SOURCE DETECTION USING REPETITIVE STRUCTURE

R. Mitchell Parry and Irfan Essa

Georgia Institute of Technology
College of Computing / GVU Center
Atlanta, Georgia 30332-0760, USA

ABSTRACT

Blind source separation algorithms typically require that the number of sources are known in advance. However, it is often the case that the number of sources change over time and that the total number is not known. Existing source separation techniques require source number estimation methods to determine how many sources are active within the mixture signals. These methods typically operate on the covariance matrix of mixture recordings and require fewer active sources than mixtures. When sources do not overlap in the time-frequency domain, more sources than mixtures may be detected and then separated. However, separating more sources than mixtures when sources overlap in time and frequency poses a particularly difficult problem. This paper addresses the issue of source detection when more sources than sensors overlap in time and frequency. We show that repetitive structure in the form of time-time correlation matrices can reveal when each source is active.

1. INTRODUCTION

Blind source separation is the problem of separating a number of source signals (e.g. musical instruments) from a set of mixture signals (e.g. a song). For example, stereo music provides any number of sources within a two channel recording. Often, the case that the number of instruments playing at once exceeds the number of mixtures. We represent this scenario via an instantaneous linear mixing model:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{1}$$

where $\mathbf{x} = [x_1(t), \cdots, x_M(t)]^T$ is a time varying vector representing the mixtures, $x_i(t)$, $\mathbf{s} = [s_1(t), \cdots, s_N(t)]^T$ represents the sources, $s_i(t)$, and $\mathbf{A}$ is the $M \times N$ real mixing matrix. The ith column of $\mathbf{A}$ defines the mixing parameters for source s_i (*i.e.,* its spatial position).

A class of algorithms that addresses blind source separation is independent component analysis [1]. However, assuming that the sources are independent is not enough; additional source structure is required. The classic information theoretic example of source structure is a non-Gaussian probability distribution [1]. However, temporal structure can enhance the separation of time-varying signals. For example, sources often have smoothly changing variance (*i.e.,* non-stationarity) or are correlated at time-lags (*i.e.,* autocorrelated) [1]. We are investigating a new type of temporal structure that has not been addressed in the literature: repetitive structure.

Source signals that change over time and are highly correlated at different points in time exhibit repetitive structure. For example, music contains repeated notes, melodies, and sections. While musical repetition is carefully constructed and often periodic, other source signals repeat in less predictable ways. For example, the sound of a bell ringing or a door closing depends on external factors, but is highly correlated every time it sounds. We believe that many interesting source signals contain repetitive structure that can be leveraged in various signal processing applications.

2. RELATED WORK

Although the ultimate goal is to separate source signals from the mixture signals, the vast majority of blind source separation algorithms require that the number of sources is known in advance and that $N \leq M$. The problem is more complicated when the number of active sources changes over time. Therefore, most source separation algorithms require source number estimation, or more generally source detection, before separation can begin. Generally, source number estimation is performed on a correlation matrix computed on the mixtures [2, 3, 4]. Estimating the number of sources from a correlation matrix requires that the number of sources is strictly less than the number of mixtures. Other techniques estimate the source signals and the source number concurrently [5, 6]. However, all of these techniques require $N \leq M$, whereas we consider the case of more sources than mixtures ($N > M$).

When the time-frequency representations of the sources do not overlap, more sources than mixtures can be detected and separated [7, 8]. The correlation matrix at a time-frequency point containing only one active source reveals the source's spatial position in the mixture. The number of unique positions indicates the number of sources. However, if sources overlap in time and frequency it becomes difficult to determine which sources are active even if the source positions are known [9]. Our approach uses repetitive structure and known

source positions to estimated which sources are active at each point in time.

Recently, repetitive structure has been used for segmentation [10], summarization [11], and compression [12] of audio signals. Foote visualizes repetitive structure in audio and video in a two-dimensional self-similarity matrix [13]. By comparing every pair of short audio frames, the self-similarity matrix captures multiple levels of repetitive structure. Repetitive structure has recently been used for blind source separation [14] when the number of sources, N, is not greater than the number of mixtures, M. Here, we show its utility for source detection when $N > M$.

3. REPRESENTATION

Temporal structure is often captured by correlation matrices. Local correlation matrices computed within short time frames capture the non-stationarity of the sources. Global covariance matrices computed at time-lags capture the autocorrelation of the sources. We combine parts of each approach to form correlation matrices between short time-frames at different time-lags. Analyzing different time frames within a signal leads to a time-time representation analogous to the self-similarity matrix mentioned above. The only difference is that a time-time representation contains a correlation *matrix* for every pair of frames instead of a scalar similarity value. The specific time-time representation that we use is derived from the pseudo Wigner distribution [14]:

$$\mathbf{D_{xx}}(t_1, t_2) = \int h(\tau)\mathbf{x}(t_1 + \frac{\tau}{2})\mathbf{x}^H(t_2 - \frac{\tau}{2})\, d\tau, \qquad (2)$$

where h is a localization window. We evaluate $\mathbf{D_{xx}}$ at equally spaced points in time, thereby correlating every windowed frame with every time-reversed windowed frame. When $t_1 = t_2$, the representation mimics the local correlation matrices that capture non-stationarity, except that one frame is time-reversed. The time-reversal enhances temporal precision at the cost of time-lag precision.

We relate the time-time representation of the mixtures to the time-time representation of the sources:

$$\mathbf{D_{xx}} = \mathbf{A}\mathbf{D_{ss}}\mathbf{A}^H \qquad (3)$$

where ^{H} is the Hermitian transpose. By applying a whitening matrix $\mathbf{W}$ we have

$$\begin{aligned} \mathbf{z}(t) &= \mathbf{WA}\mathbf{s}(t) && (4)\\ \mathbf{D_{zz}} &= \mathbf{WA}\mathbf{D_{ss}}\mathbf{A}^H\mathbf{W}^H && (5)\\ &= \mathbf{U}\mathbf{D_{ss}}\mathbf{U}^H && (6) \end{aligned}$$

where $\mathbf{U} = \mathbf{WA}$ is an $M \times N$ pseudounitary matrix (*i.e.,* $\mathbf{U}^{\#} = \mathbf{U}^H$ where $^{\#}$ is the Moore-Penrose pseudoinverse).

4. SOURCE DETECTION

The time-time representation of the sources, $\mathbf{D_{ss}}$, contains all the information required for source detection. If $\mathbf{D_{ss}}(t_1, t_2)_{ij}$ is nonzero, source i is active at t_1 and source j is active at t_2. However, because $N > M$, we cannot simply invert $\mathbf{U}$ to find $\mathbf{D_{ss}}$. Instead, we must isolate time pairs that reveal parts of $\mathbf{D_{ss}}$. For example, if source i is the only source active at t_1 and t_2, $\mathbf{D_{zz}}(t_1, t_2) = \mathbf{D_{ss}}(t_1, t_2)_{ii}\mathbf{u}_i\mathbf{u}_i^H$, where $\mathbf{u}_i$ is the ith column of $\mathbf{U}$ representing the whitened mixing parameters of source i. In this case, $\mathbf{u}_i$ can be estimated up to a scale factor by the eigenvector of $\mathbf{D_{zz}}(t_1, t_2)$, thus detecting source i at t_1 and t_2. This is a special case because $\mathbf{D_{zz}}(t_1, t_2)$ happens to be rank-one and symmetric. In the general case, $\mathbf{D_{zz}}(t_1, t_2)$ is a linear combination of the product of all pairs of whitened mixing parameters:

$$\mathbf{D_{zz}}(t_1, t_2) = \sum_{ij} \mathbf{D_{ss}}(t_1, t_2)_{ij}\mathbf{u}_i\mathbf{u}_j^H \qquad (7)$$

Although reconstructing $\mathbf{D_{ss}}$ from $\mathbf{D_{zz}}$ is generally not possible, we can hope to estimate one element of $\mathbf{D_{ss}}(t_1, t_2)$, revealing one source active at t_1 and one source active at t_2. If one element of $\mathbf{D_{ss}}(t_1, t_2)$ dominates the rest (*i.e.,* $|\mathbf{D_{ss}}(t_1, t_2)_{qr}| \gg |\mathbf{D_{ss}}(t_1, t_2)_{ij}|\ \forall\ i \neq q$ or $j \neq r$), we can approximate $\mathbf{D_{zz}}(t_1, t_2)$:

$$\mathbf{D_{zz}}(t_1, t_2) \approx \mathbf{D_{ss}}(t_1, t_2)_{qr}\mathbf{u}_q\mathbf{u}_r^H \qquad (8)$$

Using singular value decomposition, we can compute a rank-one approximation of $\mathbf{D_{zz}}(t_1, t_2)$:

$$\mathbf{D_{zz}}(t_1, t_2) \approx d\mathbf{v}_1\mathbf{v}_2^H \qquad (9)$$

where $\|\mathbf{v}_1\| = \|\mathbf{v}_2\| = 1$ and d is the largest singular value of $\mathbf{D_{zz}}(t_1, t_2)$. Source $\hat{k}$ is likely to be active at t_1, if its mixing parameters, $\mathbf{u}_{\hat{k}}$ are the closest to $\mathbf{v}_1$:

$$\hat{k} = \arg\max_k \frac{\mathbf{v}_1^H\mathbf{u}_k}{\|\mathbf{u}_k\|} \qquad (10)$$

We collect this evidence in a two-dimensional function c_n for each source n:

$$c_n(t_1, t_2) = \begin{cases} d, & n = \hat{k} \\ 0, & otherwise \end{cases} \qquad (11)$$

Because $\mathbf{D_{zz}}(t_1, t_2) = \mathbf{D_{zz}}(t_2, t_1)^H$, we need only consider the upper or lower triangular elements of $\mathbf{D_{zz}}$ to completely fill in c_n. The function $c_n(t_1, t_2)$ contains the evidence that source n is active at time t_1 given $\mathbf{D_{zz}}(t_1, t_2)$. To aggregate this information we construct the activation function $g_n(t)$:

$$g_n(t) = \int c_n(t, \tau)\, d\tau \qquad (12)$$

Applying a threshold classifier to a smoothed version of this function would then provide the source activations.

We explore the following algorithm for source detection:

Time	0	1	2	3	4	5	6	7
Source 1	x			x		x		x
Source 2			x	x			x	x
Source 3					x	x	x	x

Fig. 1. Activation sequence of sources.

1. Compute the whitened time-time representation of the mixture signals from Equation 2.
2. For each t_1 and t_2
   (a) Compute the rank-one approximation according to Equation 9.
   (b) Classify the resulting singular vectors according to Equation 10 to find the source k_1 associated with t_1.
   (c) Assign $c_n(t_1, t_2)$ to the largest singular value of $\mathbf{D_{zz}}(t_1, t_2)$, d.
3. Construct the activation function, g_n according to Equation 12.

5. RESULTS

To demonstrate our algorithm, we analyze a two channel mixture of three sources with overlapping frequency content. The sources are drawn from a Gaussian distribution with zero mean and unit variance and then filtered by a conjugate pair filter. Sources 1, 2, and 3 were filtered using normalized center frequencies of 0.20, 0.25, and 0.30, respectively. Figure 2 shows the frequency content of each of the sources. The distributions show considerable overlap in frequency. We construct the repetitive structure by activating each source in a different pattern, shown in Figure 1. Each activation from the same source is randomly generated using the same distribution and filter. Thus, the repetitions are not identical, only highly correlated.

We generate the mixtures, $\mathbf{x}(t)$, via Equation 1 using the following mixing matrix:

$$A = \begin{bmatrix} 0.4403 & 0.5499 & 0.9068 \\ -0.8978 & 0.8352 & 0.4215 \end{bmatrix} \quad (13)$$

We compute the time-time representation of the whitened mixtures using Equation 2 and fill in the collection function, c_n for each source. Figure 3 shows the collection function for source 1. Each row, t_1, contains the evidence for source 1 being active at time index t_1. The darker squares indicate that $\mathbf{D}_{zz}(t_1, t_2)$ provides more evidence for source 1 activity when source 1 is present at both t_1 and t_2. Figure 4 shows the activation function for each of the sources. As expected, when one source is active, only the correct source receives evidence of activation. For example, only source 3 is active from 3-4 seconds in Figure 4. When two sources are active,

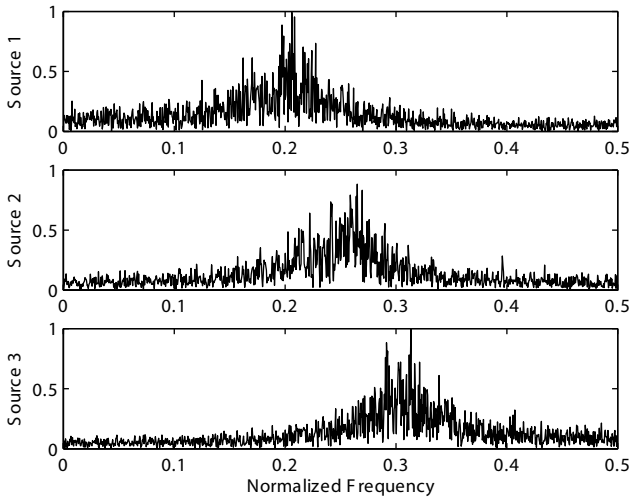

Fig. 2. Normalized frequency of the sources.

they sometimes combine to approximate the remaining inactive source. For example some activation energy is shown for source 1 between 5-6 seconds, source 2 between 4-5 seconds, and source 3 between 2-3 seconds. When all three sources are active, the activation function is high for all three sources.

6. CONCLUSION

We investigate the use of repetitive structure for source detection. The time-time representation captures information about source activation, and provides a means to retrieve it blindly from the mixtures. This method of analysis applies to the situation of more sources than mixtures.

7. REFERENCES

[1] A. Hyvärinen, *Independent Component Analysis*, New York: Wiley, 2001.

[2] S. Aouada, A. M. Zoubir, and C. M. S. See, "A comparative study on source number detection," in *Proceedings of the International Symposium on Signal Processing and its Applications*, Paris, July 2003, vol. 1, pp. 173–176.

[3] H.-T. Wu, J.-F. Yang, and F.-K. Chen, "Source number estimators using transformed Gerschgorin radii," *IEEE Transactions on Signal Processing*, vol. 43, no. 6, pp. 1325–1333, 1995.

[4] K. Yamamoto, F. Asano, W. F. G. van Rooijen, E. Y. L. Ling, T. Yamada, and N. Kitawaki, "Estimation of the number of sound sources using support vector machines and its application to sound source separation," in *Proceedings of the IEEE International Conference on*

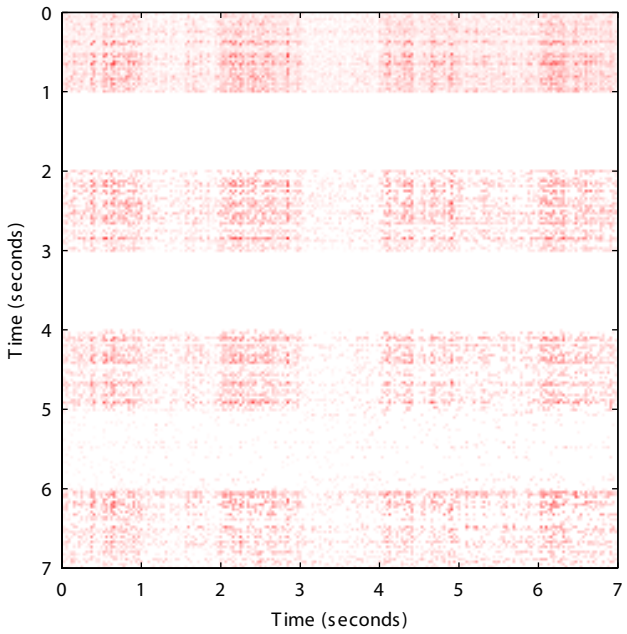

Fig. 3. Collection function for source 1.

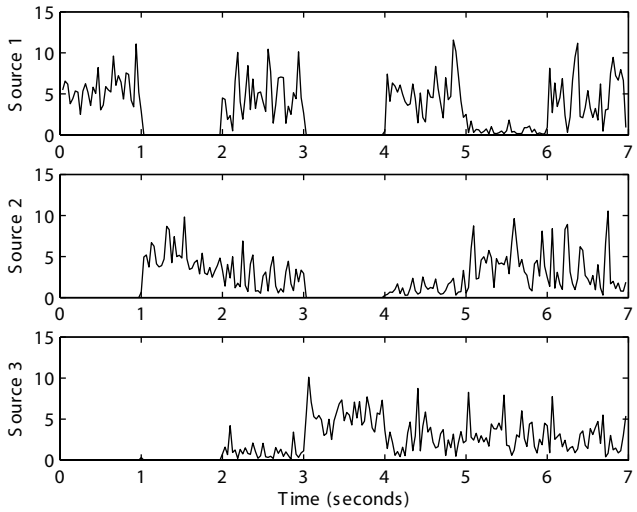

Fig. 4. Activation function for each source.

Acoustics, Speech, and Signal Processing, 2003, vol. 5, pp. 485–488.

[5] S.-T. Lou and X.-D. Zhang, "Blind source separation for changing source number: A neural network approach with a variable structure," in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation*, San Diego, CA, December 2001, pp. 384–389.

[6] J.-M. Ye, X.-L. Zhu, and X.-D. Zhang, "Adaptive blind separation with an unknown number of sources," *Neural Computation*, vol. 16, pp. 1641–1660, 2004.

[7] L.-T. Nguyen, A. Belouchrani, K. Abed-Meraim, and B. Boashash, "Separating more sources than sensors using time-frequency distributions," in *Proceedings of the International Symposium on Signal Processing and its Applications*, Kuala Lumpur, Malaysia, August 2001, pp. 583–586.

[8] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[9] A. S. Master, "Bayesian two source modeling for separation of N sources from stereo signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 2004, vol. 4, pp. 281–284.

[10] J. Foote and M. Cooper, "Media segmentation using self-similarity decomposition," in *Proceedings of SPIE*, 2003.

[11] M. Cooper and J. Foote, "Summarizing popular music via structural analysis," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2003.

[12] T. Jehan, "Perceptual segment clustering for music description and time-axis redundancy cancellation," in *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, October 2004, pp. 124–127.

[13] J. Foote, "Visualizing music and audio using self-similarity," in *Proceedings of ACM Multimedia*, Orlando, FL, November 1999, pp. 77–80.

[14] R. M. Parry and I. Essa, "Blind source separation using repetitive structure," in *Proceedings of International Conference on Digital Audio Effects*, Madrid, Spain, September 2005, pp. 143–148.