Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

Demixing Commercial Music Productions via Human-Assisted Time-Frequency Masking

MarC Vinyes[1], Jordi Bonada[1], Alex Loscos[1]

[1] *Pompeu Fabra University, Audiovisual Institute, Music Technology Group, Barcelona, 08003, Spain*

Correspondence should be addressed to MarC (mvinyes@iua.upf.edu)

ABSTRACT

Audio Blind Separation in real commercial music recordings is still an open problem. In the last few years some techniques have provided interesting results. This article presents a human-assisted selection of the DFT coefficients for the Time-Frequency Masking demixing technique. The DFT coefficients are grouped by adjacent pan, inter-channel phase difference, magnitude and magnitude-variance with a real-time interactive graphical interface. Results prove an implementation of such technique can be used to demix tracks from nowadays commercial songs.
Sample sounds can be found at http://www.iua.upf.es/~mvinyes/abs/demos.

1. INTRODUCTION

1.1. What do we mean by Audio Blind Separation?

We understand ABS (Audio Blind Separation) as extracting from an input audio signal, without additional information, a set of audio signals whose mix is perceived similarly[1] *to the original audio signal*[2].

This problem has infinite solutions, however, a human being would only find a reduced set of those solutions meaningful. These are the ones that we would like to extract.

1.2. Demixing tracks

Commercial music is often produced using a set of recorded mono or stereo audio tracks which are afterwards mixed instantaneously. Using an analog mixer or a digital audio workstation, audio tracks are usually processed separately in groups with specific pan, equalization, reverb and other digital or analog effects. With this procedure, the sound engineer

[1] We assume that when comparing a pair of sounds where one is a moderately equalized and compressed version of the other, their perceptual similarity will be very high.

[2] Note that we don't stick to the Audio Blind Source Separation definition, in which the original signal is to be exactly equal to the sum of the extracted signals. Instead we suggest a perceptual interpretation of this equality, which is more general.

often pursues to favor their perception as different *auditory streams* (see [1] for better understanding of this term). Because of this reason, when we listen to their mix, we perceive separately these audio tracks most of the times and, consequently, we find them meaningful.

On the other hand, if we were able to extract audio signals that are perceived similarly to these audio tracks, they would be a solution of the ABS problem because their remix would also be perceived similarly to the original mixture. *Therefore, in this article, our solution of the ABS problem pursues the extraction of audio signals which are perceived similarly to the audio tracks used to produce the mix.*

1.3. Notation

We are considering stereo songs as inputs. We use L to label variables related to the left channel and R for the right channel. We will refer to the two channels of the mixture as $out_L[k]$, $out_R[k]$, the stereo tracks whose instantaneously mix produces the mixture will be labeled $in_i^R[k]$, $in_i^L[k]$ and $s_i^R[k]$, $s_i^L[k]$ will denote the extracted stereo sounds. Hence,

$$\begin{pmatrix} out_L[k] \\ out_R[k] \end{pmatrix} = \begin{pmatrix} \sum_i in_i^L[k] \\ \sum_i in_i^R[k] \end{pmatrix}$$

And we pursue,

$$\begin{pmatrix} s_i^L[k] \\ s_i^R[k] \end{pmatrix} \simeq \begin{pmatrix} in_i^L[k] \\ in_i^R[k] \end{pmatrix}$$

1.4. Algorithm overview

Our algorithm uses TFM (Time-Frequency Masking) as a mechanism to generate candidate solutions of the ABS problem from the input data (section 2). Some criteria are developed to choose only the ones that are real meaningful solutions of the ABS problem. As we discussed previously, signals that are likely to match the tracks used in the production of the song will often be meaningful solutions of the ABS problem, so a set of mathematical characterizations of the sound of a track are presented in section 3 and used in section 4 to choose between the candidate solutions generated by TFM. Finally two additional selection procedures not based on these characterizations are presented.

2. GENERATION OF CANDIDATE SOLUTIONS OF THE ABS PROBLEM

We generate candidate solutions of the ABS problem as follows:

1. First we take a set of P overlapped time frames of size N from the mixture:

   $out_L[0] \cdots out_L[N-1]$
   $out_R[0] \cdots out_R[N-1]$
   $\cdots$
   $out_L[(P-1) \cdot M] \cdots out_L[(P-1) \cdot M + N - 1]$
   $out_R[(P-1) \cdot M] \cdots out_R[(P-1) \cdot M + N - 1]$

   Frames will have an overlap of $(N-M)$ samples.

2. Each frame is windowed in order to avoid spectral leaking and the DFT (Discrete Fourier Transform). Because the input signal is real, the DFT frame has hermitian symmetry, therefore all the information is stored in the first $N/2+1$ coefficients. We will refer to their values as:

   $DFT_0(out_L)[0] \cdots DFT_0(out_L)[N/2]$
   $DFT_0(out_R)[0] \cdots DFT_0(out_R)[N/2]$
   $\cdots$
   $DFT_{P-1}(out_L)[0] \cdots DFT_{P-1}(out_L)[N/2]$
   $DFT_{P-1}(out_R)[0] \cdots DFT_{P-1}(out_R)[N/2]$

3. Next, the DFT frames of $s_i^R[k]$,$s_i^L[k]$ are built keeping some the obtained DFT coefficients and setting the others to 0. In other words, we apply a binary mask to the DFT coefficients. Let $p \in 0..P-1$,

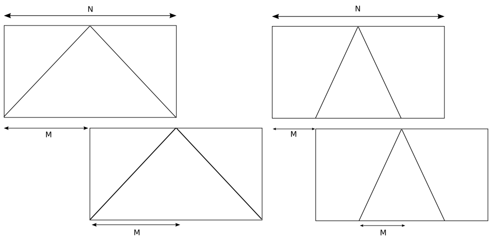
   $$DFT_p(s_i^L)[f] = \begin{cases} 0 \\ DFT_p(out^L[f]) \end{cases}$$

   $$DFT_p(s_i^R)[f] = \begin{cases} 0 \\ DFT_p(out^R[f]) \end{cases}$$

   This is the step where different parameters can be chosen to generate different sounds. For each frame and each DFT coefficient, we can choose whether to set it to zero or keep its value. Consequently, a huge family of candidate solutions can be generated: up to $2^{(N/2+1) \cdot P}$ different sounds.

4. The IDFT (Inverse Discrete Fourier Transform) of these frames is performed after filling their coefficients from $N/2+2$ to $N-1$ with values that force the hermitian symmetry (the IDFT output must be a real signal). Next, we multiply it by the inverse of the window values used before computing the DFT.

5. Finally we overlap and add those frames to obtain $s_i^R[k]$,$s_i^L[k]$. The overlap is performed using a triangular window placed around the center and padded with zeros when $M < N/2$. Two examples are displayed in figures 1(a),1(b).



(a) M=N/2 (b) M=N/4

Fig. 1: Overlap and add with a triangular window with different values of M

Some of the articles ([2]) that introduced the idea of applying a binary mask to the DFT coefficients referred to the derived processing mechanism as Time-Frequency Masking. We will continue to use this term.

We don't know how many meaningful solutions of the ABS problem are included in the space of candidate solutions produced by TFM. Moreover, this could vary among different input data. However some experiments reveal that for many real commercial music productions at least some meaningful solutions are included. In this article we present some examples of songs where our algorithm finds meaningful solutions (see section 5). Additionally, we have built a web site ([3]) whose forum collects reports of successful audio blind separation using this technique. For speech signals, ([2]) shows that the mixed voices are usually recovered.

In fact, if the original tracks had non-overlapping nonzero DFT coefficients, we would be able to generate them perfectly with TFM of the mixture (assuming loss-less DFT-IDFT). Unfortunately, most of the music sounds don't verify this first hypothesis. However it seems that a track is perceived similarly when some of its DFT coefficients are replaced by their corresponding DFT coefficients of the mix, where more than a track may overlap. This may be explained by the experience that moderate equalization and compression (that may occur in frequency bands where two tracks overlap) don't alter significantly our perception of a sound.

On the other hand, we can think of cases where tracks will be hard to extract. In particular, when two tracks contain two performances of the same instrument playing the same music (often some vocals and some guitars are recorded twice), then both tracks will undoubtedly overlap in frequency. However, in such cases, we are not usually able to perceptually distinguish both tracks either, so we will only pursue the extraction of their mixture.

3. MATHEMATICAL CHARACTERIZATIONS OF THE SOUND OF A TRACK

3.1. Pan

The fact that some *mono*[1] tracks are mirrored in the two stereo channels when they are panned in the mixing process is helpful to decide whether a sound may correspond to a track or it doesn't.

Let $in_i[k]$ be the original mono tracks of one mixture, mixers mirror them in two channels as follows:

$$\begin{pmatrix} in_i^L[k] \\ in_i^R[k] \end{pmatrix} = \begin{pmatrix} \alpha_i^L \cdot in_i[k] \\ \alpha_i^R \cdot in_i[k] \end{pmatrix}$$

Therefore, the mixture can be obtained with the following expression,

$$\begin{pmatrix} out_L[k] \\ out_R[k] \end{pmatrix} = \begin{pmatrix} \alpha_1^L & \dots & \alpha_n^L \\ \alpha_1^R & \dots & \alpha_n^R \end{pmatrix} \cdot \begin{pmatrix} in_1[k] \\ \vdots \\ in_n[k] \end{pmatrix}$$

[1] Note that when a stereo reverberation effect is applied to one mono track, the output of the process is stereo, so these tracks are not considered here.

Additionally, we found that the majority of analog and digital mixers use the following pan law, where $x \in [0, 1]$ is the value of the analog or digital pan knob:

$$\begin{cases} \alpha_i^L = \cos\left(x \cdot \pi/2\right) = \sqrt{\frac{1}{1+(\alpha_i^R/\alpha_i^L)^2}} \\ \alpha_i^R = \sin\left(x \cdot \pi/2\right) = \sqrt{\frac{(\alpha_i^R/\alpha_i^L)^2}{1+(\alpha_i^R/\alpha_i^L)^2}} \\ x = \arctan\left(\frac{\alpha_i^R}{\alpha_i^L}\right) \cdot 2/\pi \end{cases} \tag{1}$$

Hence if the extracted sound $s_i^L[k]$,$s_i^R[k]$ is one of the original stereo tracks, it will verify (assuming $in_i[k] \neq 0$):

$$\frac{s_i^R[k]}{s_i^L[k]} = \frac{in_i^R[k]}{in_i^L[k]} = \frac{\alpha_i^R}{\alpha_i^L} = constant$$

Moreover, the DFT coefficients of both channels will still verify this expression because the DFT is a linear transformation. Consequently, our requirement for the extracted stereo sounds $s_i^L[k], s_i^R[k]$ may be extended as follows to their DFT coefficients:

$$\frac{DFT_p(s_i^R)[f]}{DFT_p(s_i^L)[f]} = constant \ \forall f \in 0 \ldots N/2 \tag{2}$$

$$\text{if } DFT_p(s_i^R)[f] \neq 0 \text{ or } DFT_p(s_i^L)[f] \neq 0$$

It is worthwhile to mention that this mechanism is particularly good because it allows the discrimination between mono tracks that have been panned with different $\frac{\alpha_i^R}{\alpha_i^L}$ ratio. On the other side of the coin, there are many kinds of tracks whose characterization with this procedure is not valid:

- Stereo tracks
- Mono tracks with artificial stereo reverberation
- Mono tracks with "automated" pan

3.2. Inter-channel Phase Difference

Another consequence of mirroring mono tracks in the two stereo channels is that their DFT phase spectrum will be the same for both channels.

Hence if the extracted sounds are the original tracks, they must verify:

$$|Arg(DFT_p(s_i^L)[f]) - Arg(DFT_p(s_i^R)[f])| = 0$$
$$\forall f \in 0 \ldots N/2$$

And mono tracks with artificial stereo reverberation or stereo tracks may be generally characterized by the opposite case:

$$|Arg(DFT_p(s_i^L)[f]) - Arg(DFT_p(s_i^R)[f])| > 0$$
$$\forall f \in 0 \ldots N/2$$

In this case, it may happen that the track not only contains DFT coefficients with different phase but also some with the same phase; in that case some kind of dereverberation is performed.

IPD (Inter-channel Phase Difference) is good because all mono tracks are well characterized (using either of the two complementary formulas), however it only will allow us to distinguish between mono-non-reverberated and mono-artificially-reverberated/stereo tracks.

On the other hand, the zero phase difference can be used as a prerequisite of the pan characterization because the latter presupposes that the same sound is mirrored in both channels.

4. SOLUTION SELECTION

In this section we will describe how to select the DFT coefficients that aren't set to zero in step 3 of the candidate solutions generation process. This mechanism should allow us to output one sound that is a meaningful solution of the ABS problem out of all the possible sounds generated with TFM.

We designed it to be build in a real-time application which could receive feedback from the user. That is why we named it "Human-Assisted TFM". In order to set the parameters of our algorithm, we use an interface similar to the one presented by the authors of [4]. The user is able to set a few parameters that determine the DFT coefficients that are masked to obtain sounds that are perceived similarly to the audio original tracks of the song.

The selection process is performed applying independent processing layers that we will call "Time-Frequency Filters". In each one, a different Time-Frequency binary mask is set following a different approach. The overall algorithm applies a Time-Frequency mask that is the union of the previous masks in each frame.

First two TFF (Time-Frequency Filters) based on the mathematical characterizations of the sound of a track are presented. Then, we suggest two auxiliary post-processing TFFs that may help to discriminate between some specific sounds.

4.1. Pan TFF

It is obvious that when two tracks overlap in the same DFT coefficient of frequency f, we can't expect the ratio:

$$\frac{DFT_p(out_R)[f]}{DFT_p(out_L)[f]} \tag{3}$$

to be any of the ratios of the two tracks. Moreover, the overlapping coefficients may change in time. Hence, the DFT coefficients can't be assigned to different tracks directly using the expression in 2.

However, when dealing with non-reverberated speech signals, such overlap is not very significant, and those changes in 3 only vary slightly from a value. Therefore it might make sense to define a maximum likelihood estimator of the value 3 of each track in order to select them. Although that was the approach of [2] with speech signals, in commercial music production signals the overlap is much more significant and, at the moment, some researchers select ranges of values where 3 may vary without having a clear peak. [5] defines a Gaussian window and the authors of [4] manually select a range of pan.

Our Pan TFF is based on the manual selection approach of [4], which is improved adding a mapping of the values of 3 to their corresponding estimated pan (we use the expressions set in 1). In this way their values are displayed with the same mapping the sound engineer used to define the pan. Additionally the resulting interval $[0, 1]$ of possible values is bounded while expression 3 took values in $[0, \infty)$.

Next we assist the user of our application with a visual representation of the energy of the DFT coefficients in each pan (see subsection 4.4). The range of pans that the DFT coefficients should have is then selected. If the DFT coefficient has an estimated pan out of this range, it will be set to 0. Otherwise it will be kept as it is.

4.2. Inter-channel Phase Difference TFF

For the same reasons stated above it is not possible to clearly distinguish between exactly zero and non-zero IPD (Inter-channel Phase Difference) and the characterization of section 3.2 should be adapted to work with overlapping DFT coefficients. Therefore, we define a TFF where a threshold is set to limit both situations. We assist the user of our application with a visual representation of the energy of the DFT coefficients with each possible IPD values (between $-\pi$ and π) and the threshold is defined by the user looking at this graph.

This graph might be also used to decide whether the pan discrimination is going to work. If the DFT coefficients have the same phase (all the energy is accumulated around the zero IPD value), then it makes sense to suppose that they were mono tracks mirrored in both channels with a constant ratio, otherwise some artificial reverberation may have been added and the previous criterion may be useless.

4.3. Magnitude and Magnitude-Variance TFFs

Those filters are not based on characterizations of the sound of a track. Hence we believe that they should be used to post-process the results obtained with the previously introduced TFFs.

The Magnitude TFF, first normalizes all the DFT coefficient' magnitudes using their maximum value in the current frame. A set of magnitudes between 0 and 1 are obtained. Next, ranges within those values are selected in a graph that represents their accumulated energy across multiple frames. We found this criterion useful to distinguish between percussive sounds with large and flat frequency spectrum (eg. snares and crashes) and harmonic instruments that have a spectrum with peaks of frequency components (eg. voice or guitars).

The Magnitude-Variance TFF computes for each DFT coefficient of each channel, its magnitude variance along time. Then all variances are normalized by their maximum value and their energy is displayed in a graph. A range of this variances are also selected manually. This TFF is useful to discriminate between brief and steady attacks or sounds with different magnitude variation along time.

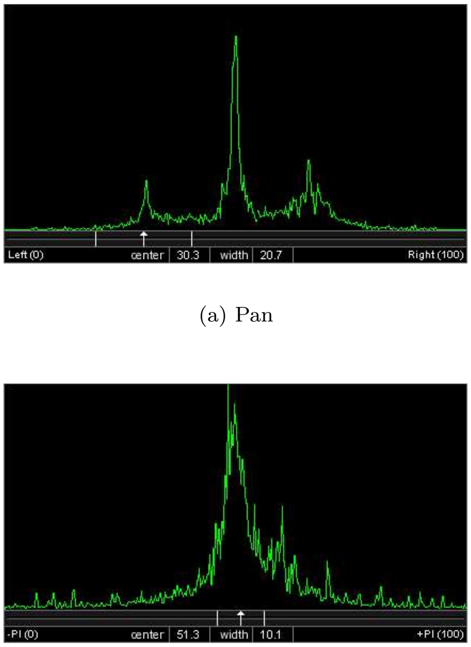
Note that these two procedures lead to Time-Frequency masks which may be different for each stereo channel. An application of these TFFs is included in example 3 of section 5.

4.4. Visual representations

All graphs are built exactly in the same way. For each DFT coefficient we compute one attribute (Pan, IPD, Magnitude or Magnitude-Variance). Then this attribute is mapped in a closed interval (generally [0,1] or [$-\pi$,π] for IPD).

This interval is divided in a finite number of small bins and for each bin we add the square modulus contributions of the DFT coefficients whose attribute value corresponds to it. Finally we average the values among several frames.

The values of the graph are normalized by their maximum value along some frames in order to achieve an optimum vertical resolution without rescaling the graph too frequently. Two examples are displayed in figures 2(a) and 2(b).



(a) Pan

(b) IPD

Fig. 2: Pan and IPD energy distribution graphs

5. EXPERIMENTS

5.1. Parameters

In our experiments we tested several values of N and we found that values above 8192 didn't improve the perceived quality of the output sound. M is set to N/4 for better quality of the sound transitions. Finally, we selected the Blackmann-Harris -92dB window in order to minimize the side-lobes which make the DFT frequency coefficients merge and change the estimation of their pan, IPD or Magnitude.

The DFT is performed with the Fast Fourier Transform algorithm and the graphs are computed with an horizontal resolution of 300 bins, average of 20 frames, and a normalization along 40 frames.

Next we will discuss some examples. Although, some waveforms and spectrograms are drawn, the reader is encouraged to download and listen to their audio at `http://www.iua.upf.es/~mvinyes/abs/demos`, where an evaluation version of our algorithm implementation is also available.

5.2. Example 1

We first tested our algorithm over a self-produced song whose tracks were available. The song was produced with synthetic sampled drums in one mono track and 3 stereo real guitar tracks (acoustic rhythm guitar, feedback-delayed acoustic guitar pickings and electric distortion guitar). The feedback-delayed acoustic guitar and the electric distortion guitar are panned to opposite sides and both drums and the rhythm guitar are panned to the center.

Using pan discrimination we were able to extract the original tracks of the guitars panned at both sides and one track consisting of the mixture of the drums and the rhythm guitar (because they shared the same pan). In figures 4(a) and 4(b) the original waveforms and spectrograms of the feedback-delayed acoustic guitar and its corresponding extracted sound are displayed together.

Note that the recovered track has a similar waveform although the amplitude varies. However when we listen to them, we perceive them to be highly similar. This example supports our decision of pursuing tracks that are *perceived similarly* to the original ones.

5.3. Example 2: Help (The Beatles)

Next we present an extraction of the vocals of the popular song Help (The Beatles). In this example

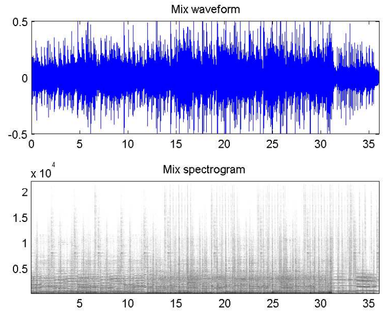
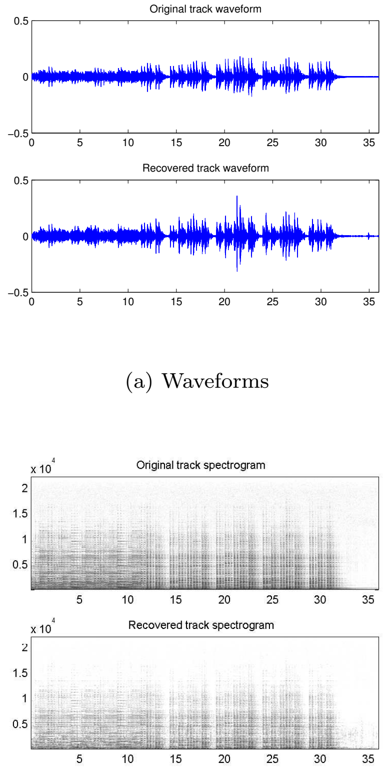


Fig. 3: S03: Mix waveform and spectrogram

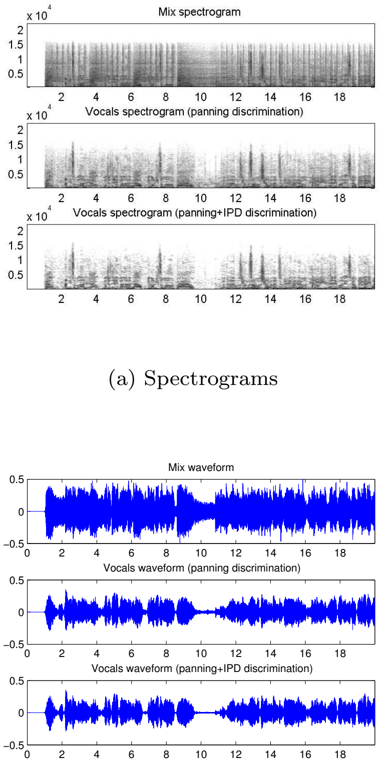


(a) Waveforms

(b) Spectrograms

Fig. 4: S03: Original vs recovered guitar track

the IPD TFF improves the separation when it is placed before the pan TFF.



(a) Spectrograms

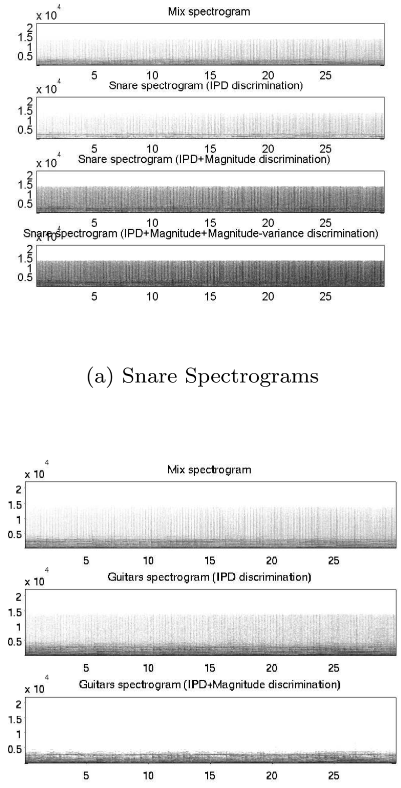
(b) Waveforms

Fig. 5: Help: Mix and extracted vocals

The IPD TFF filters many of the drums residual noise that remains present when the pan TFF is used. We guess that drums are discriminated by IPD because they were recorded in stereo tracks, while the vocals were recorded using a mono mic.

5.4. Example 3: Memorial (Explosions in the Sky)

This is an example of snare and guitar extraction. In this song, only guitars seem to be highly reverberated so a IPD TFF helps to discriminate between them and the other instruments. So we select a margin around the center (zero phase) to begin the extraction of the snare and we use its complementary range to extract the guitars. Next, in both cases we use the Magnitude TFF. In order to isolate the guitars we remove the DFT coefficients with small magnitude and we do the opposite for the snare. Finally the Magnitude-Variance TFF is applied to reduce the noise produced by the guitar pickings in our attempt to isolate the snare (those noises have greater magnitude variance than the sound of the snare).



(a) Snare Spectrograms

(b) Guitars Spectrograms

Fig. 6: Memorial: Extracted snare and guitars

6. CONCLUSIONS

Our work can be included in the group of techniques that obtain the solution of the ABS problem extracting sets of the input Discrete Fourier Transform (DFT) coefficients ([2],[4],[5],[6]). In particular we design a human-assisted mechanism as in [4] to select those sets. A graphic interface is developed to select them using their inter-channel magnitude ratio and inter-channel phase difference, with some post processing involving the magnitude of the DFT coefficients and their magnitude variance. The processing is divided in several independent layers called TFF that set a Time-Frequency binary mask in multiple steps with different criteria. We claim that this a more flexible way to select the Time-Frequency Mask than the ones developed in the previously cited articles.

Another contribution is to consider inter-channel phase difference instead of estimated time delay as introduced in [2]. [7] points out ambiguities in the estimated time delay over frequencies over 900Hz and suggests a method to resolve them in mixtures of acoustically stereo recorded sound sources. However, in commercial music productions, IPD seems to make more sense and may help to distinguish between mono tracks and stereo tracks based. Moreover, its energy distribution graph may be useful to determine whether the pan discrimination criterion is going to work and embed TFM with pan selection in automatic ABS systems.

We have shown that, in some cases, it is possible to extract from real commercial music productions the original tracks that were used in their mixing. It is often the case that vocals and other music instruments are recorded in different audio tracks. Consequently, our algorithm can be useful in karaoke systems to remove the voice of some songs, and it can be applied as a DJ remixing tool or isolating instruments in a mixture. In [3] visitors are encouraged to download our software and post successful audio separations in order to evaluate its real-world performance.

7. FUTURE WORK

This method trusts TFM as a successful way to obtain all the meaningful solutions of the ABS problem. Future work may consist of evaluating experimentally the limits of this technique independently of the chosen TFFs.

Additionally we have only presented 4 TFFs (Pan, IPD, Magnitude and Magnitude-Variance), so more ways to set the TFM mask may be explored.

Finally we have noticed low frequencies have big magnitudes that alter significantly the graph even if they are not really important to our ears. Perceptual weighting might help to make the graphs better represent what we really hear and ease the selection of meaningful sounds.

8. REFERENCES

[1] Albert S. Bregman. *Auditory Scene Analysis.* MIT Press, 1990.

[2] Özgür Yilmazz and Scott Rickard. Blind separation of speech mixtures via time- frequency masking. *IEEE Transactions on Signal Processing*, 2003.

[3] MarC Vinyes, Alex Loscos, and Jordi Bonada. `http://www.iua.upf.edu/mtg/audioscanner`.

[4] Dan Barry, Bob Lawlor, and Eugene Coyle. Sound source separation: Azimuth discrimination and resynthesis. *Proc. of the 7th Int.Conference on Digital Audio Effects (DAFX 04)*, 2004.

[5] Carlos Avendano. Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications. *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2004.

[6] Michael A. Casey and Alex Westner. Separation of mixed audio sources by independent subspace analysis. *International Computer Music Conference (ICMC)*, 2000.

[7] Harald Viste and Gianpaolo Evangelista. On the use of spatial cues to improve binaural source separation. In *Proc. of the 6th Int. Conference on Digital Audio Effects (DAFX-03)*, 2003.