30-10-2005

A sequential approach to sparse component analysis

Daniel Smith
University of Wollongong

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

I. Burnett
University of Wollongong, ianb@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/infopapers>

Part of the [Physical Sciences and Mathematics Commons](#)

Recommended Citation

Smith, Daniel; Lukasiak, Jason; and Burnett, I.: A sequential approach to sparse component analysis 2005.
<https://ro.uow.edu.au/infopapers/546>

A sequential approach to sparse component analysis

Abstract

A sequential approach to sparse component analysis (SeqTIF) is proposed in this paper. Although SeqTIF employs the estimation process of the simultaneous TIFROM algorithm, a source cancellation and deflation technique are also incorporated to sequentially estimate speech signals in the mixture. Results indicate that SeqTIF's separation performance shows a clear improvement upon the simultaneous TIFROM approach, due to the less restrictive assumptions it places upon the signals in the mixture. In particular, the analysis indicates SeqTIF's data efficiency is high, enabling the sequential approach to track a time-varying mixture with much greater accuracy than the simultaneous algorithm. Furthermore, SeqTIF is a more flexible approach, free from the constraints that a simultaneous approach places upon the mixing system.

Disciplines

Physical Sciences and Mathematics

Publication Details

This paper was originally published as: Smith, D, Lukasiak, J & Burnett, I, A sequential approach to sparse component analysis, 2005 IEEE 7th Workshop on Multimedia Signal Processing (MMSP-05), Shanghai, China, 30 October-2 November 2005, 1-4. Copyright IEEE 2005.

A Sequential Approach to Sparse Component Analysis

Daniel Smith, Jason Lukasiak and Ian Burnett

School of Electrical, Computer and Telecommunications Engineering

University of Wollongong

Email: dsmith@titr.uow.edu.au

Abstract—A sequential approach to Sparse Component Analysis (SeqTIF) is proposed in this paper. Although SeqTIF employs the estimation process of the simultaneous TIFROM algorithm [1], a source cancellation and deflation technique are also incorporated to sequentially estimate speech signals in the mixture. Results indicate that SeqTIF's separation performance shows a clear improvement upon the simultaneous TIFROM approach, due to the less restrictive assumptions it places upon the signals in the mixture. In particular, the analysis indicates SeqTIF's data efficiency is high, enabling the sequential approach to track a time-varying mixture with much greater accuracy than the simultaneous algorithm. Furthermore, SeqTIF is a more flexible approach, free from the constraints that a simultaneous approach places upon the mixing system.

I. 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 [2]. Conventional BSS employs Independent Component Analysis (ICA) to address the 'cocktail party' problem, requiring speech signals to be statistically independent, non-Gaussian and stationary in order to separate them successfully [2]. However, these assumptions are not always satisfied by speech signals, thus BSS techniques have considered an alternate mechanism, known as Sparse Component Analysis (SCA) [3]. SCA as employed in [1], [4]–[9], separates speech signals by exploiting their non-overlapping (sparse) time-frequency (T-F) representations in the mixture. The approaches in [1], [4]–[7], [9], however, estimate signals from the mixture concurrently using a simultaneous approach. Each speech signal in the mixture is required to possess an adequately sparse representation to be estimated with any accuracy. In [6], the authors show that each speech signal in a mixture should possess an approximate sparse representation, as each time-frequency window consists of only one speech signal. However, this approximation becomes less valid when speech signals have formants that overlap or when the number of signals in the mixture increases [6].

This paper addresses the estimation problems associated with simultaneous SCA approaches by proposing an SCA algorithm that employs a sequential approach to separation. A sequential approach to BSS extracts signals from the mixture one by one, with the process being repeated until all signals have been retrieved [10]. We hypothesize that a sequential approach to SCA will improve the separation performance of a simultaneous approach, especially when signals in the

mixture are not sparsely represented. Furthermore, a sequential approach to SCA removes the constraints that a simultaneous approach places upon the mixing system.

While writing this paper, the authors became aware of another sequential approach to SCA, the 'successive source cancellation' method in [8]. This method, however, does not address the estimation problems of the simultaneous approach. In order to separate the signals, [8] places constraints upon the mixing system and requires all signals in the mixture to be visible, that is sparse in one of its T-F regions. Although the algorithm proposed in this paper is similar to the 'successive source cancellation' method, our framework includes a deflation technique [10], making it a more flexible, generic approach to separation. Our approach operates without the mixing and estimation constraints of the 'successive source cancellation' method.

We demonstrate that a sequential SCA approach has a separation advantage over a SCA simultaneous approach, by comparing our algorithm to the Time Frequency Ratio of Mixtures (TIFROM) algorithm [1]. The analysis compares the algorithms across stationary and time-varying mixtures containing three speech sources.

A. Problem Formulation

The BSS problem can be formulated as follows: The vector of sensor signals $X(n) = [x_1(n) \dots x_M(n)]^T$ are observations of the vector of signals $S(n) = [s_1(n) \dots s_P(n)]^T$ linearly mixed according to the system $A = \begin{pmatrix} a_{11} & \dots & a_{1P} \\ a_{M1} & \dots & a_{MP} \end{pmatrix}$:

$$\begin{aligned} X(n) &= A \cdot S(n) \\ \begin{pmatrix} x_1(n) \\ \vdots \\ x_M(n) \end{pmatrix} &= \begin{pmatrix} a_{11} \cdot s_1(n) + \dots + a_{1P} \cdot s_P(n) \\ \vdots \\ a_{M1} \cdot s_1(n) + \dots + a_{MP} \cdot s_P(n) \end{pmatrix} \end{aligned} \quad (1)$$

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 ($P=M$). Given only mixed observations $X(n)$, a $P \times M$ separation matrix W (estimating A^{-1}) must be computed and then multiplied by $X(n)$ in order to obtain a scaled permutation of the original signals $c \cdot S(n)$.

SCA approaches generally exploit the sparse representations of speech signals in the time-frequency domain [3]. The Short Time Fourier Transform (STFT) of the mixed observations $X(n)$ are computed, where $X(m, k)$ represents the T-F window of the mixed observations, centered on short

time window m and frequency k . The ratios $\xi(m, k) = [\xi_2(m, k) \dots \xi_M(m, k)]$ are computed between the corresponding T-F windows of the 1st and j th mixed observation for $j = 2$ to M :

$$\begin{aligned}\xi_j(m, k) &= \frac{x_1(m, k)}{x_j(m, k)} \\ &= \frac{a_{11}s_1(m, k) + \dots + a_{1P}s_P(m, k)}{a_{j1}s_1(m, k) + \dots + a_{jP}s_P(m, k)}\end{aligned}\quad (2)$$

Eqn (2) shows that when only a single source $s_i(n)$ (where $i \in \{1 \dots P\}$) is present in a T-F window (m, k) , $C_{ie} = [1 \ \xi(m, k)]^T$ will correspond to the source's mixing column. We define the signal $s_i(n)$ as being visible within this T-F window.

II. SIMULTANEOUS SCA APPROACH WITH TIFROM

TIFROM [1] employs a simple approach to simultaneously separate speech signals from the mixture by estimating the mixing column associated with each of the signals, under the following conditions:

- 1) For each signal $s_i(n)$, there exists a set of adjacent T-F windows where either $s_i(n)$ occurs alone or has dominance over all other signals. This will be referred to as TIFROM's visibility assumption.
- 2) Signals should be time-varying across a set of adjacent T-F windows (m, k) .

Under these two conditions, when only one signal is present (or dominant) across adjacent T-F windows, the ratios $\xi(m, k)$ are approximately constant. However, if there are two or more signals in the set of T-F windows, $\xi(m, k)$ will vary across these windows.

As a consequence of this property, TIFROM uses the variance across a series of T-F windows to estimate each mixing column. Within a series (Y_u, k) of time-adjacent windows of the ratios $\xi(m, k)$, the mean $me(Y_u, k)$ and variance $var(Y_u, k)$ are computed. All series are then searched for $var(Y_u, k)_{min}$. The first mixing ratios estimate is chosen as $R_1 = me(Y_u, k)_1$ corresponding to the series with the minimum variance $var(Y_u, k)_{min}$. For all additional mixing ratios estimates $q = \{2 \dots P\}$, $R_q = me(Y_u, k)_q$ corresponds to the $var(Y_u, k)_{min}$ from the set of series $Q \in [me(Y_u, k) - R_y] > T$, where $y = \{1 \dots q-1\}$. T is a priori threshold set to determine the minimum difference between the mixing column ratios, ensuring that the same signal is not estimated multiple times. Finally, the inverse function of the mixing columns estimates are computed in order to estimate the separation matrix W .

III. SEQUENTIAL SCA APPROACH

A. Source Cancellation Approach

The sequential algorithm proposed in this paper uses the TIFROM approach to estimate mixing ratios. However, instead of estimating each signal's mixing ratios from the mixture concurrently, only the mixing ratios $R_1 = [r_{12} \dots r_{1M}]$ of the most visible signal $s_i(n)$ is estimated. The source cancellation

method (as used in [8]) is then employed to eliminate the contribution of this signal $s_i(n)$ from the mixture $X(n)$ using R_1 as follows:

$$\begin{aligned}x_{mod,j}(n) &= x_1(n) - r_{1j} \cdot x_j(n) \\ &= (a_{11} \cdot s_1 + \dots + a_{1P} \cdot s_P) \\ &\quad - \frac{a_{1i}}{a_{ji}} \cdot (a_{j1} \cdot s_1 + \dots + a_{jP} \cdot s_P)\end{aligned}\quad (3)$$

where $j = 2$ to M and $i \in \{1 \dots P\}$. If more than one signal remains, the mixing ratios of the most visible signal in the T-F representation of $X_{mod}(n)$ are estimated and used in (3) to eliminate this signal. This process continues until only an estimate of the least visible signal $s_e(n)$ in the mixture remains.

The source cancellation approach is advantageous to simultaneous TIFROM for the separation of less visible signals. As TIFROM estimates the mixing ratios of all the signals from the original mixture, signals of poor visibility are inaccurately estimated. The source cancellation approach, however, estimates a signal by repeatedly cancelling the contribution of the most visible signal (best estimate) of the mixture. Therefore, less visible signals are not separated by their own weak estimate, but the more accurate mixing ratios estimates of the other signals. In addition, sequentially cancelling signals from the mixture improves separation, as the visibility (and hence estimation) of the remaining signals is likely to improve in the less dense mixture.

B. Deflation Technique

Once the least visible signal $s_e(n)$ is estimated, a deflation technique is employed to deflate (remove) this signal from the original mixture $X(n)$. In the source cancellation process, $s_e(n)$ is estimated up to an undetermined scaling factor of the original signal. Thus, $s_e(n)$ is removed from the mixture by:

$$X(n)_{v+1} = X(n)_v - a \cdot s_e(n) \quad (4)$$

where a is a scaling factor. The optimal solution of a is the Minimum Mean Squared Error (M.M.S.E) of $X(n)_{v+1}$. This involves computing the derivative $\frac{\delta E}{\delta a} = 0$ and then making a the subject of the expression:

$$\begin{aligned}C &= E\{X(n)_{v+1}^2\} \\ \frac{\delta C}{\delta a} &= -2 \cdot s_e(n) \cdot X(n)_v + 2 \cdot s_e(n)^2 \cdot a \\ a &= R_{xs} \cdot R_{ss}^{-1}\end{aligned}\quad (5)$$

where $E\{\cdot\}$ is the expected value, $R_{ss} = E\{s_e(n) \cdot s_e(n)^T\}$ and $R_{xs} = E\{X(n)_v \cdot s_e(n)^T\}$. The optimal value of a scales $s_e(n)$ in order to remove it from $X(n)_v$.

Incorporating a deflation technique into the SeqTIF algorithm removes the constraint that TIFROM or the 'source cancellation method' in [8] places upon the mixing system. TIFROM and [8] require restrictions to be placed on the distance between mixing ratio estimates (threshold T detailed in Section II) to ensure different signals are estimated from the mixture. This is unnecessary when a deflation approach

is employed, as each signal is permanently removed from the mixture once its estimate has been found. Furthermore, after a signal is estimated, deflating it from the mixture enhances the estimation of the remaining signals. SeqTIF deflates the least visible signal from the mixture to ensure that this weaker estimate does not contribute to the estimation of the remaining visible signals. In [8], however, weaker estimates will contribute to the estimation of all signals.

C. Outline of the Sequential Algorithm

We outline the steps of the proposed sequential SCA (SeqTIF) algorithm:

- 1) Initially, the number of signals in the mixture is $J = P$. The mixed observations are processed as series (Y_u, k) of T-F windows, as detailed in Section II. The mean $me(Y_u, k)$ and variance $var(Y_u, k)$ of these series are computed.
- 2) The mixing ratios of the most visible signal $s_i(n)$ in the mixture are estimated as $R_1 = me(Y_u, k)$ from the series with minimum variance $var(Y_u, k)_{min}$.
- 3) The contribution of this signal $s_i(n)$ is eliminated from the mixture using the elements of R_1 in (3).
- 4) Steps 1, 2 and 3 are repeatedly applied to the new mixture $J-1$ times, so that all but one signal is cancelled from the mixture.
- 5) The remaining signal is then removed from the mixture using a deflation technique. The optimal scaling factor is computed from (5) and then used to eliminate the signal from the mixture in (4). The number of signals in the mixture is decreased by one i.e. $J = J - 1$.
- 6) The process from steps 1-5 is repeatedly applied on the newly deflated mixture $P-1$ times, so that all the signals in the mixture are retrieved.

IV. RESULTS

To verify that the sequential SeqTIF approach has a separation advantage over the simultaneous TIFROM approach, we applied these two algorithms to a data set of ten different mixtures that consisted of three speech signals that were 2.5s in length and sampled at 8000 Hz. The simulation was repeated on the data set four times, with a different mixing system being applied on each occasion. Hence, there were 40 mixtures used in the analysis. All mixtures were passed to the algorithms in analysis blocks sized:

$$blocksize = overlap * framesize * (fps + 1) + overlap * framesize * (seriesnum - 1) \quad (6)$$

where the $framesize = 20ms$, $overlap = 50\%$ of a frame, number of adjacent frames per series (fps) = {6, 8} and number of series in each block of data ($seriesnum$) = {1...180}. The threshold T for TIFROM was set to ensure the minimum distance between estimated mixing ratios was 15% of each ratio's value. This did not restrict TIFROM performance in the experiment, however, as the true mixing ratios of each mixing system were well spaced.

Fig. 1. SeqTIF's and TIFROM's average SNR across across 10 stationary speech mixtures. The analysis was conducted for $seriesnum=\{1...180\}$ and $fps = 6$ and 8 .

The criteria used to measure the separation performance is the Signal to Noise Ratio (SNR). It is calculated between the estimated signal $s_e(n)$ and the corresponding original signal $s_i(n)$:

$$SNR = 10 \cdot \log_{10} \frac{\sigma^2}{E\{(s_i(n) - s_e(n))^2\}} \text{ dB} \quad (7)$$

where σ^2 is the variance of $s_i(n)$. When calculating the SNR , both $s_e(n)$ and $s_i(n)$ are normalized to a variance of 1, in order to avoid the scaling problem of $s_e(n)$ that is inherent to BSS [10]. Thus, the SNR is simplified to $-10 \cdot \log_{10} E\{(s_i(n) - s_e(n))^2\}$, such that a higher SNR corresponds to a smaller estimation error of $s_e(n)$, hence better separation performance. The average SNR between the original mixtures $X(n)$ and original signals $s_i(n)$ for this analysis was 1.2 dB.

Fig. 1 shows the average SNR achieved in separating the ten mixtures of speech signals in the experiment. It clearly illustrates that SeqTIF's separation performance is superior to TIFROM across all $seriesnum$ and fps . SeqTIF's separation advantage over TIFROM is a consequence of its estimation requirements being less restrictive. This is particularly evident for a smaller number of series $seriesnum < 100$, where SeqTIF's separation advantage over TIFROM increases, reaching a maximum SNR advantage of 16dB (at $seriesnum = 5$ for $fps = 8$). TIFROM's estimation performance is poor for a smaller number of series, as visible signals are scarce and hence it is difficult to simultaneously estimate all of the signals from the mixture. SeqTIF, however, demonstrates good data efficiency with a SNR of at least 19dB across a smaller

Fig. 2. SeqTIF's and TIFROM's average SNR across across 10 time-varying mixtures of speech. The analysis was conducted for $seriesnum=\{1...180\}$ and $fps = 6$ and 8 . The dotted vertical line on each plot represents an analysis block of 200ms, the mixing system was varied every 200ms.

number of series. Such a sequential approach is better suited to estimation with a smaller number of series, as it only requires an accurate estimate of a single signal from the mixture during each cancellation stage in (3). In addition, as signal estimates are cancelled, there is an increase in the visibility of signals in the remaining mixture.

A second experiment was conducted to demonstrate that the superior data efficiency of SeqTIF enables it to track time-varying mixtures better than TIFROM. Both algorithms were applied to the same ten speech signal trios from the previous experiment, with the mixtures being extended to 5 s in length. The mixing system applied to the speech signal trios was varied every 200ms in this experiment.

Fig. 2 shows the average SNR achieved in separating the ten time-varying mixtures of speech signals in this second experiment. The results indicate that SeqTIF has a significant separation advantage over TIFROM for smaller numbers of series, with a SNR advantage of up to 15dB for $seriesnum < 15$. This is because a sequential approach to separation has a superior data efficiency to a simultaneous approach, as shown in the first experiment. When the number of series increases, however, SeqTIF's separation performance advantage over TIFROM declines. This is because when analysis blocks increase in size from 200ms (the vertical dotted lines correspond to a 200ms block in Fig. 2) they span at least one change in the mixing system. As a consequence, SeqTIF is unable to accurately estimate signals, as the variation in the mixing system makes it impossible to cancel signals from the mixture using (3) or (4). TIFROM estimation is

also degraded for analysis blocks that possess mixing matrix variation, hence the performance of SeqTIF and TIFROM converge for $seriesnum > 50$.

V. CONCLUSION

We proposed a sequential approach to Sparse Component Analysis (SeqTIF) by combining the estimation process of TIFROM [1], with a source cancellation method [8] and deflation technique [10]. The separation performance of SeqTIF was shown to be superior to the simultaneous TIFROM approach, which used the same estimation process. In particular, SeqTIF had a significant separation advantage over TIFROM (up to 16dB) for smaller analysis blocks $seriesnum < 100$. The separation advantage of SeqTIF can be attributed to the less restrictive assumptions it places upon signals in the mixture. Whilst a simultaneous SCA approach, such as TIFROM, requires all signals to be sparsely represented (visible) in the mixture, SeqTIF focuses only on a single signal in the current mixture, and hence, only requires one signal to be visible. SeqTIF's superior data efficiency to TIFROM was again shown, as both algorithms tracked a time varying mixture. SeqTIF's SNR advantage over TIFROM reached 15dB for analysis blocks with constant mixing systems. SeqTIF's improved data efficiency, combined with its freedom from heuristics and mixing system constraints, indicates that sequential SCA has potential for application in realistic acoustic scenes where mixing systems vary.

REFERENCES

- [1] F.Abrard, Y.Deville, and P.White, "From blind source separation to blind source cancellation in the undetermined case: A new approach based on time-frequency analysis," in *Proc.3rd International Conference on Independent Component Analysis and Blind Source Separation (ICA2001)*, San Diego, December 2001, pp. 734-739.
- [2] A.Hyvarinen, J.Karhunen, and E.Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.
- [3] S.Choi, A.Cichoki, H.Park, and S.Lee, "Blind source separation and independent component analysis: A review," *Neural Information Processing*, vol. 6, no. 1, pp. 1-57, 2005.
- [4] A.Jourjine, S.Rickard, and O.Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proc.IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP2000)*, vol. 5, 2000, pp. 2985-2988.
- [5] D.Smith, J.Lukasiak, and I.Burnett, "Two channel, block-adaptive audio separation using the cross correlation of time-frequency information," in *Proc.5th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, vol. LNCS 3195. Springer, September 2004, pp. 889-897.
- [6] O.Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans on Signal Processing*, vol. 52, no. 7, pp. 1830-1847, 2004.
- [7] A.Blin, S.Araki, and S.Makino, "A sparseness-mixing matrix estimation (smme) solving the undetermined bss for convolutive mixtures," in *Proc.IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP2004)*, vol. 4, May, 2004, pp. 85-88.
- [8] F.Abrard and Y.Deville, "A time-frequency blind signal separation method applicable to undetermined mixtures," *Signal Processing*, In Press, online April 2005.
- [9] R.Balan, J.Rosca, and S.Rickard, "Scalable non-square blind source separation in the presence of noise," in *Proc.IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2003)*, vol. 5, 2003, pp. 293-296.
- [10] A.Cichoki and S.Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, 2002.