AN EXTENSION FOR SOURCE SEPARATION TECHNIQUES AVOIDING BEATS

Harald Viste

Communication Systems Department
Swiss Federal Institute of Technology Lausanne
harald.viste@epfl.ch

Gianpaolo Evangelista

Dep.of Physical Sciences, Univ. of Naples, Italy
Swiss Federal Institute of Technology Lausanne
gianpaolo.evangelista@epfl.ch

ABSTRACT

The problem of separating individual sound sources from a mixture of these, known as Source Separation or Computational Auditory Scene Analysis (CASA), has become popular in the recent decades. A number of methods have emerged from the study of this problem, some of which perform very well for certain types of audio sources, e.g. speech. For separation of instruments in music, there are several shortcomings. In general when instruments play together they are not independent of each other. More specifically the time-frequency distributions of the different sources will overlap. Harmonic instruments in particular have high probability of overlapping partials. If these overlapping partials are not separated properly, the separated signals will have a different sensation of roughness, and the separation quality degrades.

In this paper we present a method to separate overlapping partials in stereo signals. This method looks at the shapes of partial envelopes, and uses minimization of the difference between such shapes in order to demix overlapping partials. The method can be applied to enhance existing methods for source separation, e.g. blind source separation techniques, model based techniques, and spatial separation techniques. We also discuss other simpler methods that can work with mono signals.

1. INTRODUCTION

When instruments play together, their signals are mixed together. Source separation is simply the problem of obtaining the original source signals from the recorded mixture. The problem with music signals, as opposed to other types of signals like e.g. speech, is that the sources are normally not independent. First of all, the instruments are dependent among each others in time, due to the fact that they all follow the same underlying tempo and rhythm of the music piece. In addition, for melodic instruments which have a pitch and harmonic structure, the notes are often related by harmonic intervals as well. When two partials fall within one critical band, the ear will not hear two separate sounds, but rather one combined sound. This is explained in [1]: *"When two sinusoids with slightly different frequencies are added together, they resemble a single sinusoid, with frequency equal to the mean frequency of the two components, but whose amplitude fluctuates at a regular rate. These fluctuations in amplitude are known as 'beats'."* These beats occur at a rate equal to the frequency difference of the two components.

If the beats are slow, this results in audible loudness fluctuations. In the combined sound these fluctuations sounds natural, but in the separated signals such fluctuations can be very annoying if present. For faster beats the fluctuations can not be heard separately, but rather as an increase in the roughness of the sound. This roughness is related to the consonance [2], and depends on frequency. Maximum roughness occurs for beat frequencies in the range 30-70 Hz [3]. However, partials that are 30-70 Hz apart, are quite well handled by existing source separation methods. We will therefore concentrate on slow beats.

Figure 1 shows the spectrogram of two trumpet notes, an F at about 349 Hz, and a C at about 523 Hz. Every third partial of the F overlaps every second partial of the C. For the partials slightly above 1, 2 and 3 kHz, one can see the beats as amplitude fluctuations. By counting the number of fluctuations per second, we can see that the beat frequencies for these are about 3, 6 and 9 Hz, respectively. All the other partials have quite constant amplitude during their duration. Obviously, the fluctuations we see in the

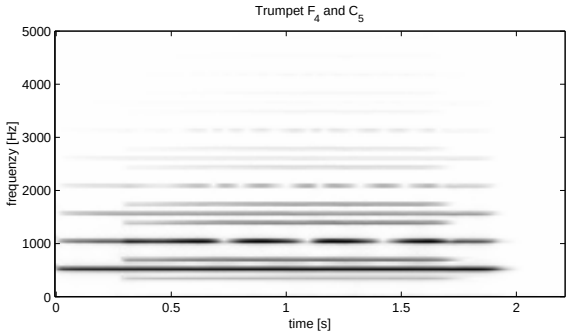

Figure 1: *Spectrogram of two trumpet notes with overlapping partials and beats (just above 1, 2 and 3 kHz).*

figure is an effect of the time-frequency representation we choose. If we had chosen longer time windows, the "overlapping" partials could have been seen as two separate partials with very close frequencies. However, since we study frequency separations of less than 10 Hz, this would require a window that is longer than the sound itself, and is not feasible. Also, we see that the period of the loudness fluctuations is not constant, and effectively the two partials may cross each other over time. Still it is interesting to note that this time-frequency representation matches the psychoacoustical effect that beats represent.

Clearly, the separation problem is an ill-posed problem. However, a human may be able to follow the tune of one of the instruments in the mixture, and this gives the motivation for attacking the problem.

This article is organized as follows. In section 2, we give an overview of the existing types of methods for source separation, and discuss the shortcomings of these for separation of music signals. Section 3 is devoted to our work on how to overcome these

problems. In particular, it explains the method we propose for separation of overlapping partials in multi-channel recordings. We also propose methods for partial shape reconstruction that can be used for mono signals as well. Finally, there is the conclusion.

2. EXISTING SOURCE SEPARATION METHODS

There are several approaches to the problem of separating sound sources. We briefly present three different types of separation methods, and discuss their drawbacks and shortcomings with respect to music signals.

2.1. Model based source separation

There are several methods that use a model of the sound sources to help separating them. The models range from low-level sinusoidal represenatations [4] to specific instrument models and higher level perceptual and cognitive models. Still, most methods in this category are based on sinusoidal models, e.g. [5, 6, 7, 8, 9]. These methods extract sinusoidal tracks from some time-frequency representation of the signal, and then apply grouping principles to assign these tracks to the different sources. Typically, these methods operate on mono signals.

There are two major problems with these methods with respect to instrument separation. Firstly, the methods do not work well for sources that don't exhibit a sinusoidal structure (partials). Secondly, even when the sources have partials, each partial is allocated to one (or more) sources. Thus, any beats due to overlapping partials will be present in the separated signals, and this can be perceptually very annoying.

2.2. Blind source separation

In the last ten years, a number of different approaches to blind source separation have evolved [10]. Early works studied the separation of instantaneous mixtures under various frameworks, e.g. neural networks [11], information theory [12], and statistics [13]. Later the methods were extended to convolved mixtures, [14, 15].

Although these approaches come from different areas of interest, they are all built on similar principles. Usually, iterative algorithms update the coefficients in a deconvolution matrix by minimizing some error measurement or by maximizing some information measurement. Then this matrix is used to separate the sources (deconvolve the mixtures). The basic assumption these methods are built on is that the sources are statistically independent. Even though this is not true in general for music, the methods can perform quite well. However, there are some drawbacks. They only work for multichannel mixtures, and in general there can be no more sources than sensors (number of channels). In addition, for real world mixtures, the FIR demixing filters need to be sufficiently long, which makes the computational complexity very high.

The costly deconvolution algorithm in the time-domain can be tranformed into a set of instantaneous demixing algorithms in time-frequency [16, 17]. This makes the computational cost more affordable, but introduces a new problem. In general the instantaneous demixing does not converge for the frequency bands where there are overlapping partials, due to the fact that the sources are not statistically independent.

2.3. Spatial separation

There exist some methods that exploits the physical locations of the sources in order to separate them, like e.g. beamforming [18]. Recently, the DUET method for separating sources in the time-frequency domain based on spatial cues was introduced in [19]. It is fast, simple and still performs quite well for music signals. In the following we are explaining this method in more detail. The method takes a stereo recording of the audio scene, $x_L(t)$ and $x_R(t)$, and computes the Short Time Fourier Transforms (STFT) for both channels, $X_L(\omega, t)$ and $X_R(\omega, t)$. Based on these, the relative amplitudes,

$$A(\omega, t) = \left| \frac{X_L(\omega, t)}{X_R(\omega, t)} \right|, \tag{1}$$

and phase delays,

$$D(\omega, t) = \frac{\angle \left(\frac{X_L(\omega,t)}{X_R(\omega,t)} \right)}{\omega}, \tag{2}$$

between the channels are computed. From these one can generate a 2-D histogram, as shown in figure 2. In this example, there is one

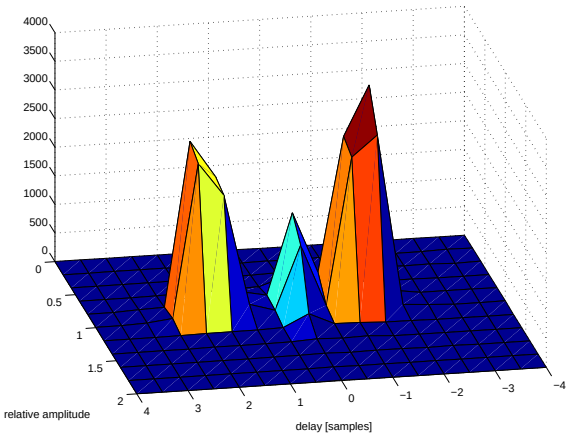

Figure 2: *2-D Histogram showing where the sources are located in (phase delay, relative amplitude) space.*

source on the left side (delay of two samples between left and right microphone), and one source on the right side ("delay" of minus one sample between left and right microphone). We also note that there is a phantom source in the middle (about 0 delay). This is in fact due to the overlapping partials, since the net effect of one sinusoid coming from the left and one from the right is a sinusoid coming from somewhere in between.

By applying a clustering algorithm to this histogram, one decide where the sources are, and how the bins of the STFT should be assigned in the separation phase. For each time-frequency bin, the relative amplitude and phase delay between the channels are used to assign the bin to one source. There are two interesting properties of this. Firstly this means that the method can work with sources that don't exhibit a strong partials structure, which was a major limitation of the sinusoidal models. Secondly, there may be more sources than there are sensors, which is a strong limitation in blind separation techniques. Although the method is very promising in these respects, we note the following constraints: The method is

based on the assumption that the sources are W-disjoint orthogonal [19]. Basically this means that each time frequency bin contains energy from only one source. Apparently this is not at all true for music. Also, the fact that each bin is assigned to one source exclusively can introduce new beats or binary "on/off" effects (the signal "comes" and "goes" in one frequency band as time evolves). Figure 3 clearly shows this in the bands where the partials overlap; the bins in these bands are assigned to either source in an alternating way.

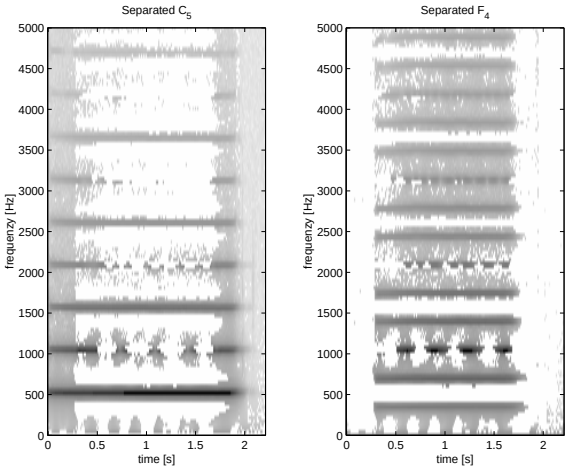

Figure 3: *Separated signals using the DUET method. Notice how the overlapping partials (just above 1, 2 and 3 kHz) are exclusively assigned to either source in an alternating way. The other partials are correctly assigned to one source only.*

Despite the mentioned drawbacks, the DUET method shows promising results, and we will se how the quality of the separated signals can be improved by using our proposed methods.

3. DEALING WITH OVERLAPPING PARTIALS

In the previous section we saw that the main problem with separation of instruments is related to overlapping partials. In this section we first propose a method for separation of 2 overlapping partials in a stereo mixture. Then we discuss other methods for partial shape reconstruction that can be used with mono signals as well. We also discuss how the separation quality can be improved when there are overlapping partials but no beats.

The methods we are introducing are based on the observation that the partials of a harmonic instrument have similar shapes. Figure 4 shows the envelope of the first 5 harmonics for the trumpet mixture (F_4 and C_5). We see that all the nonoverlapping partials have similar shapes (up to a scaling factor), and that we can deduce the wished shape for the overlapping partials from these

3.1. Partial Separation

To be able to separate overlapping partials, we need a multi channel mixture. At this point, the method we propose is only suitable for 2 overlapping partials, and we therefore consider the case with 2 sources and 2 sensors. Note, however, that since the method works on the frequency bands independently, it can still be used in

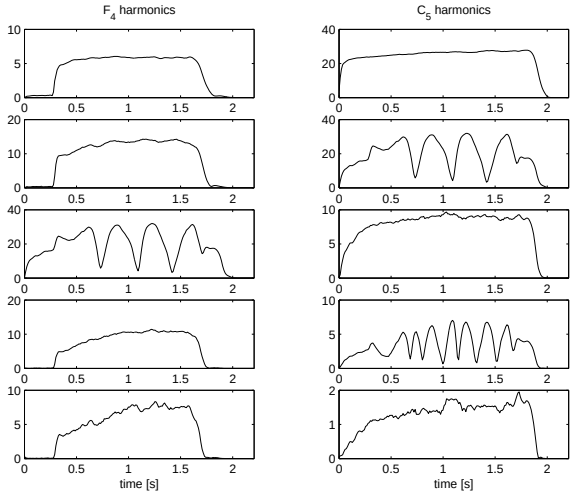

Figure 4: *First five partials of the F_4 (left) and the C_5 (right) in the trumpet mixture.*

systems with more than 2 sources. The assumption is that for each frequency band, only two of the sources are significant.

In the time domain, the mixing of the sources can be described as follows:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} h_0(t) & h_1(t) \\ h_2(t) & h_3(t) \end{pmatrix} * \begin{pmatrix} s_1(t) \\ s_2(t) \end{pmatrix}. \quad (3)$$

where x_i are the mixture signals, s_i are the source signals, and h_i are the mixing filters. We assume that these filters are time invariant. Without loss of generality, we can assume that h_0 and h_3 are the identity filters $\delta(t)$. If we take the short time Fourier transform (STFT), (convolution in time corresponds to multiplication in frequency), (3) becomes:

$$\begin{pmatrix} X_1(\omega,t) \\ X_2(\omega,t) \end{pmatrix} = \begin{pmatrix} 1 & H_1(\omega) \\ H_2(\omega) & 1 \end{pmatrix} \begin{pmatrix} S_1(\omega,t) \\ S_2(\omega,t) \end{pmatrix}. \quad (4)$$

Now, we consider a fixed ω that corresponds to a frequency band with overlapping partials. We observe the following:

- If the frequency band is narrow enough, the $H_1(\omega)$ and $H_2(\omega)$ in (4) can be assumed to be constant over the frequency range, and are thus simply two complex numbers. For simplicity we therefore omit the parameter ω in the notation.
- With a suitable sensor setup (i.e., two microphones closely spaced), the relative signal strengths between the sensor signals are close to 1. This means that H_1 and H_2 both lie close to the unit circle. In other setups, the relative signal strengths can be estimated from the relative amplitude $A(\omega,t)$ for the neighbouring non-overlapping partials.

We estimate the values of the two complex unknowns in the mixing matrix, $\hat{H}_1$ and $\hat{H}_2$. The inverse of this estimated mixing matrix (unscaled) can then be applied to estimate the separated sources $\hat{S}_i$

$$\begin{pmatrix} \hat{S}_1(\omega,t) \\ \hat{S}_2(\omega,t) \end{pmatrix} = \begin{pmatrix} 1 & -\hat{H}_1 \\ -\hat{H}_2 & 1 \end{pmatrix} \begin{pmatrix} X_1(\omega,t) \\ X_2(\omega,t) \end{pmatrix}, \quad (5)$$

which can be written out as:

$$\begin{pmatrix} \hat{S}_1(\omega,t) \\ \hat{S}_2(\omega,t) \end{pmatrix} = \begin{pmatrix} 1-\hat{H}_1H_2 & H_1-\hat{H}_1 \\ H_2-\hat{H}_2 & 1-\hat{H}_2H_1 \end{pmatrix} \begin{pmatrix} S_1(\omega,t) \\ S_2(\omega,t) \end{pmatrix}. \quad (6)$$

The closer one of these estimated values is to the real value, the less the presence of the unwanted signal in the corresponding separated source. When it reaches the optimal value, the separated partial is simply a scaled versions of the original partial before mixing. This applies to the two unknowns independently.

The estimation algortihm we propose is a simple recursively refined search for the two unknowns on the unit circle. For any chosen point, we calculate the separated partials. Then we calculate distance measurements between the normalized shapes of these partials and the desired shapes (deduced from neighbouring non-overlapping partials).

We start with values coarsly distributed on the unit circle, and then refine on shorter intervals in the search for the global minimum. Figure 5 shows the partial shape distance for the demixing of the first overlapping partial. We see that this is in line with what we found in the histogram. The optimal demixing coefficients are found for delays of 2 and -1.

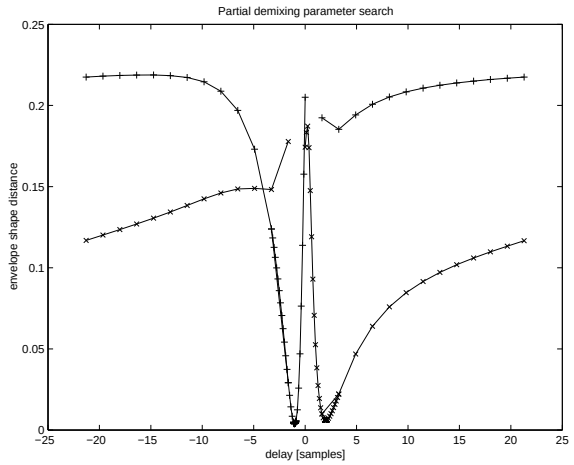

Figure 5: *Iteratively refined search for the demixing coefficients $\hat{H}_1$ and $\hat{H}_2$.*

Figure 6 shows the envelopes of the corresponding separated partials. In the top the shapes of the neighbouring non-overlapping partials are shown, the second harmonic of F_4, and the first harmonic of C_5. Both have quite constant amplitude over the duration of the signal. When the two notes are played together, these partials overlap, and there are beats, as seen in the second plot. The third plot shows how the DUET method performs for this case. Notice that the beats are still present. In addition the partial is exclusively assigned to either source in an alternating way. This introduces additional artefacts. In the bottom we see the envelope of the separated partials when using our partial separation algorithm.

The result when we separate the three first overlapping partials to improve the DUET separation can be seen in Fig. 7. Compare the separated partials (frequencies just above 1, 2 and 3 kHz) with those in Fig. 3.

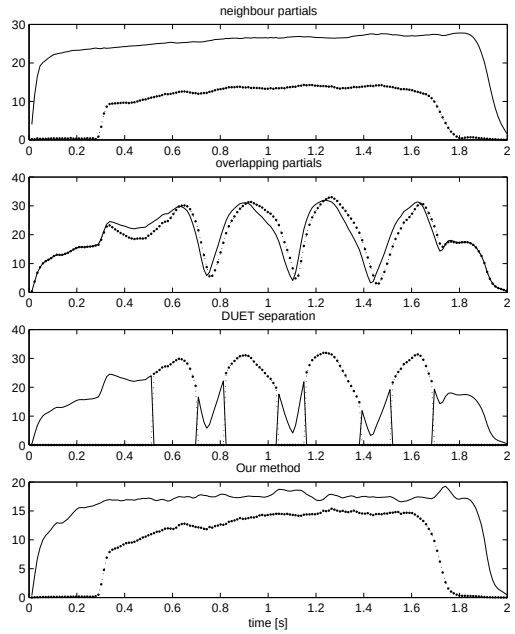

Figure 6: *Neighbouring non-overlapping partials (top), partials in mixed signals with beats (next-to-top), DUET separation (next-to-bottom), and separated partials using our method (bottom).*

3.2. Partial shape reconstruction

When one has only a mono signal to work with, clearly the partial separation method can not be used. The same applies for the case where there are more than two sources that overlaps in a given frequency band. We briefly discuss some other methods that can be used in these cases, even though they are not as good as the separation algorithm.

3.2.1. Partial removal

The first method is partial removal. When beats are detected, this can be avoided in the separated signals simply by removing the partial in question. This leads to a less rich sound, but can be less annoying than the beats. For just one partial, this may be an acceptable method, but when you remove more partials, the quality of the separated signals degrades very fast.

3.2.2. Partial flattening

If the overlapping partials never totally cancel each other, i.e. the valleys in the partials of the mixed signals are not too deep, we can scale up/ amplify the valleys to produce a flatter envelope. There are limits as to how much this can be done without introducing artefacts (phase problems). Also, the partial envelopes are only correct up to a scale factor.

3.2.3. Partial synthesizing

This method deduces the shape of the partial in question from the neighbouring non-overlapping partials, and then modulate this. Here one must be careful with the phase. If the original partial has

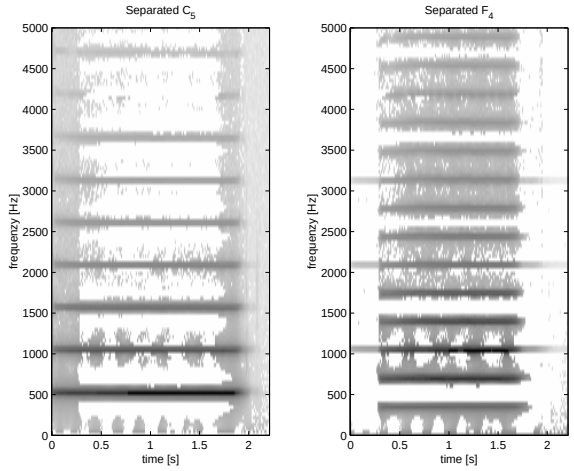

Figure 7: *Final result by improving DUET with our partial separation algorithm. The first three overlapping partials (just above 1, 2 and 3 kHz) have been separated. Compare with Fig. 3.*

frequency fluctuations (vibrato, etc), or if the frequency resolution is not fine enough, the synthesized partial may actually cause the full separated source to sound out of tune. Pitch tracking will not be useful to avoid this, since the overlapping partials interfere. There is also the scale factor uncertainty.

3.2.4. Partial splitting

The methods mentioned above all deal with beats. Basically, slow beats ($<$20 Hz) are the most annoying. For faster beats, and for beats masked by other frequency fluctuations (e.g. vibrato), it is sufficient to simply split the partial energy among the implied sources.

4. CONCLUSION

In this article we have presented a method to separate overlapping partials in stereo signals, and seen how this can be used to improve many of the existing source separation techniques. Other techniques for beat removal/ partial reconstruction that can work for other cases have also been discussed. We have applied our methods to mixtures of instruments and seen how it can be used to improve the perceptual quality of the separated signals. Sound examples can be found at http://lcavwww.epfl.ch/˜viste.

5. REFERENCES

[1] Brian C.J. Moore, *An introduction to the psychology of hearing*, Academic Press, 1997.

[2] R. Plomp and W.J.M. Levelt, "Tonal consonance and critical bandwidth," *Journal of the Acoustical Society of America*, vol. 38, pp. 548–560, 1965.

[3] E. Zwicker and H. Fastl, *Psychoacoustics, 2nd edition*, Springer, 1999.

[4] Robert J. McAulay and Thomas F. Quatieri, "Speech analysis/sysnthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[5] Tero Tolonen, "Methods for separation of harmonic sound sources using sinusoidal modeling," in *AES 106th Convention, Munich, Germany*, 1999.

[6] Matti Karjalainen and Tero Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing, Phoenix, Arizona, USA*, 1999.

[7] P. Fernandez-Cid and F.J. Casajus-Quiros, "Multi-pitch estimation for polyphonic musical signals," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, 1998.

[8] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing, Istanbul, Turkey*, 2000, pp. 765–768.

[9] A. Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing, Salt Lake City, Utah, USA*, 2001.

[10] K. Torkkola, "Blind separation for audio signals – are we there yet," in *Proceedings of International workshop on Independent Component Analysis and Blind Signal Separation, Aussois, France*, January 11-15 1999.

[11] Christian Jutten and Jeanny Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture," *Signal Processing, Elsevier*, vol. 24, pp. 1–10, 1991.

[12] Anthony J. Bell and Terrence J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[13] P. Comon, "Independant component analysis - a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[14] Anthony J. Bell Te-Won Lee and Russell H. Lambert, "Blind separation of delayed and convolved sources," *Advances in Neural Information Processing Systems, MIT Press*, 1997.

[15] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proc. IEEE Workshop on Neural Networks for Signal Processing, Kyoto, Japan*, September 4-6 1996, pp. 423–432.

[16] Paris Smaragdis, "Blind separation of convolved mixtures in the frequency domain," in *Int. Workshop on Independence and Artificial Neural Networks*, 1998.

[17] Shiro Ikeda and Noburu Murata, "A method of ICA in time-frequency domain," in *Proceedings of International workshop on Independent Component Analysis and Blind Signal Separation, Aussois, France*, 1999.

[18] Barry D. Van Veen and Kevin M. Buckley, "Beamforming - a versatile approach to spatial filtering," *IEEE Acoustics, Speech and Signal Processing Magazine*, pp. 4–24, 1988.

[19] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing, Istanbul, Turkey*, 2000.