmitters (Pedersen et al. 2007). Another less obvious application for blind source separation is to separate images that have been mixed (Amari and Cichocki 1998), though this does not apply to CTRANCs.

One trivial way of solving this problem is to use a Widrow-Hoff least mean-squares (LMS) filter and an approximation of the noise signal to remove the noise from the mixture. However, this has the fundamental flaw that the noise signal has to be relatively signal-free. While there may be situations in which acquiring such a noise approximation is the quite plausible (for example, in a jet cockpit, where the engine noise can be obtained with negligible speech crosstalk), in the majority of everyday situations this assumption cannot be justified.

To overcome this problem, Zinser *et al.* (1985) proposed a cross-talk resistant adaptive noise canceller. The basic premise was that cross-coupling two LMS filters could result in an adaptive noise canceller that was not susceptible to crosstalk from the desired signal in the noise estimate.

In this paper, we propose a novel adaptation to the cross-talk resistant noise canceller that utilizes vector-LMS to increase the computational efficiency of the algorithm when dealing with 2^k input signals, where $k > 1$. We then show that the proposed algorithm actually slightly outperforms the multiple-input multiple-output (MIMO) cross-talk resistant adaptive noise canceller proposed by Mei and Yin (2004) in terms of input-output signal-to-noise ratios with while reducing computational complexity.

This paper is organized as follows. In section 2, the background information of all of the components required for the development of a CTRANC based of vector-LMS are discussed. Section 3 shows the derivation of the novel algorithm, and compares its computational complexity to the CTRANC proposed by Mei and Yin (2004). The experimental set-up and results are discussed in section 4, and the paper is then concluded in section 5.

2 Background Information

The Mixing System

We will first consider the case of a two-input, two-output (TITO) system. In matrix form this is

$$\mathbf{x}(t) = \mathbf{G}^T(t)\tilde{\mathbf{S}}(t) \qquad (1)$$

where $\mathbf{x}(t) = [x_1(t), \quad x_2(t)]^T$, the superscript ^{T} denotes the transpose operator,

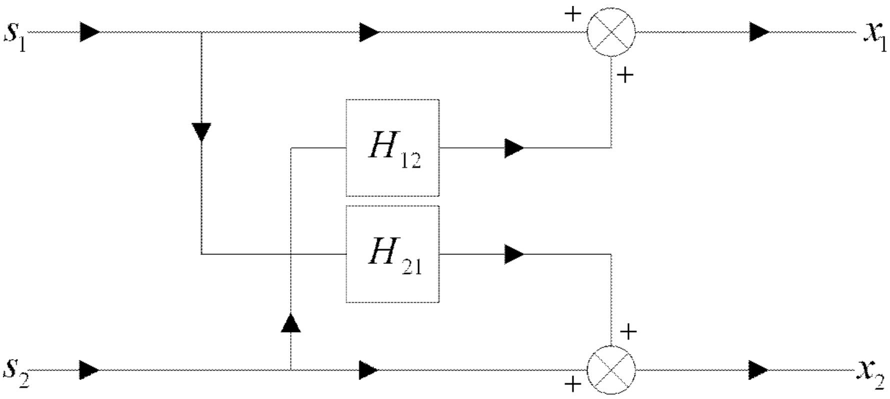

Fig. 1. *The simplified mixing system*

t denotes the time index,

$$\mathbf{G}(t)=\left[G^{0}(t)\quad G^{1}(t)\quad \ldots\quad G^{n-1}(t)\right]^{T}$$

is the mixing matrix with n taps (the superscript number indicates the tap index), and

$$\begin{aligned}\tilde{\mathbf{S}}(t)\ &=\ \left[\tilde{s}_1(t)\quad \tilde{s}_2(t)\quad \tilde{s}_1(t-1)\quad \tilde{s}_2(t-1)\right.\\ &\qquad \left.\ldots\quad \tilde{s}_1(t-n+1)\quad \tilde{s}_2(t-n+1)\right]\end{aligned}$$

However, because we are more interested in the separation of the signals rather than the deconvolution of them, we take the assumption that the channels between each source and the closest microphone are simply the Kronecker delta function. This simplifies the problem because it means that we only have to account for two unknown filters rather than four. Fig. 1 shows the simplified mixing system.

On the other hand, this means that at best, we will separate the signals only up to filtered versions of the original. In order to find the original unfiltered versions of the sources, blind dereverberation is needed. This is a very difficult problem when only given one instance of the filtered speech; temporal whitening is not recommended since pure speech is naturally temporally correlated (Douglas 2003), and temporal whitening would make the speech sound unnatural. On the other hand, temporal decorrelation may have its uses in applications where listening to the signal is not needed - e.g. automatic speech recognition.

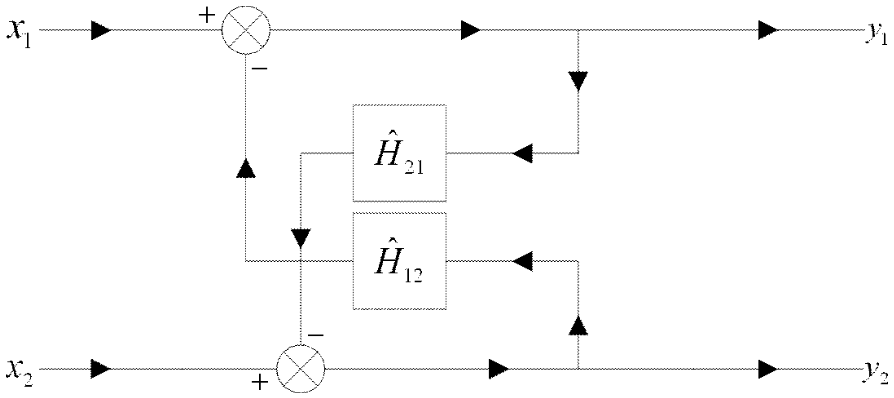

Fig. 2. The separating system

The Cross-talk Resistant Adaptive Noise Canceller

Zinser et al. proposes an adaptation to the LMS filter in order to make it more resilient to cross-talk. Rather than feeding the noise estimate directly into the LMS filter as a reference, he describes how a second LMS filter can be used to remove any of the crosstalk from the noise estimate, resulting in a better noise approximation (Zinser et al. 1985). Fig. 2 shows a block diagram of the backward-separation system. It can be seen that as $\hat{H}_{12}$ and $\hat{H}_{21}$ converge to H_{12} and H_{21} respectively, y_1 and y_2 will converge to s_1 and s_2 respectively. Note that permutation of the order of inputs to outputs cannot occur. This permutation occurs where there is no guarantee that any specific source will be mapped to a specific output. The inability to permute differentiates the CTRANC from other methods of blind separation (such as independent component analysis (Comon 1995)), which is based purely on the independence of the outputs. However, this is based on the assumption that each microphone is the closest microphone to a unique source.

Mei and Yin expand on this idea to derive the following simplified equation updates for the filters $\hat{H}_{12}$ and $\hat{H}_{21}$ (Mei and Yin 2004).

$$\hat{H}_{12}(t+1) = \hat{H}_{12}(t) - \mu_1 y_1(t)\mathbf{Y}_2(t)$$

$$\hat{H}_{21}(t+1) = \hat{H}_{21}(t) - \mu_2 y_2(t)\mathbf{Y}_1(t)$$

where μ_1 and μ_2 are the positive learning rates, $y_1(t)$ and $y_2(t)$ are the estimates of the separated signals at time t, and

$$\mathbf{Y}_1(t) = \left[y_1(t-1),\quad y_1(t-2),\quad \ldots,\quad y_1(t-n)\right]^T$$

$$\mathbf{Y}_2(t) = \left[y_2(t-1),\quad y_2(t-2),\quad \ldots,\quad y_2(t-n)\right]^T$$

Vector-LMS

While ordinary LMS will find the transversal filter weights when given both the input and output of a filter, vector-LMS will find the mixing system given the inputs and outputs of the mixing system. For example, if we applied vector-LMS to two-input two-output system shown in equation (1), the matrix-polynomial of the filter would converge to $\mathbf{G}$. Batra and Barry show the derivation of the vector LMS algorithm

$$\widehat{\mathbf{G}}(t+1)=\widehat{\mathbf{G}}(t)+\mu\mathbf{S}(t)\mathbf{e}(t)^T$$

where $\widehat{\mathbf{G}}(t)$ is the estimate at time t of the mixing polynomial matrix $\mathbf{G}$, μ is the step size, $\mathbf{S}(t)=\left[\mathbf{s}^T(t),\quad \mathbf{s}^T(t-1),\quad \dots,\quad \mathbf{s}^T(t-n)\right]^T$ is a vector of length $2n$ of the inputs where n is the filter order, and $\mathbf{e}(t)$ is a length-2 vector of the errors between the desired filter output $\mathbf{x}$ and its actual output $\hat{\mathbf{x}}$ where $\hat{\mathbf{x}}=\mathbf{G}^T\mathbf{S}$ (Batra and Barry 1995).

In this paper, we develop a crosstalk resistant adaptive noise canceller that utilizes vector-LMS to obtain a multibranched-recursive structure, creating a more modular algorithm with increased computational efficiency.

3 The Cross-coupled Vector-LMS

In order to show the working of the CTRANC based on vector-LMS, we will consider the situation of four inputs and four outputs. In Fig. 3 we have a matrix polynomial representation of the mixing system, where $\tilde{\mathbf{s}}_1=\left[\tilde{s}_1,\quad \tilde{s}_2\right]^T$ and $\tilde{\mathbf{s}}_2=\left[\tilde{s}_3,\quad \tilde{s}_4\right]^T$ are the four inputs multiplexed into two vectors, $\mathbf{x}_1=\left[x_1,\quad x_2\right]^T$ and $\mathbf{x}_2=\left[x_3,\quad x_4\right]^T$ are the four outputs multiplexed into two vectors, and $\mathbf{G}_{11}$, $\mathbf{G}_{12}$, $\mathbf{G}_{21}$, and $\mathbf{G}_{22}$ are all mixing polynomial matrices representing the entire mixing system. Note that these should not be confused with their scalar counterparts. Using the same reasoning as with the ordinary CTRANC, we derive the following update equations for the separating polynomial matrices $\widehat{\mathbf{H}}_{12}$ and $\widehat{\mathbf{H}}_{21}$.

$$\widehat{\mathbf{H}}_{12}(t+1)=\widehat{\mathbf{H}}_{12}(t+1)+\mu_1\mathbf{Y}_2(t)\mathbf{y}_1^T(t)$$

$$\widehat{\mathbf{H}}_{21}(t+1)=\widehat{\mathbf{H}}_{21}(t+1)+\mu_2\mathbf{Y}_1(t)\mathbf{y}_2^T(t)$$

where μ_1 and μ_2 are convergence weights, $\mathbf{y}_1$ and $\mathbf{y}_2$ are the length-2 output

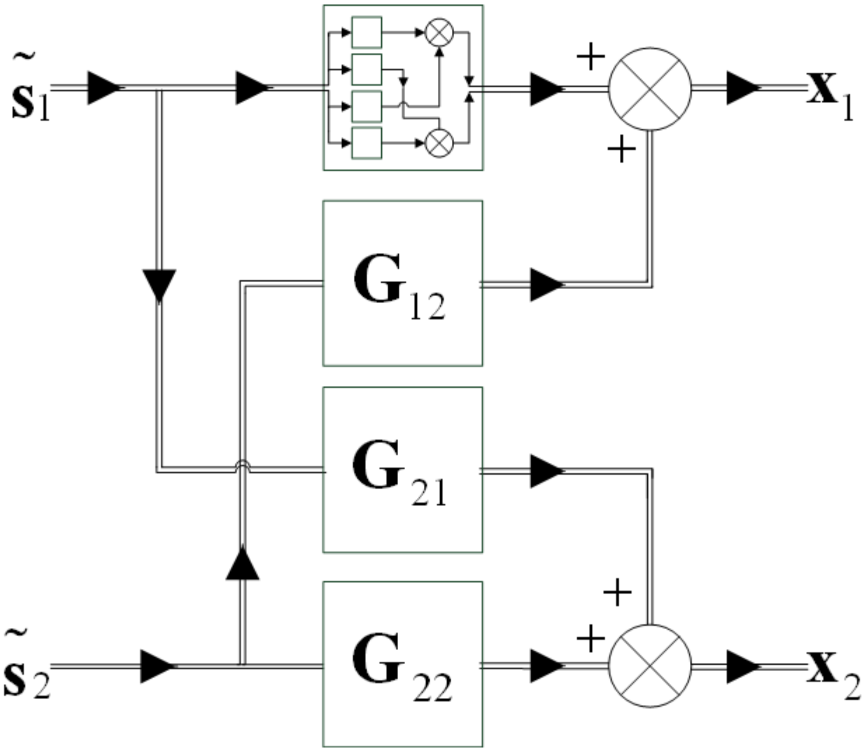

Fig. 3. The four input mixing system

vectors $[y_1, \quad y_2]$ and $[y_3, \quad y_4]$ respectively, and the length-$2n$ vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are defined by

$$\mathbf{Y}_1 = \left[\mathbf{y}_1^T(t-1), \quad \mathbf{y}_1^T(t-2), \quad \ldots, \quad \mathbf{y}_1^T(t-n)\right]^T$$

$$\mathbf{Y}_2 = \left[\mathbf{y}_2^T(t-1), \quad \mathbf{y}_2^T(t-2), \quad \ldots, \quad \mathbf{y}_2^T(t-n)\right]^T$$

Essentially what this algorithm will do is separate a system of four mixed sources into two systems of two mixed sources. One can then apply the algorithm from an ordinary CTRANC to separate each of the sources into approximations of the original individual signals.

Number of Inputs	Proposed Method		Mei and Yin Method
	Separate all	Extract one	
4	$24n+32$	$20n+26$	$24n+36$
8	$112n+136$	$84n+98$	$112n+168$
16	$480n+544$	$340n+370$	$480n+720$

Table 1. Multiplication operations required.

Number of Inputs	Proposed Method		Mei and Yin Method
	To separate all	To extract one	
4	$12n+20$	$10n+16$	$12n+24$
8	$56n+80$	$42n+56$	$56n+112$
16	$240n+304$	$170n+200$	$240n+480$

Table 2. Addition/subtraction operations required.

Computational Efficiency

Mei and Yin (2004) proposed an adaptation to the TITO CTRANC that extended it for use with more than two input signals. This was simply an extension of the two-channel case. For example, with three sources each input needed two LMS filters removing the crosstalk from the other two channels. Thus the computational complexity of their algorithm was equivalent to $L(L-1)$ LMS algorithms. The multibranched recursive approach that we propose is more efficient under certain conditions as will now be shown.

We will now consider the computational requirements for the proposed algorithm. With $L=2^k$ inputs, it requires two 2^{k-1}-vector LMS algorithms, four 2^{k-2}-vector LMS algorithms, etc. The number of multiplication and addition/subtraction operations for each vector LMS algorithm is given by the following equations.

$$2(n+1)M^2+M \qquad \text{multiplications}$$

$$(n+1)M^2+M \qquad \text{additions/subtractions}$$

where n is the filter size and M is the size of the input/output vectors. These equations also work for scalar LMS, when $M=1$.

Another advantage in the proposed method is that its modular structure allows the removal of portions that may be unnecessary. For example, in an eight-input system, if only one source needs to be extracted, and it is known which output channel that source maps to, then six scalar LMS and two 2-vector LMS algorithms can be discarded. This allows for further computational savings.

Tables 1 and 2 show the multiplication and addition/subtraction requirements for 4, 8, and 16 input systems for the cases where all sources need to be extracted, and when only one needs to be extracted. These results show that the current method is more computationally efficient than that proposed by Mei and Yin for all given cases.

4 Separation Performance

We conducted a simple experiment to discover the relative separation of the proposed method to the method in (Mei and Yin 2004).

Experimental Procedure

The experiment was set up as follows: four microphones were placed as four corners of a 0.2m × 0.2m square near the middle of a 4m × 7m room furnished with a lounge suite, a piano and a dining room suite. There were three noise sources, all samples of a car assembly line from the file labeled ‘factory floor noise 2’ from the NOISEX database. The speech was created by using a loudspeaker playing the speech sample in the package ‘Lunatick-20080326–cc.tgz’ from the VoxForge speech corpus. The algorithm was implemented using NI LabVIEW.

Using the described set-up, we used the proposed algorithm to reduce the noise level. Each filter had 1000 tap-weights. We chose this number because increasing the number of tap weights beyond 1000 increased computational complexity with a negligible increase in SNR, while decreasing the number of tap-weights adversely affected the results. Because we do not have the power of the desired signal by itself, to calculate the SNR, we net to use the following formula

$$SNR = 10\log_{10}\left(\frac{P_{SN} - P_N}{P_N}\right) \tag{2}$$

where P_{SN} is the combined power of the speech with the noise and P_N is the power of the noise. This is based on the assumption that the noise and the speech are statistically independent.

Results

Using the formula for calculating signal-to-noise ratios given in equation (2), we obtained the results as shown in Table 3. In an informal listening test, we also found that the speech was more comprehensible in the separated signals than in the mixed signals.

	Input SNR	Output SNR
Proposed Method	7.7 dB	14.2 dB
Mei and Yin Method	7.7 dB	13.9 dB

Table 3. Increases in SNR

There is a modest gain in the SNR for the proposed method when comparing it to the method described by Mei and Yin. This indicates that the proposed method can perform separation at least as well as the method proposed by Mei and Yin, while saving in computational complexity.

5 Conclusion

One solution to the blind source problem is to use a cross-talk resistant noise canceller to separate the signals. This paper describes an adaptation to the CTRANC algorithm to increase its computational efficiency. Experimental data shows that there is a modest increase in performance due to these adaptations. It also has the advantage that it is potentially even more computationally efficient if there is only one desired source, and it is known which channel it will be separated to. In future studies we propose to incorporate this method with an automatic speech recognition system, and evaluate its performance in that capacity.

6 References

L. Zhang and A. Cichocki, (2000) Blind deconvolution of dynamical systems: A state space approach," Journal of Signal Processing, vol. 4, no. 2, pp. 111-130.

M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, (2007) "A survey of convolutive blind source separation methods," Multichannel Speech Processing Handbook.

S. Amari and A. Cichocki, (1998) "Adaptive blind signal processing-neural network approaches," Proceedings of the IEEE, vol. 86, no. 10, pp. 2026– 2048.

S. V. Gerven and D. V. Compernolle, (1995) "Signal separation by symmetric adaptive decorrelation: stability, convergence, and uniqueness," IEEE Transactions on Signal Processing, vol. 43, no. 7, pp. 1602–1612.

S. Douglas, (2003) "Convolutive blind separation of speech mixtures using the natural gradient," Speech Communication, vol. 39, no. 1-2, pp. 65–78.

R. Zinser, G. Mirchandani, and J. Evans (1985) "Some experimental and theoretical results using a new adaptive filter structure for noise cancellation in the presence of crosstalk," in Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85., vol. 10, 1985, pp. 1253– 1256.

P. Comon, (1994)"Independent component analysis, a new concept?" Signal processing, vol. 36, no. 3, p. 287-314.

T. Mei and F. Yin, (2004) "Blind separation of convolutive mixtures by decorrelation," Signal Processing, vol. 84, no. 12, pp. 2297–2313.

A. Batra and J. Barry, (1995) “Blind cancellation of co-channel interference,” in Proceedings of GLOBECOM ’95, Singapore, 1995, pp. 157–162