28-8-2005

# An investigation of temporal modeling in blind signal separation

Daniel Smith
*University of Wollongong*

Jason Lukasiak
*University of Wollongong*, jl01@ouw.edu.au

Ian Burnett
*University of Wollongong*, ianb@uow.edu.au

# An investigation of temporal modeling in blind signal separation


## Abstract

This paper investigates the performance of blind signal separation (BSS) algorithms that exploit the temporal predictability of speech. Specifically, the investigation considers how the separation performance of two BSS algorithms will be affected when the length of the AR process (used in the algorithms to model speech) is varied. The investigation concludes that the length of the AR process (prediction order) has a significant impact on separation performance. In particular, the separation performance of both algorithms is degraded, if the AR model's prediction order, over fits, or under fits, the temporal structure of the speech. It is revealed that a prediction order of 30- 50 provides maximum separation performance for natural speech, however a prediction order of 10 is more applicable if computational cost is a consideration.



## Disciplines

Physical Sciences and Mathematics

## Publication Details

This paper originally appeared as: Smith, D, Lukasiak, J & Burnett, I, An investigation of temporal modeling in blind signal separation, Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, 28-31 August 2005, vol 2, 503-506.

# AN INVESTIGATION OF TEMPORAL MODELING IN BLIND SIGNAL SEPARATION

*Daniel Smith, Jason Lukasiak and Ian Burnett*

School of Electrical, Computer and Telecommunications Engineering
University of Wollongong
dsmith@titr.uow.edu.au

## ABSTRACT

This paper investigates the performance of blind signal separation (BSS) algorithms that exploit the temporal predictability of speech. Specifically, the investigation considers how the separation performance of two BSS algorithms will be affected when the length of the AR process (used in the algorithms to model speech) is varied. The investigation concludes that the length of the AR process (prediction order) has a significant impact on separation performance. In particular, the separation performance of both algorithms is degraded, if the AR model's prediction order, over fits, or under fits, the temporal structure of the speech. It is revealed that a prediction order of 30-50 provides maximum separation performance for natural speech, however a prediction order of 10 is more applicable if computational cost is a consideration.

## 1. INTRODUCTION

Over the past decade, Blind Signal Separation (BSS) has been a major area of interest within speech processing research. This is largely due to its potential to solve the "cocktail party" problem, where any speaker in an acoustic environment can be retrieved from a mixture of other speakers and noise [1]. Conventional BSS employs Independent Component Analysis (ICA) to address the "cocktail party" problem, however fails to exploit any a priori knowledge of the production mechanisms of speech. This failure has been addressed by approaches developed specifically for speech in [2, 3], however a group of BSS techniques developed for more general application [1, 4, 5, 6, 7], also have an inherent connection to speech production mechanisms. These approaches separate on the basis of a signal's temporal predictability and can be successfully applied to speech separation due to the existence of temporal correlation (structure) in the speech signal [8].

The BSS approaches of [1, 4, 6, 7] model the temporal structure of signals in the mixture using an auto-regressive (AR) process. The AR process predicts a signal $S_j(t)$ as a linear combination of its previous $P$ samples:

$$S_{jp}(t) = \sum_{i=1}^{P} b_{ji} \cdot S_j(t-i) \qquad j = 1...N \qquad (1)$$

where $S_{jp}(t)$ is the predicted signal and $b_j = [b_{j1}.....b_{jP}]$ is a $1 \times P$ vector of prediction coefficients. As the length of prediction filter $b_j$ ($P$, also known as the prediction order) used to model underlying signals in the mixture must be determined a priori, the choice of prediction order is important. This is because the prediction order has a significant impact on how the temporal structure of underlying signals in the mixture are modeled, which in turn, effects the separation performance [4]. This was demonstrated in [6] for a mixture of 4 physiological signals with different temporal structures. As the prediction order was varied between 1 and 400, the signal extracted from the mixture changed, as did the signal's separation performance. It was concluded that a signal was more likely to be extracted from the mixture if the prediction order was large enough to capture the temporal structure, yet not so large that the signal was no longer predictable.

While [6] informally reports on the effect that prediction order has on signal extraction, in this paper we consider the issue in greater detail and with particular emphasis on speech separation. This analysis provides detailed insight into the relationship between the prediction order and separation performance of BSS algorithms that exploit the temporal correlation of signals. Mixtures of artificially generated speech, constrained with fixed AR structures, are used as inputs to two different BSS algorithms. The first algorithm [4] (ARalg) models the AR structure of the speech signals exclusively, while the second algorithm (AR-F0alg) recently proposed in [9], provides a more complete model of speech production mechanisms. The latter jointly models the AR structure and periodicity of the signal.

The final stage of this analysis investigates the influence that the prediction order has on the separation of natural speech. We propose a range of prediction orders that are, in general, suitable for the separation of natural speech.

### 1.1. Problem Formulation

The BSS problem can be formulated as follows: The vector of sensor signals ($X(t)$) are observations of the vector of signals ($S(t)$) linearly mixed according to the system $A$:

$$X(t) = A \cdot S(t) \qquad (2)$$

where $X(t) = [X_1...X_M]^T$ is a $M \times 1$ vector of mixed observations, $S(t) = [S_1....S_N]^T$ is an unknown

0-7803-9243-4/05/$20.00 ©2005 IEEE

$N \times 1$ vector of signals and $A$ is an unknown $M \times N$ non-singular matrix. In this approach it is assumed that A contains scalar elements (instantaneous mixing) and the system is square, i.e. the number of signals is equal to the number of sensors ($N$=$M$).

In order to obtain a scaled permutation of the original signals $c \cdot S(t)$, given only mixed observations $X(t)$, an $N \times M$ separation matrix $W$(estimating $A^{-1}$) must be computed and subsequently multiplied by $X(t)$. In contrast to simultaneously estimating the entire separation matrix, the two algorithms used in this investigation are sequential approaches that estimate each column of the separation matrix ($W_j$) and separated signal ($S_{je}(t) = W_j^T \cdot X(t)$) individually.

## 2. DESCRIPTION OF BSS ALGORITHMS

The BSS algorithms employed in this investigation extract speech signals from a mixture by exploiting the following assumption:

(a) *A single speaker has more temporal correlation than any linear combination of mixed speakers*

The first BSS approach (ARalg) represents the signal's temporal correlation with an AR model (shown in (1)). The AR model is incorporated into ARalg's cost function ($W_j^T \cdot xa(t)$), which is the first term of $\xi(t)$ in (3). A gradient descent approach is used to adapt system parameters $W_j$ and $b_j$ towards the minima of the cost function. When the minima of the cost function is reached, the mixture has maximum temporal correlation, which in accordance with assumption (a), indicates that a clean signal estimate has been obtained. A full derivation of the learning rules and specific details of ARalg can be found in [4].

Our proposed BSS approach [9] (AR-F0alg) is an extension of the ARalg model, which provides a more complete representation of speech temporal correlation. AR-F0alg jointly models the AR structure and periodicity of speech ($F0^{-1}$) in the cost function $C(W_j, b_j, B_j)$ as:

$$\begin{aligned} C(W_j, b_j, B_j) &= 1/2 * E[\xi(t)^2] \\ \xi(t) &= W_j^T \cdot xa(t) - B_j \cdot W_j^T \cdot xl(t) \end{aligned} \quad (3)$$

where $\xi(t)$ is the error function of the joint AR-F0 model, $xa(t) = X(t) - \widehat{X}(t) \cdot b_j^T$, $\widehat{X}(t) = [X(t-1)...X(t-P)]$ is a $M \times P$ matrix, $B_j$ is the long term prediction gain and $xl(t) = X(t - F0^{-1}) - \widehat{X}(t - F0^{-1}) \cdot b_j^T$.

### 2.1. Derivation of the Joint AR-F0 Algorithm's Learning Algorithm

In order to reach the minima of AR-F0alg's cost function in (3), stochastic gradient descent is used to derive the adaptation rules for the parameter set $W_j$, $b_j$ and $B_j$.

The initial step in deriving adaption rules, involves computing the partial derivatives of $C(W_j, b_j, B_j)$ with respect to each of the parameters $W_j$, $b_j$ and $B_j$. The partial derivatives are calculated as:

$$\begin{aligned} \frac{\delta C(W_j, b_j, B_j)}{\delta W_j} &= E[\xi(t) \cdot (xa(t) - B_j \cdot xl(t))] \\ \frac{\delta C(W_j, b_j, B_j)}{\delta b_j} &= E[-\xi(t) \cdot W_j^T \cdot \widehat{x}] \\ \frac{\delta C(W_j, b_j, B_j)}{\delta B_j} &= E[-\xi(t) \cdot W_j^T \cdot xl(t)] \end{aligned} \quad (4)$$

where $\widehat{x} = \widehat{X}(t) - B_j \cdot \widehat{X}(t - F0^{-1})$. The learning rules in (5) are derived by substituting the derivatives from (4) into the stochastic gradient descent approach:

$$\begin{aligned} W_{j+1} &= W_j - uW \cdot E[\xi(t) \cdot (xa(t) - B_j \cdot xl(t))] \\ b_{j+1} &= b_j + ub \cdot E[\xi(t) \cdot W_j^T \cdot \widehat{x}] \\ B_{j+1} &= B_j + uB \cdot E[\xi(t) \cdot W_j^T \cdot xl(t)] \end{aligned} \quad (5)$$

where $uW$, $ub$ and $uB$ are the step sizes, and $W_{j+1}$, $b_{j+1}$ and $B_{j+1}$ are the parameters to be used in the next iteration of the gradient descent. The adaptation concludes after the algorithm converges to $C(W_j, b_j, B_j)_{min}$. Under assumption (a), at $C(W_j, b_j, B_j)_{min}$, $W_j$ can be used to estimate a scaled version of an original signal $S_{je} = c \cdot S_j$.

$X(t)$ are initially whitened, so that $W$ is constrained to the space of orthonormal matrices. This is particularly beneficial in ill conditioned problems. In addition, $W_j$ is normalised i.e. $\frac{W_j}{\|W_j\|^2}$, after each iteration of the gradient descent, so that the estimated signal is constrained to $E[S_{je}^2] = 1$. This ensures that the trivial solution $S_{je} = 0$ is avoided when finding $C(W_j, b_j, B_j)_{min}$. $F0$ is also estimated from a clean speech estimate $W_j \cdot X(t)$ after each iteration of the gradient descent.

## 3. PREDICTION ORDER AND SEPARATION PERFORMANCE INVESTIGATION

The data set used in this investigation consisted of both artificially generated signals with known AR structures of order 10 imposed upon them, and natural speech signals. The artificial signals were generated as basic stationary models of voiced and unvoiced speech, by imposing 30 different AR structures (derived from speech), upon 30 different periodic and i.i.d Gaussian excitations respectively. The periodic excitations were generated by applying a linear prediction filter [8] to real vowels. The prediction order was chosen to be slightly less than the $F0^{-1}$ of the voiced signal, so that all of the temporal correlation apart from the period of the vowels was removed. The AR structures were then imposed upon the periodic and Gaussian excitations, by applying an IIR 10th order linear prediction synthesis filter [8] with coefficients derived from speech.

The artificial signals were 1s in duration, while the speech signals were 3s long. All signals were sampled at 8KHz. The 30 artificial unvoiced signals were grouped

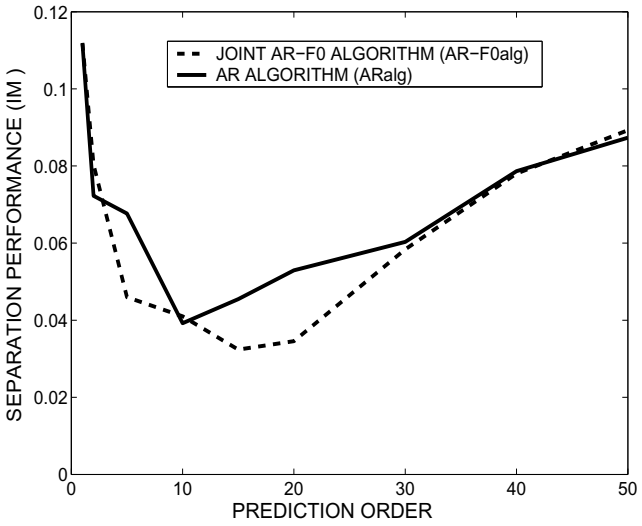


**Fig. 1**. Average $IM$ across 15 mixed pairs of artificial unvoiced speech. Prediction order ranges from 1-50.

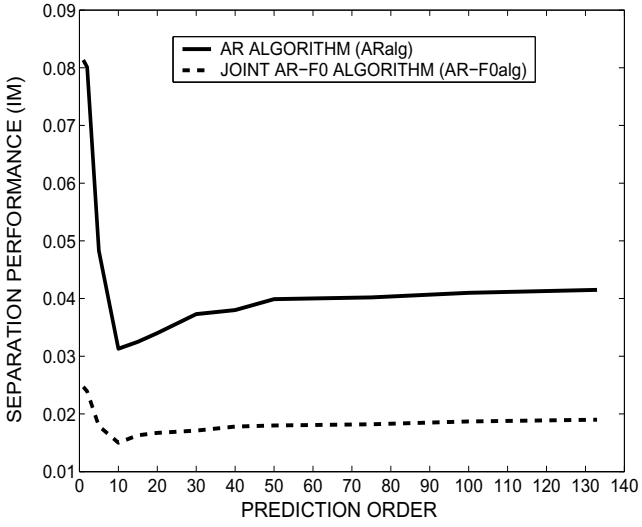


**Fig. 2**. Average $IM$ across 15 mixed pairs of artificial voiced speech. Prediction order ranges from 1-133.

into pairs and then mixed together by the same stationary mixing system $A$, forming 15 mixed pairs. The same mixing process was employed for the artificial voiced and natural speech independently, so that 15 pairs of artificial voiced and 15 pairs of natural speech mixtures were also generated. The investigation was conducted by applying 30ms non-overlapped frames of the mixtures to the BSS algorithms AR-F0alg and ARalg, across a range of prediction orders that varied from 1 to 133. In the unvoiced mixtures however, the range of prediction orders were limited to 1 to 50. The step sizes of $uW = ub = uB = 0.05$ were employed in both algorithms.

The measure of separation performance used in this investigation was the Interference Measure $(IM)$. For the extraction of a single signal, $IM = \frac{(pp^T - max(p)^2)^{\frac{1}{2}}}{max(p)}$ where $p = W_j^T \cdot A$ and $IM = 0$ indicates perfect estimation of a signal from the mixture.

### 3.1. Artificial Voiced-Unvoiced Speech Investigation

The short, fixed temporal structures of the artificial signals don't strictly correspond to the temporal structure of natural speech, which is time-varying and often significantly longer than an AR order of 10 [8]. The following investigation, however, enables firm conclusions to be drawn on the impact of prediction order on modeling signals in the mixture, and hence the effect on separation performance.

Figure 1 compares the average separation performance of the ARalg and AR-F0alg algorithms for the 15 mixed pairs of the artificial unvoiced speech. The results show that the separation performance $(IM)$ of ARalg is maximised for a prediction order of 10 and steadily degrades for prediction orders both greater than and less than this value (the order of 10 corresponds exactly to the imposed AR 10 structure). When the prediction order exceeds the inherent order of the temporal structure of the signal to be separated, the AR parameterisation over fits the signal and potentially captures formants of additional signals. This results in weakened separation performance. In contrast, when the prediction order is less than that of the temporal structure of the underlying signal, the separation model is inadequate; hence, degraded separation performance results.

Figure 1 shows that the relationship between the prediction order and separation performance of the AR-F0alg algorithm resembles ARalg. The most significant difference is that the prediction order that provides the best separation performance for AR-F0alg is 15. The apparent contradiction of best modeling an AR 10 signal with a prediction order of 15, can be attributed to the interaction of jointly modeling the AR structure and periodicity of artificial unvoiced speech that doesn't possess long term temporal correlation (periodicity).

Figure 2 compares the average separation performance of the ARalg and AR-F0alg algorithms for the 15 mixed pairs of the artificial voiced speech. The ARalg and AR-F0alg models also support the results from Figure 1, as they exhibit the same maximum separation performance at a prediction order of 10, and a degradation in performance as the prediction order alters from 10.

The results from Figure 2 also indicate that the AR-F0alg algorithm's separation performance is 52-75% superior to ARalg across all prediction orders. This is because AR-F0alg provides a more complete model of the temporal structure of the voiced speech, jointly modeling their AR-10 structure and periodicity. Even when ARalg is given sufficient order to model some of underlying signal's periodicity (prediction orders greater than 25), AR-F0alg sustains higher performance. As mentioned in the discussion of unvoiced speech, this is due to long prediction filters which over fit the signal and hence model formants in additional signals.

### 3.2. Speech Analysis

Figure 3 compares the average separation performance of the ARalg and AR-F0alg algorithms for the 15 mixed

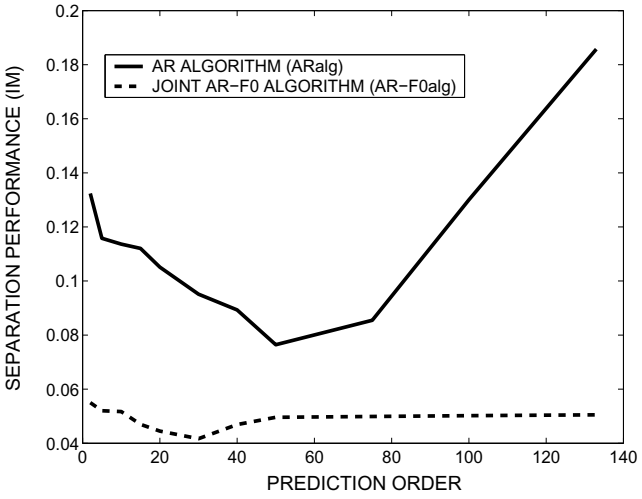


**Fig. 3**. Average $IM$ across 15 mixed pairs of natural speech signals. Prediction order ranges from 1-133.

pairs of natural speech. It shows that the AR modeling of natural speech is influenced by the same factors as the artificial signals in Section 3.1.

The prediction orders which best capture the temporal structure (in an average sense) and produce maximum average separation performance for ARalg and AR-F0alg are 50 and 30 respectively. Using these optimal prediction filters within the algorithms is costly however, due to the exponential increase in computational complexity [8] with prediction order. If computational cost is a consideration in the algorithm's application, a prediction order of around 10 should be employed; this is commonly used in speech linear prediction coding at 8kHz [8]. This is suitable for AR-F0alg in particular, as the $IM$ of AR-F0alg (in Figure 3) only increases by 0.01 between a prediction order of 10 and the optimal order of 30.

## 4. CONCLUSION

In this paper, we investigated the relationship between the separation performance of BSS algorithms and the AR structure that the algorithms employ to model the signal's temporal correlation. The analysis indicated that the prediction order of the AR model had significant impact on the separation performance of the AR-F0alg and ARalg algorithms. It was revealed that separation performance is improved by using an AR model with a long enough prediction order to capture the temporal structure of signals in the mixture. However, increasing the prediction order of the AR model does not necessarily correspond to improved separation performance. This is because an excess of AR parameters can over fit the underlying signal in the mixture, in such cases, the AR model captures formants of additional signals. It was also shown that the separation performance of the joint AR-F0alg model was superior to ARalg for all signals. Even when ARalg captured the periodicity of the signals (by employing longer prediction filters), the performance of AR-F0alg was superior, as it explicitly incorporated the periodicity into its model.

Finally, the analysis revealed that the prediction orders that provide maximum separation performance for speech were quite long (order of 30-50) and computationally expensive. Thus it was proposed that a smaller prediction order of 10 would be suitable, especially if computational complexity was a major concern, as in the case of real time applications.

## 5. REFERENCES

[1] A. Cichoki and S. Amari, *Adaptive Blind Signal and Image Processing : Learning Algorithms and Applications*, John Wiley & Sons, 2002.

[2] A. Acero, S. Altschuler, and L. Wu, "Speech/Noise Separation Using Two Microphones and a VQ Model of Speech Signals," in *Proc.Int Conf on Spoken Language Processing*, Beijing, Oct 2000, pp. 613–619.

[3] F. Tordini and F. Piazza, "A Semi-Blind Approach to the Separation of Real World Speech Mixtures," in *Proc. IJCNN02*, Honolulu, May 2002, pp. 1293–1298.

[4] R. Thawonmas and A. Cichoki, "Blind Signal Extraction of Arbritary Distributed but Temporally Correlated Signals-Neural Network Approach," *IEICE Trans.Fundamentals*, vol. E82-A, no. 9, pp. 1834–1844, Sept 1999.

[5] A. Barros and A. Cichoki, "Extraction of Specific Signals with Temporal Structure," *Neural computation*, vol. 13, pp. 1995–2003, 2001.

[6] D. Mandic and A.Cichocki, "An Online Algorithm for Blind Extraction of Sources with Different Dynamical Structures," in *Proc.ICA2003*, Nara, Apr 2003, pp. 645–650.

[7] B. Pearlmutter and L. Parra, "Maximum Likelihood Blind Source Separation: A Context-Sensitive Generalization of ICA," in *Advances in Neural Information Processing Systems 9*, MIT Press, Ed., Denver, Dec 1996, pp. 613–619.

[8] L. Rabiner and R.Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[9] D. Smith, J. Lukasiak, and I. Burnett, "Blind Speech Separation using a Joint Model of Speech Production," *Accepted. IEEE Signal Proc.Letters*.