BLIND SOURCE SEPARATION OF MONAURAL MUSICAL SIGNALS USING COMPLEX WAVELETS

José Ramón Beltrán Blázquez
Dpt. of Electronic Engineering and Communications,
University of Zaragoza
Zaragoza, Spain
jrbelbla@unizar.es

Jesús Ponce de León Vázquez
Dpt. of Electronic Engineering and Communications,
University of Zaragoza
Zaragoza, Spain
jponce@unizar.es

ABSTRACT

In this paper, a new method of blind source separation of monaural signals is presented. It is based on similarity criteria between envelopes and frequency trajectories of the components of the signal, and on its onset and offset times. The main difference with previous works is that in this paper, the input signal has been filtered using a flexible complex band pass filter bank that is a discrete version of the Complex Continuous Wavelet Transform (CCWT). Our main purpose is to show that the CCWT can be a powerful tool in blind separation, due to its strong coherence in both time and frequency domains. The presented separation algorithm is a first approximation to this important task. An example set of four synthetically mixed monaural signals have been analyzed by this method. The obtained results are promising.

1. INTRODUCTION

In the last years, *Blind Audio Source Separation* (BASS) has been receiving increasing attention. BASS tries to recover the source signals from their mixtures, when the mixing process is unknown. “Blind” means that very little information is needed to carry out the separation, although some assumptions are always necessary. Several techniques for solving the BASS problem have been developed, such as *Computational Auditory Scene Analysis* (CASA) [1], [2], *Independent Component Analysis* (ICA) [3] or *Sparse Decompositions* [4]. General methods for signal separation, like BASS, require multiple sensors. For example, in the stereo (DUET) separation, the delay information between the left and right channels can be used to detect the number of sources present in the mixture and some kind of *“scene”* situation [5]. But for other applications, a monaural solution is needed. Monaural separation is more difficult because we only have information of a single channel. But even in this case, the human auditory system itself can segregate the acoustic signal into separate streams, according to principles of auditory scene analysis (ASA) [6]. One of the most important applications is monaural speech enhancement and separation [7]. They are generally based on some analysis of speech or interference and subsequent speech amplification or noise reduction. Monaural BASS of musical signals have been developed in several ways [8], [9], [10].

Most of the authors [5], [8], [9], [10], use the STFT to analyze the mixed signal in order to obtain its main components or *partials*. In this work we have tried a different approach: the BASS of synthetically mixed signals using a complex band pass filtering of the signal.

A capital part of the separation process is based on some kind of statistical treatment of the available information. In order to find the real importance of this statistical process, we have avoided some of the usual limitations for this kind of separation, for example the non-overlapping spectra of the different sources. Although the purposed technique pretends to be as general as possible in the future, all the analyzed signals correspond to two different musical instruments synthetically mixed. The algorithm is not yet implemented in a frame-to-frame context, so we analyze the whole signal in a single step. This way it is possible to use the onset and offset time of each detected component of the mixed signal to obtain a first separation of sources. On the other hand, sometimes a single partial incorrectly separated can cause a non negligible error.

This paper is divided as follows: in Section 2 we have included a brief introduction to the CCWT, the interpretation of its results and the additive synthesis process. The proposed algorithm is presented in Section 3 and the experimental results are shown in Section 4. The main conclusions and actual and future lines of work are presented in Section 5.

2. COMPLEX CONTINUOUS WAVELET TRANSFORM AND ADDITIVE SYNTHESIS

The CCWT can be defined in several ways [11]. The most common is given by the expression:

$$W_x(a,b) = \int_{-\infty}^{+\infty} x(t)\Psi^*_{a,b}(t)dt \qquad (1)$$

where * is the complex conjugate and $\Psi_{a,\,b}\ (t)$ is the *mother* wavelet, or wavelet *atom*, frequency scaled by a factor *a* and temporally shifted by a factor *b* (both continuously varying):

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right) \qquad (2)$$

Therefore, the wavelet transform of a signal is equivalent to a band pass filtering of the signal. The convolution computes the wavelet transform using the dilated band pass filters. We have used the Morlet's wavelet as mother wavelet. In the frequency domain, Morlet's wavelet is a simple Gaussian filter:

$$\hat{\psi}_a(\omega) = C\, e^{-\sigma^2 \frac{(a\omega - \omega_0)^2}{2}} \tag{3}$$

In this equation, a is the scale parameter, C is a normalization constant and ω_0 is the central position (frequency or band / scale) of the filter.

We have developed a flexible filter bank *of constant Q* using these Gaussian filters. This work was partially presented in [12]. Once overlapped the handicap of developing the CCWT in an algorithmic (and so *discrete*) process, the result of filtering a signal through this band pass filter bank is a matrix of complex numbers (the wavelet coefficients), *c(a, t)*. These coefficients carry the information of the temporal evolution (envelopes) and frequency trajectories of each partial of the signal (see Figure 1). The coefficient matrix has a size *NxM* where *N* is the number of analysis bands (vertical axis) and *M* is the number of samples of the signal (horizontal axis).

One example of the modulus of *c(a, t)* can be seen in Figure 1, for a signal composed by the mixture of a guitar and a sax. The frequencies of the partials of the guitar are the oscillating trajectories, while the evolution of the sax is much less variable. These differences between partial trajectories will be used later to separate the original sources.

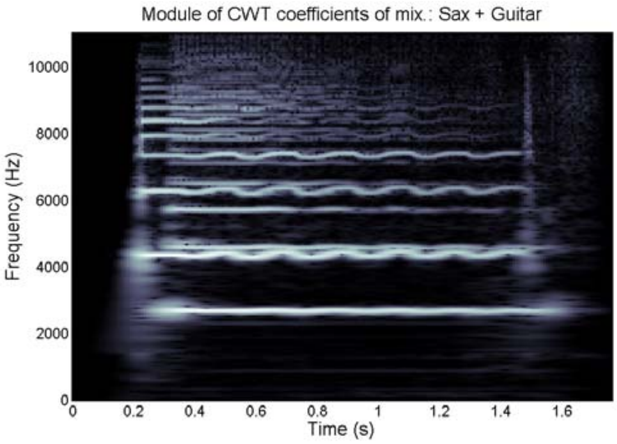

Figure 1: *The module of the CCWT coefficients of the signal obtained mixing the guitar and the sax signals.*

The next step is to obtain the scalogram of the signal. It is equivalent to the spectrogram of the Fourier analysis. It is obtained through the addition of the module of the wavelet coefficients in the temporal (abscissa) axis. Two examples of scalogram related to a flute and a clarinet are shown in Figure 2. Observe that some components are common to both signals (for example, bands 104-109) and therefore they will be indistinguishable in the mixed signal, whose scalogram is the sum.

Once the scalogram is obtained, we extract the partials of the signal. For each peak *i* of the function, we detect its associated upper and lower limit bands, $B_{sup}(i)$ and $B_{inf}(i)$, (by searching the closest minima). Then we perform the summation of the *complex* coefficients *c(a, t)* between these band limits. The result for the i^{th} peak is a complex valued function $P_i(t)$ whose modulus is the temporal envelope of the partial, and whose phase carries the instantaneous frequency information of the component.

$$P_i(t) = \sum_{n=B_{\inf}(i)}^{B_{\sup}(i)} c(b_{n(i)}, t) \tag{4}$$

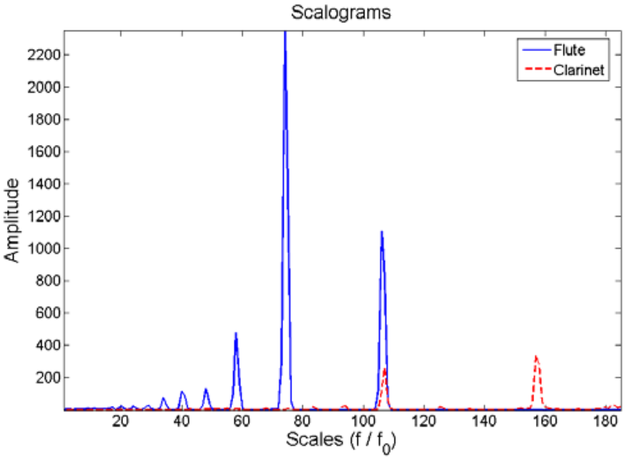

Figure 2: *The scalogram of the flute (continuous trace) and the clarinet (dotted line).*

Through a simple *additive synthesis* method, the original signal *x(t)* can be obtained performing the summation of these partials:

$$x(t) = \sum_{i=1}^{m} P_i(t) \tag{5}$$

The strength of this technique is that the obtained additive model of the signal results to be highly coherent in both time and frequency domains. This allows us to extract the instantaneous frequency of each component with remarkable precision, and to obtain the temporal error of the synthetic signal by means of the simple subtraction of the original and the synthetic waveforms [12].

Now the subject is how to use this information to somehow separate the mixed signal into its different sources.

3. ALGORITHM DESCRIPTION AND LIMITATIONS

As advanced in Section 1, for purposes of simplicity all the analyzed signals are synthetic mixtures of two different sources. This makes the separation process especially instructive.

First of all, the band pass filtering is performed, obtaining the wavelet coefficients matrix *c(a ,t)*, and the scalogram of the signal . Using this information, the partials of the signal are calculated.

A certain partial can be part of one of the sources, or can be shared by both of them. Examples of these three possibilities can be seen in the Figure 2. The goal is to find at least the partials clearly related to each source.

In order to separate the sources, some kind of criteria of *resemblance* between partials is necessary. This information can be found in the instantaneous frequency and the instantaneous envelope of the partials. The objective is to obtain some kind of *similarity pattern* between partials of the same source. This pattern may be obtained by evaluating how differs the envelope and instantaneous frequency evolution of each partial with respect to all other [8].

Different partials have a wide range of amplitude values. However, scaling each partial by its average, the resulting medium envelopes remain quite close each other. Due to the nature of sound, the information given by the instantaneous frequency does not oscillate as much as envelopes, and so it is more relevant. Figure 3 shows the envelope evolution and the instantaneous frequency of the three most important partials of the mixture of a flute and a clarinet. Note that the instantaneous frequency of these three partials evolves in a similar manner, despite the normalized amplitudes are not so similar.

Most of the signals present a high amplitude modulation (especially at high frequency) reaching sometimes values close to zero. In these points, there is no information enough to obtain the instantaneous frequency of the partial with accuracy, and it can present great oscillations (Figure 3). If the amount of sampled points with this problem is too high, the difference in frequency evolution can be a source of error. Fortunately this is not our case because we are dealing with signals with enough temporal duration.

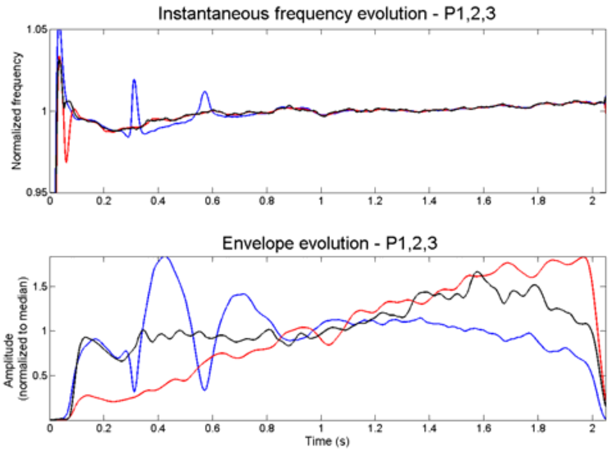

Figure 3: *Temporal evolution of the instantaneous frequency and the envelope (both normalized to their median value) of the first three partials of the clarinet plus flute signal.*

The distance between amplitudes and frequency trajectories (calculated over a time *Δt*) can be obtained from the mean square error [8]:

$$d_m(i,j)=\frac{1}{\Delta t+1}\sum_{t=0}^{\Delta t}\left(\frac{m_i(t)}{\widehat{m}_i}-\frac{m_j(t)}{\widehat{m}_j}\right)^2 \qquad (6)$$

$$d_f(i,j)=\frac{1}{\Delta t+1}\sum_{t=0}^{\Delta t}\left(\frac{f_i(t)}{\bar{f}_i}-\frac{f_j(t)}{\bar{f}_j}\right)^2 \qquad (7)$$

where scaling coefficients $\widehat{m}_i, \widehat{m}_j, \bar{f}_i$ and $\bar{f}_j$ are the average value of amplitude and frequency of the i^{th} and j^{th} partials, respectively, while $m_{i,j}(t)$ and $f_{i,j}(t)$ are the respective instantaneous amplitude and frequency of the same partials.

The global distance between partials is a weighted summation of both parameters:

$$d(i,j)=w_f d_f(i,j)+w_m d_m(i,j) \qquad (8)$$

where w_f and w_m are respectively the weights associated to the instantaneous frequency and the envelope mean square errors. In particular, the used *ad hoc* values for the weights are $w_f=0.9$ and $w_m=0.1$. The separation results do not depend significantly with respect to the exact value of these parameters.

Equation (8) is a distance between pairs. If the signal has *m* partials we obtain a square *mxm* matrix with all the information of distances between different trajectories, with zeros in its main diagonal (obviously).

The objective now is how to interpret this information to separate the partials corresponding to each source. The problem is that *a priori* we do not know the number of sources, and we only know the distance between partials, but not if a certain partial is part of a source or it is shared. Hence it is not easy to obtain a mapping where different partials from the same source (whose distance between pairs tends to zero) appear close each other. It is necessary to divide the partials into as many different families or categories as sources, having a minimum error between members of a class [8]. For signals with non-overlapping spectra, it is:

$$\min\left(\frac{1}{|S_1|}\sum_{i,j\in S_1}d(i,j)+\frac{1}{|S_2|}\sum_{k,l\in S_2}d(k,l)+\ldots\right) \qquad (9)$$

where S_1, S_2... are sets of partials and $|S_{1,2}|$ is the cardinality of each set. $S=S_1 U S_2$ is the whole set of founded partials and $S_1 \cap S_2=ø$.

The ideal solution would be to calculate all the possible permutations of Equation (9), and to choose the best one. But the number of possibilities to evaluate is too high, so we need a different approach.

A good first approach to the final solution uses the following hypothesis: The set of partials related to the same source must have *approximately* the same onset (and less important, offset) time. Different events [13] can create onsets in the audio signal. In our case, the instantaneous envelope of each partial contains the information about the onset and offset times. Before the first onset and after the offset, the amplitude is negligible. Between the onset and the offset times, the instantaneous envelope can vary in quite different shapes. In the same Figure 3, the three depicted partials are part of the same source, in this case the flute note. The amplitudes are dissimilar, but note that the onset and offset times remain quite close. Figure 4 shows the envelopes of the four most important partials of the (isolated) clarinet note. Although the shape of the envelopes can vary, the onset and offset times are similar.

We used a simple onset detection algorithm based on a more complex version presented in [14].

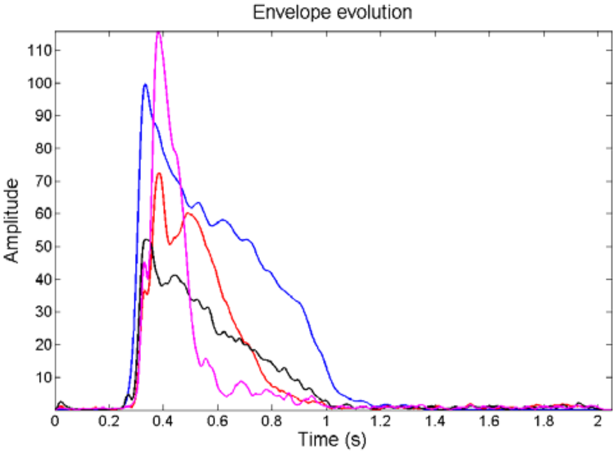

Figure 4: *Temporal evolution of the amplitude (no normalized) of the first four partials of the isolated clarinet signal.*

The onset detection tries to distinguish partials from different sources, assuming that a single instrument only produces a single event at a certain time instant. Unfortunately, neither the offset nor the onset time, are constant even for a single isolated musical instrument. For example in Figure 1, observe how the partials related to one of the sources (in this case, the sax signal) does not start *exactly* at the same time (the onset time increases with the median frequency of the partial). A similar situation is founded looking at the offset time. However, the partials of the guitar present clearly marked onset and offset times. So onsets and offsets are somehow signal-dependent information. It is not possible then to use *only* these two parameters to separate the different sources. However, they can be used to find a set of preliminary candidates for the different sources involved in the mix. We only need two or three partials from each source to develop a statistical search of the other partials related to the same source. The greater the energy of the partial, the better it defines the envelope and specially the instantaneous frequency evolution of its related source. The algorithm evaluates the onset and offset times of each partial, ranked accorded to energy, creating families of partials with similar characteristics. Each family is considered as a source. Then, taking the most energetic partials, we can evaluate the distances between each one and the rest of the partials of the signal through Equations (6), (7) and (8). In a last step, partials not included as part of any source can be proportionally divided in as much parts as sources, according for example to the mean distance to each family of partials.

Taking the upper and lower scales (or frequencies) that characterize each one of the m partials of a given source, s_k, we obtain a 2D mask for the source that can be used to synthesize the source directly from the wavelet coefficients, using:

$$s_k(t) = \sum_{i=1}^{m} P_i(t) = \sum_{i=1}^{m} M\left(a_{i,k}, t\right) .* c(a,t) \qquad (10)$$

where the operator **.*** must be interpreted as in *Matlab*®, and $a_{i,k}$ are the scales related to the i^{th} partial of the k^{th} source.

3.1. Limitations

The limitations of this technique result evident, and they will be detailed in the next Section. As a brief resume, we can found two main limitations. One of them is inherent to the analysis itself, the other is related with the analyzed signal.

As explained in Section 2, a partial is synthesized through the summation of complex wavelet coefficients inside the bands related to a certain peak in the scalogram of the signal. As the scalogram is calculated using the information of the whole signal, every partial has the same duration of the signal, independently of the real duration of the sound (see for example Figure 4). This makes that sources of shorter duration tend to carry some information of the longer duration sources that becomes more evident when the main sound decays.

On the other hand, if the separated signals present overlapping spectra, some partials are shared by different sources. In such a case, the separation process can fail.

4. EXPERIMENTAL RESULTS

The algorithm was tested with a set of four synthetically mixed signals from a short database of real musical instruments, chosen in order to obtain the limitations of this technique. Results are shown in Figures 5 to 8. In these figures five different waveforms are depicted: The first and the second are the original (isolated) signals. The third and the fourth are the separated sources, and the last waveform is the error signal, obtained as the temporal subtraction of one of the original signals and its related source. These figures should be considered a figure of merit of the results.

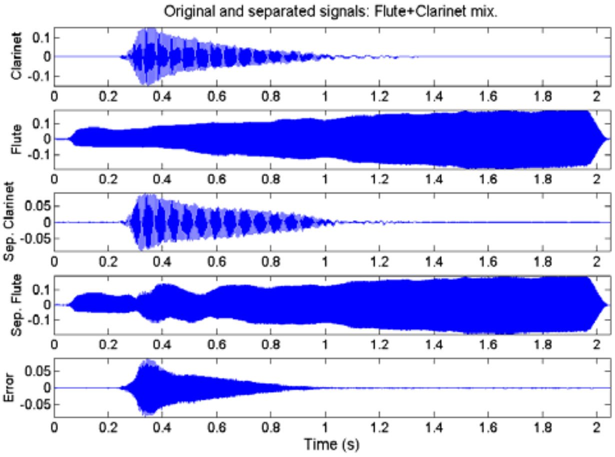

Figure 5: *Experimental results for the signal composed by the mixture of a flute and a clarinet.*

The first separation corresponds to the clarinet and the flute mix. It presents one of the limitations of the technique, introduced in Sections 2 and 3.1. In Figure 2 we presented the individual scalograms of these two signals. Observe that the second partial (in amplitude) of the clarinet (scalogram plotted in dotted line) is situated approximately in the same bands (frequencies) as the second partial of the flute. That means that these partials will merge into a single one, whose separation will be impossible, using the same filter bank. Results of the separation are shown in

Figure 5. Observe that the algorithm considered the merged partial as part of the flute note. This single partial represents approximately 40% of the energy of the clarinet note, as can be checked looking at the values of the first and third waveforms of the figure. This error results audible, and becomes more evident in the resynthesized flute signal.

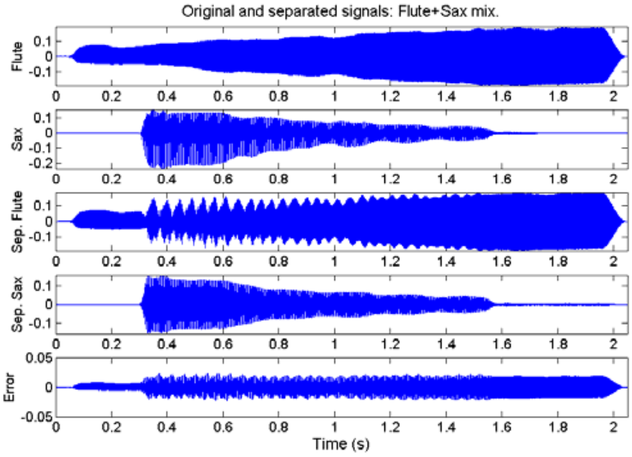

Figure 6: *Experimental results for the signal composed by the mixture of a flute and a sax.*

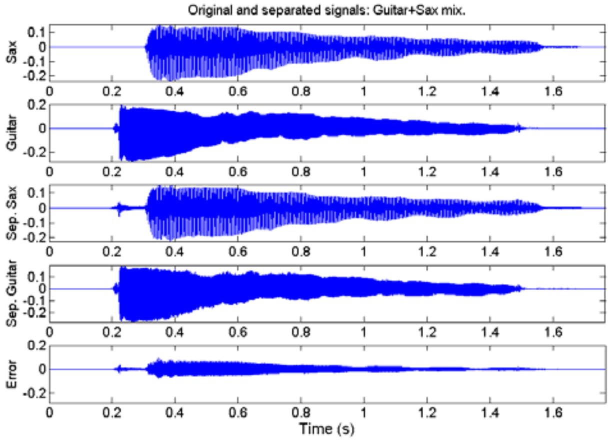

Figure 7: *Experimental results for the signal composed by the mixture of a guitar and a sax.*

As we explained briefly in Section 2, the bandpass filter bank that we generate is of constant *Q*. It means that in low frequency, filters are narrow while their bandwidth grows with frequency. So, partials in the high frequency are more difficult to separate as we can see in the flute and sax mix. Figure 6 shows the graphical results. In this signal, some high frequency shared partials were assigned to the flute, revealing one of the limitations of this technique. The final separation presents a clearly defined error signal and an audible high frequency sax distortion in the separated flute. Some possible solutions to this problem will be presented in Section 5.

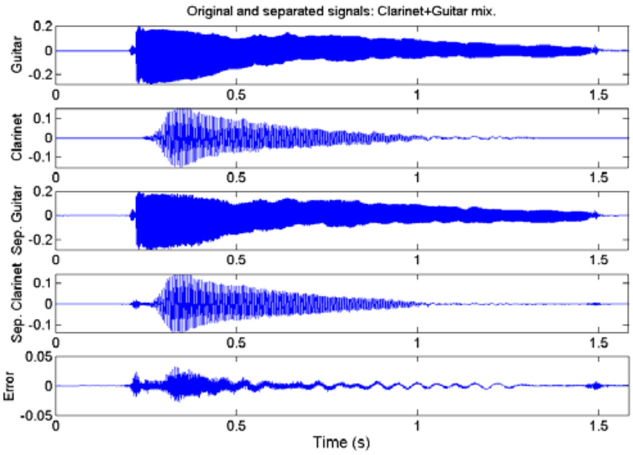

Figure 8: *Experimental results for the signal composed by the mixture of a clarinet and a guitar.*

In Figures 7 and 8, we show the results for the guitar-sax signal and the clarinet-guitar signal, respectively. In both signals, the separation process works almost perfectly. Only a little high energy information was not properly separated. The acoustic difference between the original signals and the separated sources is almost indistinguishable.

5. CONCLUSIONS

In this work a new technique of BASS of monaural signals is presented. The signal is analyzed through a complex band pass filter bank that comes from the CCWT.

The purposed method evaluates a distance between components of the mix signal. To make a first distinction of the number of present sources and some candidates to be components of a given source, we used an algorithm of onset and offset detection.

This technique has one main limitation: the sources have not to evolve similarly. That is, onset and/or offset times must be different. A musical piece with several instruments playing synchronously will be not properly separated.

The overlapping spectra of the signals, is another source of error. It can be smoothed using an appropriate (and application dependent) filter bank structure.

This algorithm is in an early stage. We are working in several lines, trying to develop a more complete separation technique. A first line of work combines a partial tracking algorithm, and a third distance term in Equation (8), related with the harmonic distance between partials. The onset detection is then not necessary, assuming that a single instrument can play a single note at a certain time (that is, polyphonic instruments are not compatible with this new separation technique).

On the other hand, we need to obtain standard numerical results of the quality separation process, in order to compare the proposed technique with previous works. It is also necessary to test the accuracy of the separation process using monaural not synthetic mixed signals. The necessity of non overlapping spectra of the involved sources is an important (and hard to solve) limitation of this task.

6. ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministry of Education and Science under the project TEC2006-13883-C04-01.

7. REFERENCES

[1] D. L. Wang and G. J. Brown, Eds., “Computational Auditory Scene Analysis: Principles, Algorithms, and Applications”, *Wiley, 2006*.

[2] G. J. Brown and M. Cooke “Computational auditory scene analysis,” *Computer speech & language (Print) 8:44, 297-336, Elsevier, 1994*.

[3] J. F. Cardoso, “Blind signal separation: Statistical principles,” *in Proceedings of the IEEE. October 1998, vol. 86, pp. 2009–2025, IEEE Computer Society Press*.

[4] M. G. Jafari, M. D. Abdallah, M. D. Plumbey, and M. E. Davies, “Sparse coding for convolutive blind audio source separation,” *in ICA 2006, Charleston, SC, USA, March 2006, pp. 132–139, Springer-Verlag.*

[5] Ö. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” IEEE Trans. On Signal Processing, Vol.52, No.7, July 2004.

[6] A. S. Bregman, “Auditory Scene Analysis”. *Cambridge, MA: MIT Press. 1990*.

[7] G. Hu and D. Wang “Monaural speech segregation based on pitch tracking and amplitude modulation,” *IEEE Trans. On Neural Networks, vol. 15, No.5, September 2004*.

[8] T. Virtanen, and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, *vol.2, no., pp.II765-II768 vol.2, 2000*.

[9] T. Virtanen “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Trans. On Audio, Speech and Lang. Processing, vol.15, No.3, March 2007.*

[10] G. Cauwenberghs, “Monaural separation of independent acoustical components,” *Circuits and Systems, 1999. ISCAS '99. Proceedings of the 1999 IEEE International Symposium on*, *vol.5, no., pp.62-65 vol.5, 1999*.

[11] I. Daubechies, “Ten Lectures on wavelets”, *vol. 61 of CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, 1992*.

[12] J. R. Beltrán and J. Ponce de León, “Analysis and Synthesis of Sounds through Complex Bandpass Filterbanks,” *Proc. of the 118th Convention of the Audio Engineering Society (AES’05). Preprint 6361, May. 2005*.

[13] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davis, and M. Sandler, “A Tutorial on Onset Detection in Music Signals,” *IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 1035–1047, Sep. 2005*.

[14] J. R. Beltrán, J. Ponce de León, N. Degara and A. Pena. "Localización de Onsets en Señales Musicales a través de Filtros Pasobanda Complejos," *in XXIII Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2008). Septiembre 2008*.