BLIND SOURCE SEPARATION USING REPETITIVE STRUCTURE

R. Mitchell Parry and Irfan Essa

College of Computing / GVU Center
Georgia Institute of Technology, Atlanta, GA USA
[parry|irfan]@cc.gatech.edu

ABSTRACT

Blind source separation algorithms typically involve decorrelating time-aligned mixture signals. The usual assumption is that all sources are active at all times. However, if this is not the case, we show that the unique pattern of source activity/inactivity helps separation. Music is the most obvious example of sources exhibiting repetitive structure because it is carefully constructed. We present a novel source separation algorithm based on spatial time-time distributions that capture the repetitive structure in audio. Our method outperforms time-frequency source separation when source spectra are highly overlapping.

1. INTRODUCTION

Source separation techniques attempt to decompose a set of time-aligned mixture signals (*e.g.,* a song) into their constituent source signals (*e.g.,* instrument tracks). The usual assumption is that all sources are active at all times. However, many sources exhibit repetitive structure in the form of activation patterns. We utilize this structure in order to separate sources.

As a motivating example, consider a repeating source such as a bell tower or public address system that obscures the separation of other local signals such as people talking. Because the bell tower chimed an hour ago among a different mix of sounds, we expect to better separate it in the current instance, *e.g.,* either by producing a cleaner recording of the bells or by removing their contribution to the mixture. A bell tower sounds very similar every time it chimes whether or not the exact melody is duplicated. We identify when a source repeats itself and use this to separate it from a mixture. This is in contrast to blind source separation (BSS) techniques that utilize correlations between mixtures computed globally on an entire signal or locally at different time or time-frequency points. These techniques decorrelate time-aligned mixture signals, whereas we decorrelate between mixtures at different points in time.

In general, BSS attempts to separate N source signals from M mixtures by estimating the sources and mixing matrix according to the following:

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

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 $\mathbf{A}$ is the spatial *position* of s_i in the mixture, *i.e.,* its contribution to each mixture channel.

Independent component analysis (ICA) is a class of algorithms for BSS that assume sources are statistically independent. Earlier techniques additionally assume that sources are stationary (*i.e.,* do not change over time) [1, 2, 3, 4]. These algorithms operate on a single correlation matrix computed on the entire multichannel mixture signal. Because source signals are assumed to be statistically independent, the correlation matrix computed on the source signals is diagonal. Therefore, diagonalizing these second-order mixture correlations via a whitening transform is an important first step for source separation. Because independence implies higher-order decorrelation, these techniques use additional criteria such as information maximization, minimum mutual information, and higher-order decorrelation to separate stationary sources. However, many real source signals are not stationary, and this non-stationarity can be leveraged for source separation.

Non-stationary signals have statistical properties that change over time, *e.g.,* signal or spectral energy. Correlations between the time-varying energy of signals are 4th-order relationships that are explicitly minimized in the stationary case [3]. For non-stationary signals, changes in energy affect the local 2nd-order correlations. Instead of diagonalizing a single global correlation matrix and optimizing an additional criterion, sources can be separated by joint diagonalization of multiple correlation matrices computed within different time blocks [5, 6, 7].

The energy of a non-stationary signal may also change within a frequency band. Techniques that isolate these changes apply to time-frequency distributions [8, 9, 10]. A correlation matrix is computed for each time-frequency point. Points that correspond to single source contributions are isolated and jointly diagonalized to separate sources.

The previous techniques do not consider the repetitive structure of audio. Non-stationary techniques only benefit from source inactivity if it uncovers time-frequency points containing only one source. Our method maximizes the utility of repetitive structure by pinpointing unique source repetitions and isolating their spatial positions.

Repetitive structure informs other tasks including segmentation [11], summarization [12], and compression [13]. Foote visualizes repetitive structure in audio and video in a two-dimensional self-similarity matrix [14]. This representation is a time-time energy distribution. Just as correlation matrices at time-frequency points separate sources with unique spectral shape, we show that time-time correlations separate signals exhibiting unique repetitive structure.

2. SPATIAL TIME-FREQUENCY SEPARATION

Time-frequency distributions (TFD) estimate the energy of a signal at time-frequency points. The spectrogram is often used to estimate the energy content in a single signal. Other distributions enable the estimation of the energy shared between two signals,

e.g., the pseudo Wigner distribution [15]:

$$D_{x_1x_2}(t,f) = \int h(\tau)x_1(t+\frac{\tau}{2})x_2^*(t-\frac{\tau}{2})e^{-j2\pi f\tau}d\tau \quad (2)$$

where h is a time window and superscript * is the complex conjugate. Some blind source separation techniques leverage the unique TFDs of source signals. Belouchrani and Amin construct a TFD for every pair of mixture signals. These are viewed as an $M \times M$ spatial correlation matrix for every time-frequency point [8]:

$$[\mathbf{D_{xx}}(t,f)]_{ij} = D_{x_ix_j}(t,f) \quad (3)$$

The correlation matrices of the mixtures are related to those of the sources according to the following equation [8]:

$$\mathbf{D_{xx}}(t,f) = \mathbf{A}\mathbf{D_{ss}}(t,f)\mathbf{A}^H, \quad (4)$$

where $\mathbf{D_{xx}}$ is the $M \times M$ mixture correlation matrix, $\mathbf{D_{ss}}$ is the $N \times N$ source correlation matrix, and superscript ^{H} indicates the Hermitian transpose. The whitened correlation matrices,

$$\mathbf{D_{zz}}(t,f) = \mathbf{W}\mathbf{D_{xx}}(t,f)\mathbf{W}^H, \quad (5)$$

can be constructed from the mixtures using the whitening transform $\mathbf{W}$, or from the sources by applying a unitary transform $\mathbf{U}$:

$$\mathbf{D_{zz}}(t,f) = \mathbf{U}\mathbf{D_{ss}}(t,f)\mathbf{U}^H. \quad (6)$$

If only one source is active at a time-frequency point, $\mathbf{D_{ss}}(t,f)$ is quasi-diagonal [16]. These single-source time-frequency points are called autoterms. Matrix $\mathbf{U}$ can be estimated as the unitary matrix that jointly diagonalizes $\mathbf{D_{zz}}$ at all time-frequency autoterms. This requires that each source generates at least one autoterm. Belouchrani and Amin [8] estimate $\mathbf{A}$ as

$$\hat{\mathbf{A}} = \mathbf{W}^{\#}\mathbf{U}, \quad (7)$$

where superscript $^{\#}$ indicates the pseudoinverse.

Autoterm candidates are estimated using the energy and rank-oneness at each time-frequency point [8, 9, 10]. When the time-frequency distributions of sources do not overlap, more sources than mixtures can be extracted [16, 17]. However, the performance of time-frequency separation degrades as source distributions become more overlapping. In the extreme case, there are no time-frequency autoterms and therefore $\mathbf{A}$ cannot be estimated. Our approach leverages a source's repetitive structure in order to overcome this shortcoming.

3. REPETITIVE STRUCTURE

Many audio signals exhibit structure in the form of repetition. Music is the most obvious example because the structure is carefully constructed. Different combinations of instruments play at different times and the notes they play are repeated over the course of a song. Repetitive structure also exists in other audio signals such as speech and natural recordings. Words, syllables, and phonemes are repeated in a conversation. The sounds of keyboards, telephones, and printers permeate an office building. Because each sound repeats in a different pattern and emanates from the same physical location, we expect to more easily separate or cancel it from a recording.

Foote's self-similarity matrix is an example of a time-time representation that operates on a single signal [14]. The original audio is partitioned into short audio frames ($\approx$ 50 milliseconds), features are computed on these frames, and every pair of frames is compared via a similarity metric. Here we use the magnitude of the fast Fourier transform for features, and the cosine of the angle between them for similarity. This produces a matrix of comparisons that represents the structure and repetition within the audio. In Figure 1 self-similar segments appear as white (*i.e.*, similar) squares along the main diagonal of the matrix. Repetitions appear as white rectangles off the main diagonal. In this case, the first verse (25–55 seconds) is very similar to the second verse (85–125 seconds) indicated by the large off-diagonal white rectangles centered at (40,105) and (105,40). Each verse is followed by a chorus (55–75 seconds and 125–145 seconds) with off-diagonal repetition squares centered at (65,135) and (135,65).

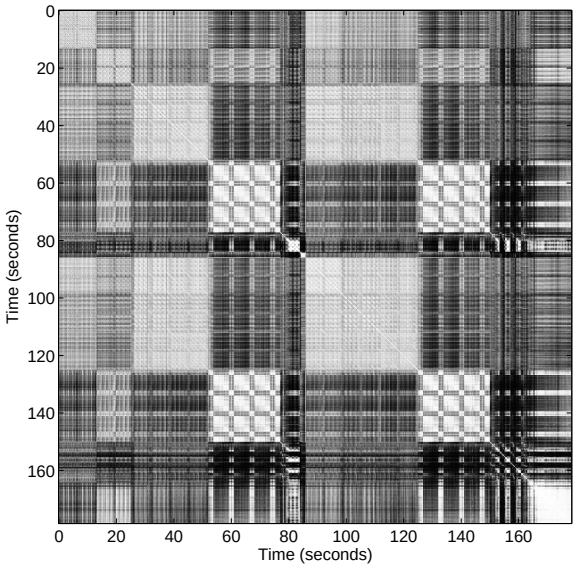


Figure 1: *Self-similarity matrix for "March of the Pigs" by Nine Inch Nails.*

4. SPATIAL TIME-TIME SEPARATION

Using the repetition in audio, we propose a novel approach to source separation: spatial time-time distribution (TTD) separation. Following the same general procedure as the spatial time-frequency separation described above, we identify time-time autoterms and estimate the mixing matrix via joint diagonalization of autoterm spatial correlation matrices.

We construct our time-time distribution by manipulating the pseudo Wigner distribution to be a function of two points in time:

$$D'_{x_1x_2}(t_1,t_2,f) = \int h(\tau)x_1(t_1+\frac{\tau}{2})x_2^*(t_2-\frac{\tau}{2})e^{-j2\pi f\tau}d\tau \quad (8)$$

To mimic the self-similarity matrix we remove the dependence on frequency:

$$S_{x_1x_2}(t_1,t_2) = D'_{x_1x_2}(t_1,t_2,0) \quad (9)$$

We focus on the application of time-time distributions, and leave the potential of time-time-frequency distributions for future work.

In the self-similarity example above, we compare the frequency components between audio frames. Here, time-time distributions compare windowed frames in the time domain, the second of which is reversed. This defines a self-similarity matrix (or time-time distribution) for every pair of mixtures. Alternatively, we represent them as $M \times M$ spatial correlation matrices:

$$[\mathbf{S_{xx}}(t_1, t_2)]_{ij} = S_{x_i x_j}(t_1, t_2) \tag{10}$$

Once again, we frame the source separation problem in terms of our time-time distribution:

$$\mathbf{S_{xx}}(t_1, t_2) = \mathbf{A}\mathbf{S_{ss}}(t_1, t_2)\mathbf{A}^H. \tag{11}$$

By applying the whitening matrix $\mathbf{W}$, we generate the whitened time-time correlation matrices:

$$\mathbf{S_{zz}}(t_1, t_2) = \mathbf{W}\mathbf{S_{xx}}(t_1, t_2)\mathbf{W}^H \tag{12}$$

As before, we estimate $\mathbf{A}$ using the unitary matrix $\mathbf{U}$ that satisfies the following:

$$\mathbf{S_{zz}}(t_1, t_2) = \mathbf{U}\mathbf{S_{ss}}(t_1, t_2)\mathbf{U}^H, \tag{13}$$

When two points in time contain only one common source, $\mathbf{S_{ss}}(t_1, t_2)$ is nearly diagonal. Thus, we estimate $\mathbf{U}$ as the unitary matrix that jointly diagonalizes $\mathbf{S_{zz}}$ at time-time autoterm points. Alternatively, because an autoterm's principal eigenvector best diagonalizes it [16], we may construct the columns of $\mathbf{U}$ as the unique principal eigenvectors of the autoterm correlation matrices. This enables the estimation of fewer source positions, if not all sources have unique repetitions.

Figure 2 shows the time-time distribution for the same song depicted in Figure 1. In all of the following figures, higher energy content is darker. One important difference between Figure 1 and 2 is that each frame of the self-similarity matrix is normalized to unit energy. Because the time-time distribution varies with the energy in the signal, the darkness of the image trails off at the end of the song. Otherwise, much of the same structure is visible in both representations.

We identify time-time autoterms in an analogous way to time-frequency autoterms. We estimate the energy at a time-time point as the trace of its spatial correlation matrix:

$$E(t_1, t_2) = |\mathrm{Trace}[\mathbf{S_{xx}}(t_1, t_2)]| \tag{14}$$

We estimate the rank-oneness of the matrix at a time-time point as

$$R(t_1, t_2) = \frac{\max(\lambda_i)}{\sum_i \lambda_i}, \tag{15}$$

where λ_i are the singular values of $\mathbf{S_{xx}}(t_1, t_2)$. Time-time correlation matrices above an energy and rank-oneness threshold correspond to time-time autoterms. Currently, we use the mean energy as the energy threshold and 0.95 as the rank-oneness threshold.

Figure 3 illustrates autoterm selection using time-time separation (left) and time-frequency separation (right). The sources were drawn from a zero mean and unit variance Gaussian distribution and filtered using a conjugate pair filter at different normalized center frequencies, f_i:

$$\begin{aligned} r_i(t) &= N(0,1) \\ z_i &= pe^{j2\pi f_i} \\ a_i &= [1, -2\Re\{z_i\}, z_i z_i^*] \\ s_i(t) &= x(t) - a_i(2)s_i(t-1) - a_i(3)s_i(t-2) \end{aligned} \tag{16}$$

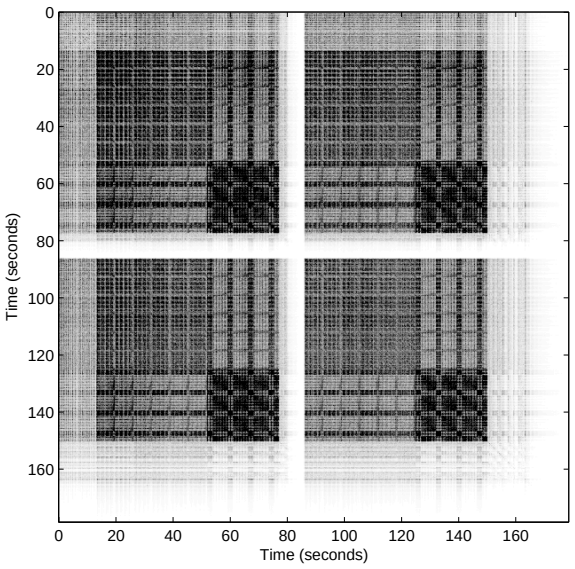
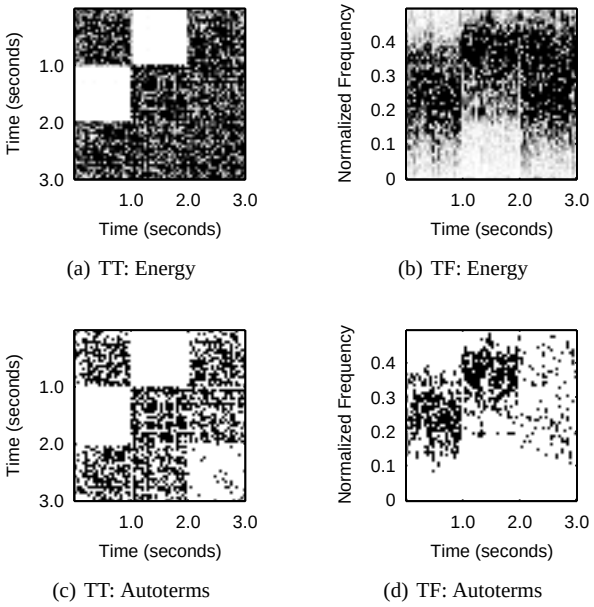


Figure 2: *Time-time distribution for "March of the Pigs" by Nine Inch Nails.*



(a) TT: Energy (b) TF: Energy

(c) TT: Autoterms (d) TF: Autoterms

Figure 3: *Autoterm selection: Time-time distribution (left) and time-frequency distribution (right)*

where superscript * indicates complex conjugate and $p = 0.85$. For this example, we use $f_1 = 0.25$ and $f_2 = 0.35$. In addition, each source exhibits a different activity pattern. Sources s_1 and s_2 activate in the patterns [on, off, on] and [off, on, on], respectively. This can be seen as a checkerboard pattern in Figures 3(a) and 3(c) and alternating frequency content in Figures 3(b) and 3(d). Figures 3(a) and 3(b) show the energy content in the two distributions. Figures 3(c) and 3(d) show the selected autoterm points. Notice that the high energy content appearing between 2 and 3 seconds (when the sources overlap) is not as likely to be chosen as an autoterm. Otherwise, high energy content is correctly identified as an autoterm.

5. RESULTS

We have described the application of spatial time-time separation in an analogous way to spatial time-frequency separation. Our algorithm provides an alternative to time-frequency separation when sources exhibit unique repetitions. When this is the case, our method outperforms time-frequency separation when sources have overlapping source spectra, and performs comparably well when they do not.

In our first experiment, we test how the similarity of sources affects their separability using time-time and time-frequency separation. We generate three random signals according to Equation 16 with $f_1 = 0.25 - \delta f$, $f_2 = 0.25$, and $f_3 = 0.25 + \delta f$. The unique activation sequences for s_1, s_2, and s_3 are [on, on, off], [on, off, on], and [off, on, on], respectively. The autoterms for $\delta f = 0.2$ are shown in Figure 4. The sections of Figure 4(a), annotated by dividing lines, indicate different source autoterms as labeled (*e.g.*, s_1 is the only source active the first two seconds). Relatively few autoterms are selected for time-time points within the same second because more than one source is active. Notice that each source's autoterms are also delineated in the time-frequency distribution of Figure 4(b).

We evaluate the quality of separation as the maximum interference-to-signal ratio (ISR) among all sources:

$$I = \max_p \frac{\sum_{q \neq p} |(\hat{\mathbf{A}}^{\#}\mathbf{A})_{pq}|^2}{|\hat{\mathbf{A}}^{\#}\mathbf{A})_{pp}|^2} \tag{17}$$

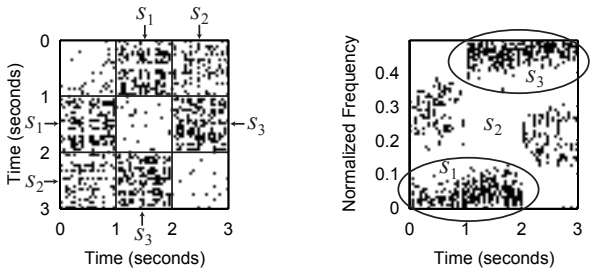
If $\hat{\mathbf{A}}$ is a good estimate of $\mathbf{A}$, $\hat{\mathbf{A}}^{\#}\mathbf{A}$ is close to diagonal, and the ISR is near zero.

We tested the performance of the separation algorithms over 500 Monte-Carlo runs. At each iteration, we drew another set of random signals $s_i(t)$ and a mixing matrix from a uniform distribution with elements in the range $(0, 1)$. We repeated this experiment with $\delta f \in [0, 0.002, 0.01, 0.05, 0.2]$. Table 1 shows the average maximum ISR for each δf. The two approaches perform comparably when the sources are sufficiently dissimilar. However, as δf approaches zero, the performance of the time-time separation improves relative to the time-frequency separation. Therefore, repetitive structure contains additional information for source separation that does not exist in spatial time-frequency distributions.

In our second experiment, we compare time-time separation and time-frequency separation using highly similar musical audio from the Iowa Musical Instrument Samples Database [18]. We extracted one-second examples of the same note played on bass clarinet, B♭ clarinet, and E♭ clarinet. These instruments produce quite similar frequency spectra as shown by the log of their time-frequency distributions in Figure 5. The range from light to dark indicates mean energy to max energy. The horizontal lines are harmonics that overlap nearly perfectly. The self-similarity or time-time distribution of the bass clarinet ($[\mathbf{S_{ss}}(t_1, t_2)]_{11}$), B♭ clarinet ($[\mathbf{S_{ss}}(t_1, t_2)]_{22}$), and E♭ clarinet ($[\mathbf{S_{ss}}(t_1, t_2)]_{33}$) are shown in Figure 6(a), 6(e), and 6(i), respectively. The cross-correlations are contained in the off-diagonal matrices of Figure 6. The matrix formed by connecting the matrices in Figure 6 is the time-time distribution of a recording containing the three instruments played consecutively. If the sources were not correlated the off-diagonal matrices would be white (*i.e.*, no correlation). Here, these sources are highly correlated. The autoterms selected for this example are shown in Figure 7. In spite of the similarity of the instruments, many time-time autoterms are identified. The alternating black and white lines perpendicular to the main diagonal indicate the fluctuating energy pattern in the clarinet sources. Each color change identifies when the energy crosses the energy threshold. The den-

δf	Time-time ISR	Time-frequency ISR
0.000	0.0832	0.1357
0.002	0.0825	0.1344
0.010	0.0767	0.1276
0.050	0.0380	0.0493
0.200	0.0082	0.0087

Table 1: *Average maximum interference-to-signal ratio (ISR) for time-time and time-frequency separation as a function of dissimilarity (δf)*



(a) Time-time (b) Time-frequency

Figure 4: *Autoterms selected from similarity experiment*

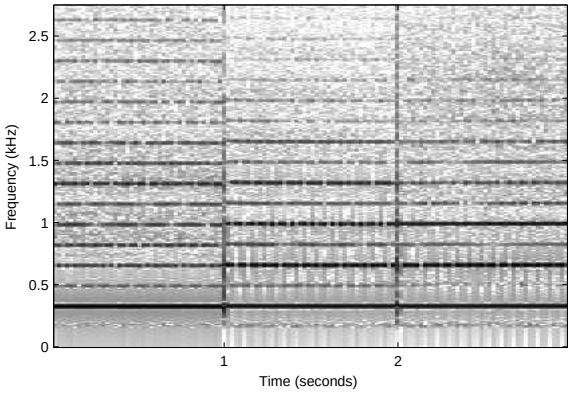
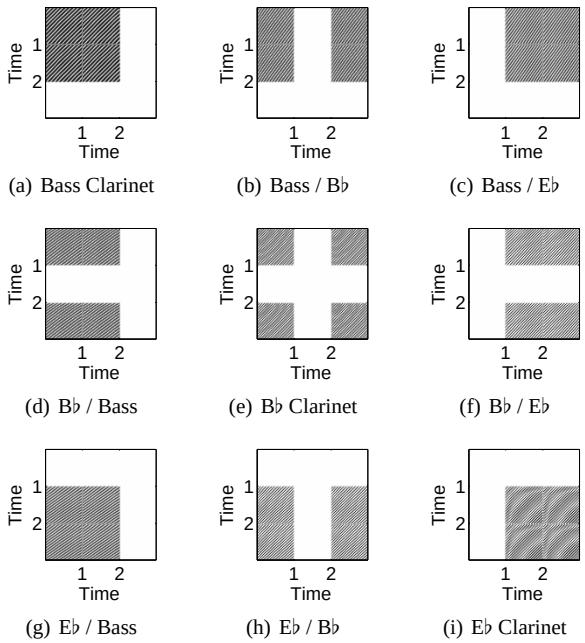


Figure 5: *Time-frequency distribution for three clarinets*



(a) Bass Clarinet (b) Bass / B♭ (c) Bass / E♭

(d) B♭ / Bass (e) B♭ Clarinet (f) B♭ / E♭

(g) E♭ / Bass (h) E♭ / B♭ (i) E♭ Clarinet

Figure 6: *Time-time distribution matrices between and within instruments*

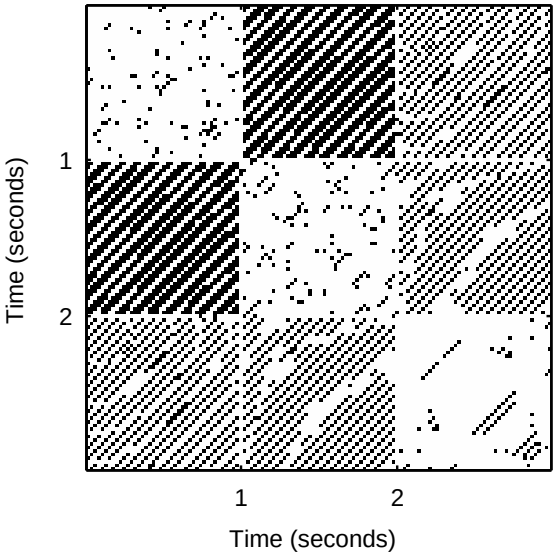


Figure 7: *Time-time autoterms (in black) from clarinets example*

sity of the autoterms reflect the same pattern as Figure 4(a) because the activation pattern is the same.

For all mixing matrices that we generated randomly, time-time separation estimates

$$\hat{\mathbf{A}}_{tt}^{\#}\mathbf{A} = \begin{bmatrix} 1.0001 & -0.0044 & -0.0044 \\ -0.0483 & 1.0012 & 0.0074 \\ 0.0267 & -0.0308 & 1.0004 \end{bmatrix} \quad (18)$$

with an ISR of 0.0488. Time-frequency separation estimates

$$\hat{\mathbf{A}}_{tf}^{\#}\mathbf{A} = \begin{bmatrix} 1.0005 & -0.0069 & 0.0310 \\ -0.0443 & 0.9874 & -0.1906 \\ -0.0178 & 0.1682 & 0.9817 \end{bmatrix} \quad (19)$$

with an ISR of 0.1982. Because the instruments are non-stationary with highly overlapping frequency components, time-time separation outperforms time-frequency separation. These results were confirmed by listening to the estimated source audio. Sections of inactivity in the original source audio are silent in a perfect reconstruction. During these sections, neither technique estimated silent sources. However, time-time estimated sources are clearly quieter than their time-frequency counterparts.

The activation patterns in the previous two experiments were constructed in order to emphasize time-time autoterms. Now, we show the performance of our algorithm on a real musical signal. Using a 20-second excerpt from a multi-track recording we artificially mix the bass guitar and organ tracks. Although the two sources overlap most of the time, there are times when only one source is active. Figure 8 shows that both algorithms leverage time points when only one source is present. The autoterms chosen by time-time separation focus on the large "plus" symbol centered at 12 seconds when the organ stops playing. These points are also chosen by time-frequency separation and illustrated by the short dense low-frequency content of the bass around 12 seconds in Figure 8(b). This is the only place where the low-frequency content is present without overlapping organ content. For all mixing matrices that we generated randomly, time-time separation estimates

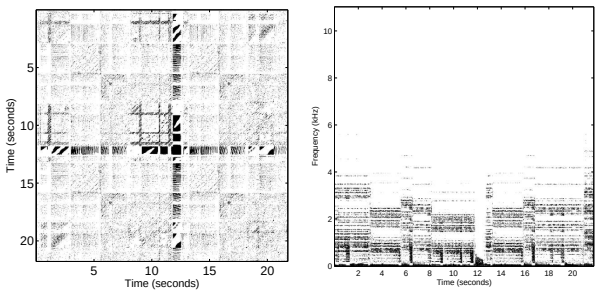
$$\hat{\mathbf{A}}_{\mathrm{tt}}^{\#}\mathbf{A} = \begin{bmatrix} 1.0000 & 0.0013 \\ 0.0120 & 1.0001 \end{bmatrix} \quad (20)$$

with an ISR of 0.0120. Time-frequency separation estimates the mixing matrix as

$$\hat{\mathbf{A}}_{\mathrm{tf}}^{\#}\mathbf{A} = \begin{bmatrix} 1.0000 & -0.0024 \\ 0.0157 & 1.0001 \end{bmatrix} \quad (21)$$

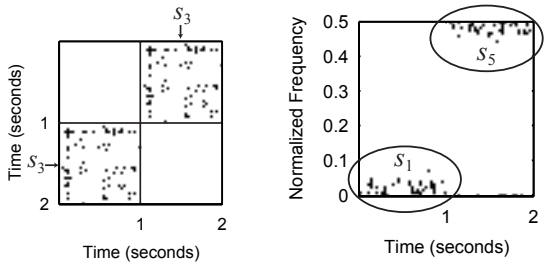
with an ISR of 0.0157. In this real musical example, repetitive structure is as informative for source separation as time-frequency structure.

Our final experiment is a synthetic version of the "bell tower" example. That is, the same source is presented twice while the sources surrounding it change. We expect this to improve the separation of the repeated source when using time-time separation. We construct 5 sources using Equation 16 with $f_1 = 0.05$, $f_2 = 0.15$, $f_3 = 0.25$, $f_4 = 0.35$, and $f_5 = 0.45$. Sources s_1 and s_2 are active for the first one-second segment. Sources s_4 and s_5 are active for the second one-second segment, and source s_3 is the obscuring source that plays the whole time. Figure 9 shows the autoterms selected by time-time and time-frequency separation. Time-frequency analysis only finds autoterms associated with sources s_1 and s_5 because there is less overlap at the edge of the spectrum. The time-time autoterms accurately identify s_3 at repetitions between each half of the signal (*i.e.*, the annotated first and third quadrant of Figure 9(a)). We use the principal eigenvector of the time-time autoterms as a column vector $\mathbf{u}$ and estimate the spatial position of s_3 as $\mathbf{W}^{\#}\mathbf{u}$. We can estimate the ISR of this spatial position by inserting it into $\mathbf{A}$ to form $\mathbf{A}'$ and computing the ISR between $\mathbf{A}$ and $\mathbf{A}'$. Over 500 Monte-Carlo trials, we estimated an ISR of 0.0176, whereas time-frequency separation could not identify any s_3 autoterms with which to separate.



(a) Time-time autoterms (b) Time-frequency autoterms

Figure 8: *Autoterm selection for bass guitar and organ*



(a) Time-time autoterms (b) Time-frequency autoterms

Figure 9: *Autoterm selection for bell tower example*

6. CONCLUSIONS AND FUTURE WORK

We present a novel spatial time-time distribution source separation algorithm that leverages the repetitive structure of sources. This requires that each source has a unique repetition (*i.e.*, time-time autoterm). Repetitions do not have to be identical, only correlated. When sources repeat uniquely, our time-time separation performs comparably to time-frequency separation. When sources have overlapping time-frequency distributions, our method outperforms time-frequency separation.

Time-time separation is an alternative to time-frequency separation when sources exhibit unique repetitions. Our future work includes combining these methods in order to leverage the repetitive and time-frequency separateness of sources. Algorithmically finding time-time-frequency autoterms is straightforward. However, the shear number of time-time-frequency points makes it unattractive to compute.

7. REFERENCES

[1] S.-I. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind source separation," in *Advances in Neural Information Processing Systems 8*, pp. 757–763. MIT Press, 1996.

[2] A.J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[3] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, 1993.

[4] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[5] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[6] A. Souloumiac, "Blind source detection and separation using second order non-stationarity," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1995, vol. 3, pp. 1912–1915.

[7] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[8] A. Belouchrani and M. G. Amin, "Blind source separation based on time-frequency signal representations," *IEEE Transactions on Signal Processing*, vol. 46, no. 11, pp. 2888–2897, 1998.

[9] C. Févotte and C. Doncarli, "Two contributions to blind source separation using time-frequency distributions," *IEEE Signal Processing Letters*, vol. 11, no. 3, pp. 386–389, 2004.

[10] A. Holobar, C. Févotte, C. Doncarli, and D. Zazula, "Single autoterms selection for blind source separation in time-frequency plane," in *Proceedings of the European Signal Processing Conference*, Toulouse, France, 2002.

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

[12] 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.

[13] 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.

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

[15] T. A. C. M. Claasen and W. F. G. Mecklenbräuker, "The Wigner distribution - a tool for time-frequency signal analysis, part 1: Continuous-time signals"," *Philips Journal of Research*, vol. 35, no. 3, pp. 217–250, 1980.

[16] 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.

[17] 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.

[18] L. Fritts, "University of Iowa Musical Instrument Samples Database," 1997, available online at http://theremin.music.uiowa.edu.