ON THE APPROXIMATE W-DISJOINT ORTHOGONALITY OF SPEECH

Scott Rickard[*] *and Özgür Yılmaz*[†]

Siemens Corporate Research, USA
Princeton University, USA

ABSTRACT

It is possible to blindly separate an arbitrary number of sources given just two anechoic mixtures provided the time-frequency representations of the sources do not overlap, a condition which we call W-disjoint orthogonality. We define a power weighted two-dimensional histogram constructed from the ratio of the time-frequency representations of the mixtures which is shown to have one peak for each source with peak location corresponding to the relative amplitude and delay mixing parameters. All of the time-frequency points which yield estimates in a given peak are exactly all the non-zero magnitude components of one of the sources. We introduce the concept of *approximate W-disjoint orthogonality*, present experimental results demonstrating the level of approximate W-disjoint orthogonality of speech in mixtures of various order, and show that even with imperfect W-disjoint orthogonality the histogram can be used to determine the mixing parameters and separate sources. Example demixing results can be found online:
http://www.princeton.edu/~srickard/bss.html

1. INTRODUCTION

The blind source separation technique introduced in [1] relied on the assumption that the sources were W-disjoint orthogonal. Despite the fact that for signals of practical interest, namely speech, this assumption is often violated, the technique still achieves good demixing results[2]. These results are possible because speech signals exhibit a level of "approximate" W-disjoint orthogonality. In this paper, we propose a definition of approximate W-disjoint orthogonality and present experimental results confirming this property for speech signals. Moreover, we show that the effects of the imperfect disjointness on the demixing algorithm are limited.

Other work which leverages properties of sources which are similar in spirit to approximate W-disjoint orthogonality include [3, 4], which considered time domain disjoint sources, and [5], which exploited the sparsity of the short-time Fourier transform when applied to speech and music signals. All of these methods, however, only considered instantaneous mixing.

The anechoic mixing model and source assumptions are discussed in Section 2. The two-dimensional power weighted mixing parameter histogram is introduced and its role in demixing is presented in Section 3. Section 4 defines approximate W-disjoint orthogonality and presents experimental results for speech mixtures. The effect of approximate W-disjoint orthogonality on the histogram is discussed in Section 5.

[*] Scott Rickard is with the Audio and Signal Processing Group, Siemens Corporate Research and the Program in Applied and Computational Mathematics, Princeton University. srickard@math.princeton.edu

[†] Özgür Yılmaz is with the Program in Applied and Computational Mathematics, Princeton University. oyilmaz@math.princeton.edu

ICASSP2002, Orlando, Florida, USA, May 13-17, 2002.

2. MIXING MODEL AND SOURCE ASSUMPTIONS

Consider measurements of a pair of sensors where only the direct path is present. In this case, without loss of generality, we can absorb the attenuation and delay parameters of the first mixture, $x_1(t)$, into the definition of the sources. The two mixtures can thus be expressed as,

$$x_1(t) = \sum_{j=1}^{N} s_j(t), \quad (1)$$

$$x_2(t) = \sum_{j=1}^{N} a_j s_j(t-\delta_j), \quad (2)$$

where $s_j(t)$, $j = 1, \ldots, N$, are the N sources, δ_j is the arrival delay between the sensors of source j, and a_j is a relative attenuation factor corresponding to the ratio of the attenuation of the paths between sources and sensors. We use Δ to denote the maximal possible delay between sensors, and thus, $|\delta_j| \leq \Delta, \forall j$.

Our goal is to recover the original sources given only the mixtures. In order to accomplish this, we assume the sources are pairwise W-disjoint orthogonal and satisfy the narrowband assumption for array processing. Both of these concepts are discussed below.

We call two functions $s_1(t)$ and $s_2(t)$ **W-disjoint orthogonal** if, for a given a windowing function $W(t)$, the supports of the windowed Fourier transforms of $s_1(t)$ and $s_2(t)$ are disjoint[1]. The windowed Fourier transform of $s_j(t)$ is defined,

$$F^W(s_j(\cdot))(\omega,\tau) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} W(t-\tau) s_j(t) e^{-i\omega t} dt, \quad (3)$$

which we will refer to as $\hat{s}_j(\omega,\tau)$ where appropriate. The W-disjoint orthogonality assumption can be stated concisely,

$$\hat{s}_1(\omega,\tau)\hat{s}_2(\omega,\tau) = 0, \forall \omega, \tau. \quad (4)$$

W-disjoint orthogonality is different and in general a stronger condition than statistical orthogonality. This difference is illustrated in Figure 1 where it is demonstrated that speech signals are approximately W-disjoint orthogonal whereas independent white noise signals, while being statistically orthogonal, are not approximately W-disjoint orthogonal. A formal definition of approximate W-disjoint orthogonality is introduced in Section 4.

When $W(t) = 1$, the following is a property of the Fourier transform,

$$F^W(s_j(\cdot - \delta))(\omega,\tau) = e^{-i\omega\delta} F^W(s_j(\cdot))(\omega,\tau). \quad (5)$$

We employ the narrowband assumption in array processing that implies for our purposes that Equation 5 holds for all δ, $|\delta| \leq \Delta$, even when $W(t)$ has finite support[6].

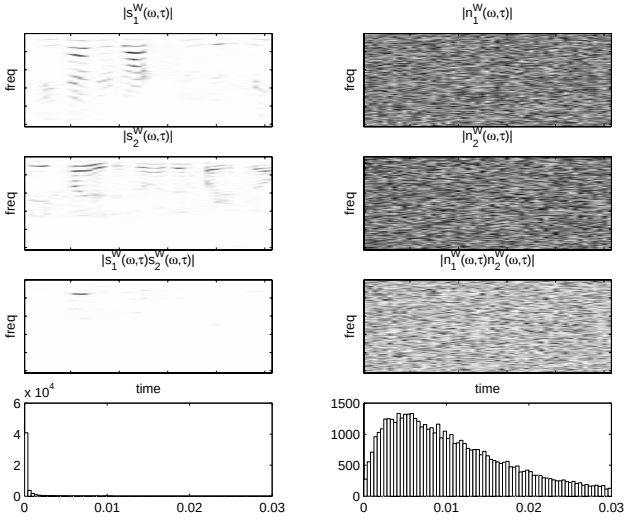


Fig. 1. Example of W-disjoint orthogonality. The top three left column figures are grey scale images of $|\hat{s}_1(\omega,\tau)|$, $|\hat{s}_2(\omega,\tau)|$, and $|\hat{s}_1(\omega,\tau)\hat{s}_2(\omega,\tau)|$ for two speech signals $s_1(t)$ and $s_2(t)$ normalized to have unit energy. The top three right column figures are grey scale images of $|\hat{n}_1(\omega,\tau)|$, $|\hat{n}_2(\omega,\tau)|$, and $|\hat{n}_1(\omega,\tau)\hat{n}_2(\omega,\tau)|$ for two independent white noise signals $n_1(t)$ and $n_2(t)$ normalized to have unit energy. A Hamming window of length 32 ms was used as $W(t)$ and all signals had length 1.5 seconds. $|\hat{s}_1(\omega,\tau)\hat{s}_2(\omega,\tau)|$ contains far fewer large components than $|\hat{n}_1(\omega,\tau)\hat{n}_2(\omega,\tau)|$. This is confirmed in the bottom row which contains histograms of the values in $|\hat{s}_1(\omega,\tau)\hat{s}_2(\omega,\tau)|$ and $|\hat{n}_1(\omega,\tau)\hat{n}_2(\omega,\tau)|$ respectively. Note, almost all values in the voice histogram are close to zero, while there are a significant number of non-zero values in the noise histogram. Thus the speech signals approximately satisfy the W-disjoint orthogonality condition while the independent white noise signals do not.

2.1. Amplitude-Delay Estimation

Using the narrowband assumption, we can rewrite the model from Equations 1 and 2 in the time-frequency domain,

$$\begin{bmatrix} \hat{x}_1(\omega,\tau) \\ \hat{x}_2(\omega,\tau) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ a_1 e^{-i\omega\delta_1} & \dots & a_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} \hat{s}_1(\omega,\tau) \\ \vdots \\ \hat{s}_N(\omega,\tau) \end{bmatrix} \tag{6}$$

Assuming the sources are pairwise W-disjoint orthogonal, at most one of the N sources will be non-zero for a given (ω,τ), and thus,

$$\begin{bmatrix} \hat{x}_1(\omega,\tau) \\ \hat{x}_2(\omega,\tau) \end{bmatrix} = \begin{bmatrix} 1 \\ a_J e^{-i\omega\delta_J} \end{bmatrix} \hat{s}_J(\omega,\tau), \tag{7}$$

where $J = j(\omega,\tau)$, in a slight abuse of notation, is the index of the source active at (ω,τ). The key observation is that the ratio of the time-frequency representations of the mixtures is a function of the mixing parameters only and does not depend on the sources. Therefore, we can calculate the relative amplitude and delay parameters associated with the source active at (ω,τ) using,

$$(a_J,\delta_J) = \left(\left| \frac{\hat{x}_2(\omega,\tau)}{\hat{x}_1(\omega,\tau)} \right|, \frac{1}{\omega} \angle \frac{\hat{x}_1(\omega,\tau)}{\hat{x}_2(\omega,\tau)} \right), \tag{8}$$

where $\angle a e^{i\phi} = \phi$, $-\pi < \phi \leq \pi$. Note, the accurate calculation of δ_J requires that,

$$|\omega\delta_J| < \pi. \tag{9}$$

3. HISTOGRAM DEFINITION AND DEMIXING

In order to demix, we construct a two dimensional weighted histogram whose peaks are in one-to-one correspondence with the amplitude and delay mixing parameters of each source. This allows for the calculation of the mixing parameters, which are then used to construct time-frequency masks which demix the mixtures.

In practice, rather than working with the continuous windowed Fourier transform, we use its discrete counterpart, $\hat{s}_j(k\omega_0, l\tau_0)$, $\forall k,l \in \mathbb{Z}$, where ω_0 and τ_0 are the frequency and time grid spacing parameters. It is well known that for any appropriately chosen window function, if ω_0 and τ_0 are small enough, $s_j(t)$ can be reconstructed from $\hat{s}_j(k\omega_0, l\tau_0), \forall k,l \in \mathbb{Z}$. For more details, consult [7].

The amplitude delay mixing parameter estimates associated with $(k\omega_0, l\tau_0)$ are,

$$(a(k,l),\delta(k,l)) = \left(\left| \frac{\hat{x}_2(k\omega_0,l\tau_0)}{\hat{x}_1(k\omega_0,l\tau_0)} \right|, \frac{1}{k\omega_0} \angle \frac{\hat{x}_1(k\omega_0,l\tau_0)}{\hat{x}_2(k\omega_0,l\tau_0)} \right). \tag{10}$$

A two dimensional weighted histogram can be constructed in (a,δ) space from the $(a(k,l),\delta(k,l))$ pairs as follows. First, we define time-frequency mask for (a,δ),

$$M_{(a,\delta,\Delta_a,\Delta_\delta)}(k,l) = \begin{cases} 1: & |\ln a(k,l) - \ln a| < \Delta_a/2 \text{ and} \\ & |\delta(k,l) - \delta)| < \Delta_\delta/2 \\ 0: & \text{otherwise} \end{cases} \tag{11}$$

where Δ_a and Δ_δ are the amplitude and delay resolution widths of the histogram. Then, the weighted histogram can be defined as,

$$\begin{aligned} h(a,\delta) = \quad & \|M_{(a,\delta,\Delta_a,\Delta_\delta)}(k,l)\hat{x}_1(k\omega_0,l\tau_0)\|^2 + \\ & \|M_{(a,\delta,\Delta_a,\Delta_\delta)}(k,l)\hat{x}_2(k\omega_0,l\tau_0)\|^2 \end{aligned} \tag{12}$$

where $\|\cdot\|$ denotes the L^2 norm. Of main interest are the locations of the histogram's peaks and the surrounding region. These shall be used to identify sources and to estimate demixing errors.

The histogram for sources which satisfy (4), (5), and (9) will consist of N peaks with rectangular support with dimensions Δ_a-by-Δ_δ centered at (a_j,δ_j), $j = 1,\dots,N$. The height of each peak will be proportional to the sum power of the corresponding source in the mixtures.

To demix, one creates the time-frequency mask corresponding to each peak in the histogram using (11) and uses it to mask one of the mixtures to produce the original source time-frequency representation. For example, for a peak located at (a_j,δ_j), one obtains source j via,

$$\hat{s}_j(k\omega_0,l\tau_0) = M_{(a_j,\delta_j,\Delta_a,\Delta_\delta)}(k,l)\hat{x}_1(k\omega_0,l\tau_0), \forall k,l. \tag{13}$$

4. APPROXIMATE W-DISJOINT ORTHOGONALITY

Clearly, the W-disjoint orthogonality assumption is not strictly satisfied for our signals of interest. We introduce here a measure of approximate W-disjoint orthogonality.

First, we define,

$$y_j(t) = \sum_{\substack{i=1 \\ i\neq j}}^{N} s_i(t), \tag{14}$$

so that $y_j(t)$ is the summation of the sources interfering with source j. Then, consider the time-frequency mask,

$$\Phi_{(j,x)}(k,l) = \begin{cases} 1 & 20\log(|\hat{s}_j(k\omega_0, l\tau_0)|/|\hat{y}_j(k\omega_0, l\tau_0)|) > x \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

and the resulting energy ratio,

$$r_j(x) = \|\Phi_{(j,x)}(k,l)\hat{s}_j(k\omega_0, l\tau_0)\|^2 / \|\hat{s}_j(k\omega_0, l\tau_0)\|^2, \tag{16}$$

which measures the percentage of energy of source j for time-frequency points where it dominates the other sources by x dB. We propose $r_j(x)$ as a measure of W-disjoint orthogonality. For example, Figure 2 shows $r_j(x)$ averaged for groups of speech mixtures of different orders and demonstrates the approximate W-disjoint orthogonality of speech in mixtures of up to 10 signals.

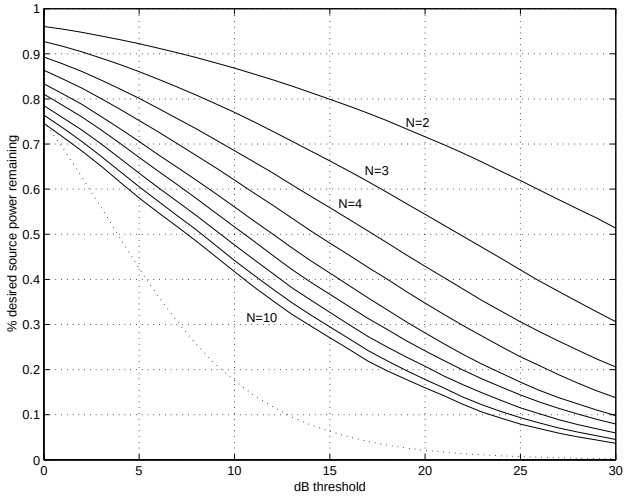
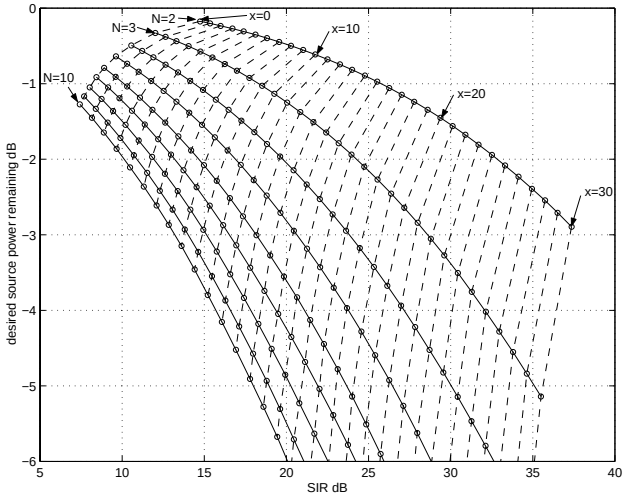


Fig. 2. Approximate W-Disjoint Orthogonality. Plot of $r_j(x)$ for $x = 0, 1, \ldots, 30$ for $N = 1, 2, \ldots, 10$. Note that $r_j(15) \approx .8$ for the $N = 2$ case, and thus say that the *speech signals in mixtures of order 2 are* 80% *W-disjoint orthogonal at 15 dB.* If we can correctly map time-frequency points with 15 dB or more single source dominance to the correct corresponding output partition, we can recover 80% of the energy of the original sources. For higher order mixtures, some level of approximate W-disjoint orthogonality is maintained. For example, comparing one source to the sum of three others others, we have 80% W-disjoint orthogonality at 5 dB. The dotted line is for two independent Gaussian white noise processes. The figure was generated using speech files taken from the TIMIT database and each line represents the average over hundreds of tests.

An important demixing performance measure is the signal to interference ratio of the outputs. We can calculate the SIR associated with a time-frequency mask using,

$$\text{SIR}_j(x) = \frac{\|\Phi_{(j,x)}(k,l)\hat{s}_j(k\omega_0, l\tau_0)\|^2}{\|\Phi_{(j,x)}(k,l)\hat{y}_j(k\omega_0, l\tau_0)\|^2}, \tag{17}$$

which measures the signal to interference ratio for time-frequency points where source j dominates the sum of the other sources by x dB. Figure 3 shows plots of $r_j(x)$ (in dB) versus $\text{SIR}_j(x)$ for mixtures of various orders and a table of $(r_j(x), \text{SIR}_j(x))$ pairs.



N	$x = 5$	$x = 10$	$x = 15$
2	$(0.92, 18.10)$	$(0.87, 21.76)$	$(0.80, 25.53)$
3	$(0.86, 15.50)$	$(0.77, 19.27)$	$(0.66, 23.19)$
4	$(0.80, 14.14)$	$(0.69, 18.05)$	$(0.56, 22.02)$
5	$(0.75, 13.31)$	$(0.62, 17.21)$	$(0.48, 21.25)$
10	$(0.58, 11.33)$	$(0.42, 15.22)$	$(0.27, 19.44)$

Fig. 3. Demixing time-frequency mask performance. Plot contains $r_j(x)$ (in dB) versus $\text{SIR}_j(x)$ for $x = 0, 1, \ldots, 30$ for $N = 1, 2, \ldots, 10$. Table contains $(r_j(x), \text{SIR}_j(x))$ for $N = 2, 3, 4, 5, 10$ for $x = 5, 10, 15$ dB. For example, using the $x = 10$ dB mask in pairwise mixing yields 21.76 dB output SIR while maintaining 87% of the desired source power.

5. PEAK SPREADING

The approximate W-disjoint orthogonality of speech signals brings about a spreading of the peak region in the (a, δ) histogram. Consider the estimation of (a, δ) mixing parameters associated with source j in the presence of interfering sources. Using (14) and defining,

$$y'_j(t) = \sum_{\substack{i=1 \\ i \neq j}}^{N} a_i s_i(t - \delta_i), \tag{18}$$

the time-frequency mixing equation (6) becomes,

$$\begin{bmatrix} \hat{x}_1(\omega,\tau) \\ \hat{x}_2(\omega,\tau) \end{bmatrix} = \begin{bmatrix} 1 + \hat{y}_j(\omega,\tau)/\hat{s}_j(\omega,\tau) \\ a_j e^{-i\omega\delta_j} + \hat{y}'_j(\omega,\tau)/\hat{s}_j(\omega,\tau) \end{bmatrix} \hat{s}_j(\omega,\tau). \tag{19}$$

Defining,

$$be^{-i\omega\beta} = \hat{y}_j(\omega,\tau)/\hat{s}_j(\omega,\tau), \tag{20}$$

$$ce^{-i\omega\gamma} = \hat{y}'_j(\omega,\tau)/\hat{s}_j(\omega,\tau), \tag{21}$$

with $b, c \geq 0$ and $\beta, \gamma \in [-\pi, \pi)$, the time-frequency ratio of the mixtures becomes,

$$\frac{\hat{x}_2(\omega,\tau)}{\hat{x}_1(\omega,\tau)} = \frac{a_j e^{-i\omega\delta_j} + be^{-i\omega\beta}}{1 + ce^{-i\omega\gamma}}, \tag{22}$$

and the estimates of the mixing parameters associated with (ω, τ) are,

$$\ln \left| \frac{\hat{x}_2(\omega,\tau)}{\hat{x}_1(\omega,\tau)} \right| = \ln a_j + \ln \left(\frac{1+2\frac{b}{a_j}\cos\omega\left(\beta-\delta_j\right)+\left(\frac{b}{a_j}\right)^2}{1+2c\cos\omega\gamma+c^2} \right)^{\frac{1}{2}} \quad (23)$$

$$\measuredangle \frac{\hat{x}_1(\omega,\tau)}{\hat{x}_2(\omega,\tau)} = \delta_j + \tfrac{1}{\omega}\arctan\left(\frac{c\sin\omega\gamma}{1+c\cos\omega\gamma}\right) - \tfrac{1}{\omega}\arctan\left(\frac{\frac{b}{a_j}\sin\omega\left(\beta-\delta_j\right)}{1+\frac{b}{a_j}\cos\omega\left(\beta-\delta_j\right)}\right). \quad (24)$$

Observe that the interfering sources bring about an error term in the mixing parameter estimates.

We can evaluate the effect of these errors on the amplitude delay estimate pairs by assuming that the phases in (20) and (21) are uniformly independently distributed from $-\pi$ to π. Using the approximate W-disjoint orthogonality results for speech mixtures, we can numerically evaluate the expected histogram peak region shape generated from (23) and (24). For example, Figure 4 shows the expected peak shape of one source for a speech mixture of order four. An example of the peak spreading is shown in Figure 5 which contains the histogram for a four speech signal mixing example. Sample mixture and demixed sound files corresponding to this histogram can be found on the webpage listed in the abstract. A more detailed presentation of these results, including the effect of noise and echoes, is being prepared.

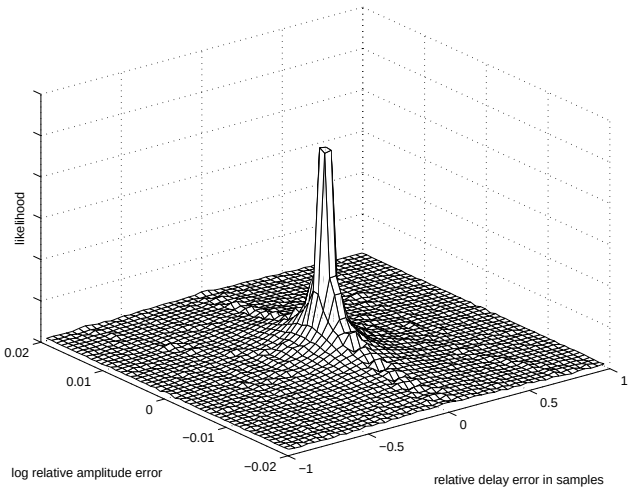


Fig. 4. Expected peak shape due to estimation errors caused by interference of three other sources assuming the phases in (20) and (21) are uniformly distributed in $[-\pi, \pi)$ and using the approximate W-disjoint orthogonality results presented in Figure 2.

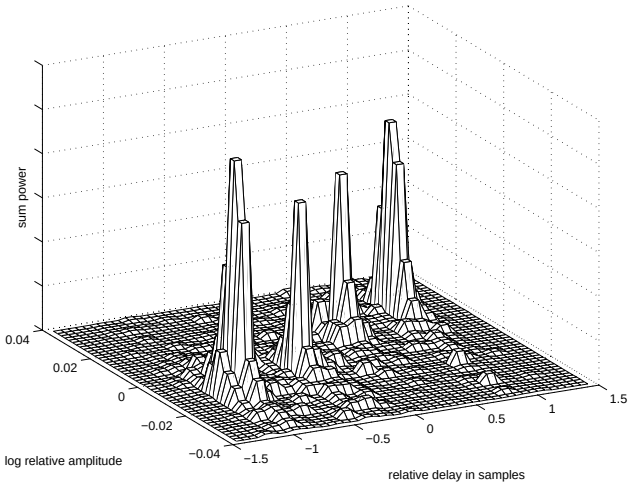


Fig. 5. Histogram for four speech signals mixed using (log-amplitude,delay) parameters, $(-0.013, -1.0)$, $(-0.009, -0.33)$, $(0.009, 0.33)$, and $(0.018, 1.0)$, which match well with the location of the histogram's peaks.

6. REFERENCES

[1] A. Jourjine, S. Rickard, and Ö. Yılmaz, "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures," in *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, June 5-9 2000, vol. 5, pp. 2985–8.

[2] S. Rickard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation," in *3rd International Conference on Independent Component Analysis and Blind Source Separation*, San Diego, CA, December 9-12 2001.

[3] M. Van Hulle, "Clustering approach to square and non-square blind source separation," in *IEEE Workshop on Neural Networks for Signal Processing (NNSP99)*, August 1999, pp. 315–323.

[4] T-W. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Letters*, vol. 6, no. 4, pp. 87–90, apr 1999.

[5] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," in *International Workshop on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland, June 19–22 2000, pp. 87–92.

[6] R. Balan, J. Rosca, S. Rickard, and J. O'Ruanaidh, "The influence of windowing on time delay estimates," in *Proceedings of the 35th Annual Conference on Information Sciences and Systems (CISS2000)*, Princeton, NJ, March 15-17 2000, vol. 1, pp. WP1–(15–17).

[7] I. Daubechies, *Ten Lectures on Wavelets*, chapter 3, SIAM, Philadelphia, PA, 1992.