Separation of Music+Effects sound track from several international versions of the same movie

Antoine Liutkus, Pierre Leveau

Audio Engineering Society

Convention Paper

Presented at the 128th Convention
2010 May 22–25 London, UK

Separation of Music+Effects sound track from several international versions of the same movie

Antoine Liutkus*[1], Pierre Leveau[2]

[1] *Telecom ParisTech, 37/39 rue Dareau 75014 Paris, France*

[2] *Audionamix, 114 avenue de Flandre 75019 Paris, France*

Correspondence should be addressed to Antoine Liutkus, Pierre Leveau
(antoine.liutkus@telecom-paristech.fr, pierre.leveau@audionamix.com)

ABSTRACT

This article concerns the separation of the music+effects (ME) track from a movie sountrack, given the observation of several international versions of the same movie. The approach chosen is strongly inspired from existing stereo audio source separation and especially from spatial filtering algorithms such as DUET that can extract a constant panned source from a mixture very efficiently. The problem is indeed similar for we aim here at separating the ME track, which is the common background of all international versions of the movie soundtrack. The algorithm has been adapted to a number of channels greater than 2. Preprocessing techniques have also been proposed to adapt the algorithm to realistic cases. The performances of the algorithm have been evaluated on realistic and synthetic cases.

1. INTRODUCTION

The movie and broadcast industries are interested in producing content available in as many languages as possible. For that purpose, they generally have at hand separated tracks for music and effects (ME) on one side, and voices on the other side. Creating the version of the content adapted to a language consists in mixing the voices track of the correct language with a generic ME track.

Sometimes, the music and effects track is not available. In that case, releasing a new version of the content in another language is not possible, unless

*Antoine Liutkus performed this study when he was working at Audionamix.

this track is recorded another time.

In this paper, a method to extract the ME and voices track using several international versions of the content is proposed. It addresses the case when several localized versions of the content exist with a good quality, while the separated tracks (voices, and music and effects) are not available. For example, given a French, an American and a Japanese version of the mixed content, it enables to extract the music and effects track and the voice track for each version.

Given a localized version, this problem can be restated as a voice extraction problem. This problem has been adressed using a number of approaches, first with sinusoidal model [1] or more recently, using a source-filter model in a Non-Negative Matrix Factorization context [2]. Both of these models were extended recently in the source-separation community to allow voice extractions within mixtures. Unfortunately, most of the work achieved in this direction so far relies on parametric models whose optimization requires very heavy computational ressources, which is prohibitive for the desired applications that ought to perform well and rapidly on very limited computing devices.

The method relies on methods inspired from stereo audio source separation, like DUET [3]. In this class of source separation methods, each channel is represented in the time-frequency domain, then each time-frequency point is allocated to a source according to the value of its left and right channel components. In this stereo case, the panpot parameter (angle between 0 and 90 degrees) can be computed from the amplitudes of the time-frequency point on each channel. The time-frequency point is then allocated to the source that has the closest angle. For example, if one wants to extract the signal of the voice in a song, the angle of the source to extract is set to 45 degrees. The voice separation is here equivalent to the extraction of the signal which is common between the two channels.

In our case, the situation is similar: the ME track is a signal which is common between several signals (assumed mono), the international versions. The proposed method extends the principle of the spatial source separation that have more than two channels, and takes into account the sparsity of the voice signals in the time-frequency domain. The use of simple statistical filtering is tried to extract the common signal. To address the realistic cases, track synchronization and filtering are also performed.

In Section 2, the problem of extracting a common ME track from several international version is detailed. Then in Section 3, the algorithm to address this problem is described. Section 4 depicts the results of the algorithm on synthetic and realistic situations.

2. PROBLEM STATEMENT

2.1. Naive modelling

The signals considered are movie soundtracks, that consist in the superposition of a ME track and a voice track. Let us call these signals *mixtures* and furthermore assume that N international versions of the same movie are available. Let $m_i(t)$ be the soundtrack of the international version i. We have :

$$\forall i \in [1:N], m_i(t) = b_i(t) + v_i(t) \qquad (1)$$

With $b_i(t)$ and $v_i(t)$ being the ME (background) and voice track respectively for international version i.

The main difference between different international versions of a movie soundtrack is the voice track. If the ME tracks are strictly identical on all versions ($\forall i \in [1:N], b_i(t) = b(t)$), a first idealized model is:

$$\forall i \in [1:N], M_i(t) = b(t) + v_i(t) \qquad (2)$$

This model can be cast into the spectral domain using a Short Time Fourier Transform (STFT) $\mathcal{F}$:

$$\forall i \in [1:N], \mathcal{F}\{m_i\} = \mathcal{F}\{b\} + \mathcal{F}\{v_i\} \qquad (3)$$

Still, even if it may be adequate in some cases, experience shows that model (2) is very unrealistic most of the time, for two main reasons :

- It is not realistic to consider that the ME signals are exactly synchronized among the localized versions. We have observed delays up to almost half a second in real international versions sold on DVDs on the market.

- The different international versions may come from different mixing processes, possibly even involving different mixing hardware or artistic choices such as dynamics or equalization.

Hence, the naive model (2) is most likely to be wrong in real world usecases. However, it will still be useful in this study as an idealized case and most of our effort on real data will be spent processing it so that it may be compatible with the naive model.

2.2. A more realistic model

As has been discussed in the previous section, it is not realistic to consider that the several localized versions of a movie share exactly the same ME track. Real international versions are notably not perfectly synchronized and are mixed differently with the voice tracks. Their level and spectral equalizations may vary between the versions. In this study, non linear effects will be put aside and differences between the different ME tracks will be modelled as coming from 1) a temporal delay between the tracks and 2) a different equalization on each of the ME tracks[1].

Hence, the naive model (2) is replaced here by :

$$\forall i \in [1:N], m_i(t) = h_i * (d^{\tau_i} b) + v_i(t) \qquad (4)$$

where h_i is a linear filter, $*$ is the convolution operator and d^{τ_i} denotes a delay operator. The more realistic model (4) thus assumes that there is a single ME track, which is filtered differently for each international version. Furthermore, the different versions are no more assumed to be perfectly synchronized. Note that τ_i can be positive or negative.
The ME track of a movie soundtrack may contain highly varying material comprising very different sorts of music, noises, bursts of sounds and complex sound effects. Thus, the *a priori* information available on the signal $b(t)$ is very weak and the model (4) encodes all we can say doubtlessly about it. For convienence, (4) can also be cast into the spectral domain to become :

$$\forall i \in [1:N], \mathcal{F}\{m_i\}_n = \mathcal{F}\{h_i\}.\mathcal{F}\{d^{\tau_i} b\}_n + \mathcal{F}\{v_i(t)\}_n \qquad (5)$$

[1] As movie soundtracks are not often hardly compressed, neglecting non-linear effects may not be such an extreme approximation.

Where n is a frame index of the STFT and . denotes element-wise multiplication.

2.3. Indeterminacies and equalizing inversion

One can guess here that there are indeterminacies in the equivalent models (4) and (5). Indeed, let g be an inversible linear filter, then $h_i * g^{-1}$ and $g * b$ instead of h_i and b lead to the same model. Likewise, if τ_i is replaced by $\tau_i - \tau_0$ and $b(t)$ by $b(t + \tau_0)$, the same model is found again.
In order to fix these indeterminacies, we can arbitrarily choose one i_0 from $[1:N]$ that will be denoted the *target version*. Then, we will set h_{i_0} to the identity and τ_{i_0} to 0. This way, we will have

$$m_{i_0}(t) = b(t) + v_{i_0}(t)$$

This can be interpreted saying that the ME track that will be estimated is the particular version found in mixture i_0.
An important remark must be made here with respect to the choice of the target version, for an important issue was actually raised choosing i_0.
When considering old movies or resampled soundtracks, it is possible and was indeed observed in practice that some international versions may be of limited bandwith compared to others even for the same movie. In this situation, it is actually impossible to consider that the ME tracks with the largest bandwith can be obtained from the damaged ones using a linear filter. Indeed, this would involve infinite gains in the frequency domain for the corresponding filters $\mathcal{F}\{h_i\}$ which cannot be allowed. Hence, the target version needs to be one with the largest bandwidth. Still, the separation process will imply the inversion of the equalizing filter h_i in order to estimate so-called *alignment filters* h_i^{-1}. This cannot be achieved correctly if some of the filters h_i are destructive in the frequency domain as may occur in case of different bandwiths. We will nevertheless consider that all the filters h_i are invertible, which basically means that there was no destructive equalization process and that all international versions roughly have the same bandwidth, which may not always be a realistic assumption. Still, we will see in 3.2.2 that the chosen method to estimate filters actually handles this instability issue even when real data does contain damaged mixes.

2.4. Sparse voice signals

In (1), we have modelled the mixtures as the sum of two components called respectively the background (ME) and voice tracks. The first one of these components has been modelled in (4) as the filtered and delayed version of a common underlying and unknown signal of interest. It is extremely difficult to make any complementary realistic assumptions about this background signal as it may contain very different sorts of audio material.
On the contrary, we will show here that it is possible to highlight important general properties of the additive voice signals $v_i(t)$. Indeed, even in realistic usecases, the additive signals mainly consist of voice sounds, whether they be sung or spoken. It would be possible to add strong *a priori* information on the signals $v_i(t)$ using all the litterature concerning models of voice signals (see Section 1), but it is out of the scope of the paper.

The property that will be particularly interesting in the scope of this paper is the generally observed *sparsity* of the voice signals in the time-frequency domain. Except in the case of whispered or cracky voiced sounds, voice signals can indeed be modelled as the sum of a limited number of sinusoids, typically 50 per time frame or less.
In order to test for the empirical validity of assessing spectral sparsity of realistic voice signals, we computed the distribution of the element-wise ratio between $||\mathcal{F}\{m_i\}||$ and $||\mathcal{F}\{b_i\}||$ computed on a 30s excerpt of a real movie soundtrack[2]. Figure 1 shows the result of this experiment and it can be seen that this distribution is very strongly peaked on 1, indicating that most of the time-frequency bins of the mixture actually correspond to their value in the desired background signal. As can be seen on this Figure, even if most of the time-frequency bins are the same in both background and mixture signals, differing bins due to the voice signal are extremely different, which is highlighted by the very strong difference between the mean and median values of the ratios. The back and front tails of the distribution represent time-frequency bins where the voice is present.

[2]In order to build such signals, an excerpt with no voice has been added to an excerpt with voices only to allow both signals b_i and v_i to be known as well as to consider a realistic situation for experiments.

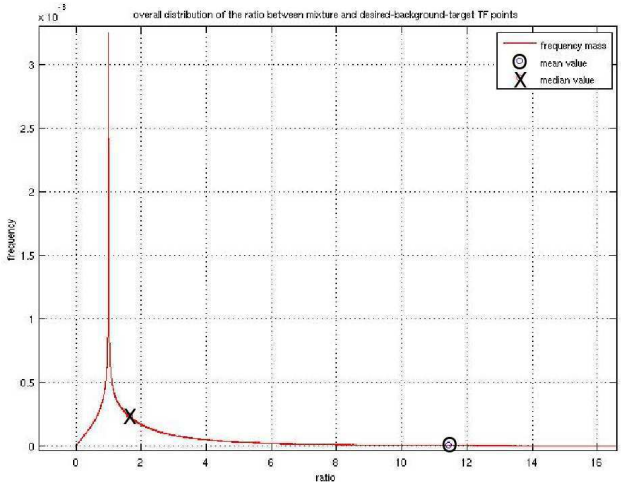

Fig. 1: Distribution of the ratio between Mixture and ME tracks TF points amplitudes .

The sparsity of the voice track in the time frequency domain induces the sparsity of the voice track activation for a given time-frequency point across the version signals. There is a low probability that a time-frequency point has a voice activity for any version. This property will be used in a method described in Section 2.5.1.

2.5. Derived extraction methods

2.5.1. Median filtering

The extraction methods used in this study will mainly involve the naive model (2) and (3). There will indeed be a first processing step called *alignment* that is reviewed in 3.2 and that will allow to transform real data in order for it to be correctly modelled by the naive model.
We thus suppose here that the observed (transformed) signals are the superposition of the *same* desired background signal with different additive voice signals that are *sparse* in the frequency domain. As mentioned in previous Section, Figure 1 showed that if one considers a random time-frequency bin from the STFT of the mixture, there is a high probability

for its value to correspond to the value of the desired background signal at that bin. If this is not the case, there is a low probability that the voice is active on that bin for several versions since the voice signals are strongly independant.

The distribution of the amplitudes of a given time-frequency point among the channels is thus strongly peaky around the background value, with sometimes outliers when the voice is active on that point. To estimate the amplitude of the background from this distribution, the median value of the set is a clever choice because the set size is very small.

As we have seen in 2.2, it is actually very unrealistic to suppose that the different voices of a movie soundtrack are perfectly synchronized (words, pitch contours are different across the languages). The fact that the additive voice signals are far from being perfectly synchronized in real situations makes the independance assumption between them less unrealistic, even if the general acoustic properties of voice signals across the different versions are similar (with respect to the gender of the speaker in particular).

2.5.2. Min filtering

The rationale behind median filtering was explained in the previous section 2.5.1. Another point of view can be adopted. It is rather common in the source separation litterature to consider that the power spectra of the different sources simply add to produce the power density spectrum of the resulting mixture, which is equivalent to consider that the sources are independent. This can be written :

$$\forall i \in [1:N], ||\mathcal{F}\{m_i\}||^2 = ||\mathcal{F}\{b\}||^2 + ||\mathcal{F}\{v_i\}||^2 \quad (6)$$

This assumption amounts to considering that an additive signal can only increase the energy of all frequency bins. Another filtering strategy can very simply be deduced from this expression. Given several observations of some time-frequency bin, the most likely to correspond to the common underlying ME track will necessarily be the one whose energy is smaller according to model 6. This idea leads to estimating the background by simply chosing the observation that has the minimal energy at each time-frequency point.
As it simply boils down to neglecting phase effects which is actually a rather coarse assumption, we will see that when N gets large, the ME track estimation performed this way will actually become very poor. This can be interpreted stating that the more different versions there are, the more probable it is that phase has a non negligible role in one of the version for some given time-frequency bin.
Still, the min filtering will be actually useful in situations where only a few observations are available. Indeed, the median filtering in this case will lead to the addition of *systematic* noise in the estimation of *every* time-frequency bin. We will also see in 4 that it is also very interesting to apply the min filter *after* having applied the median filter in order to guarantee that the estimated ME track does not have higher energy than the target ME track on any time-frequency bin.

2.5.3. Mean filtering

One may wonder why only *median* and *min* operators were considered in the preceding instead of some simple mean between the different observations for every time-frequency bin. As we will see in 4, taking the mean is actually a very bad idea, since the data is intrinsically very likely to contain *outliers* that can have a strong influence on this estimation. Figure 1 shows that the ME and mixture tracks very often share the same value, it also strongly suggests that when they do not, they actually differ greatly, a property to which the mean would be actually very sensitive to since the sample set is small, contrarily to the median filtering.

3. ALGORITHM

In section 2 we have stated the main assumptions of the proposed method, namely

- The different observed mixtures consist in the sum of a differently-filtered and delayed common underlying signal of interest. This is summed up in model (4) as well as in its spectral counterpart (5).
- The additive voice signal can be considered as being *sparse* in the time-frequency domain.

Both of these assumptions will be used extensively in the following. We will here present a complete

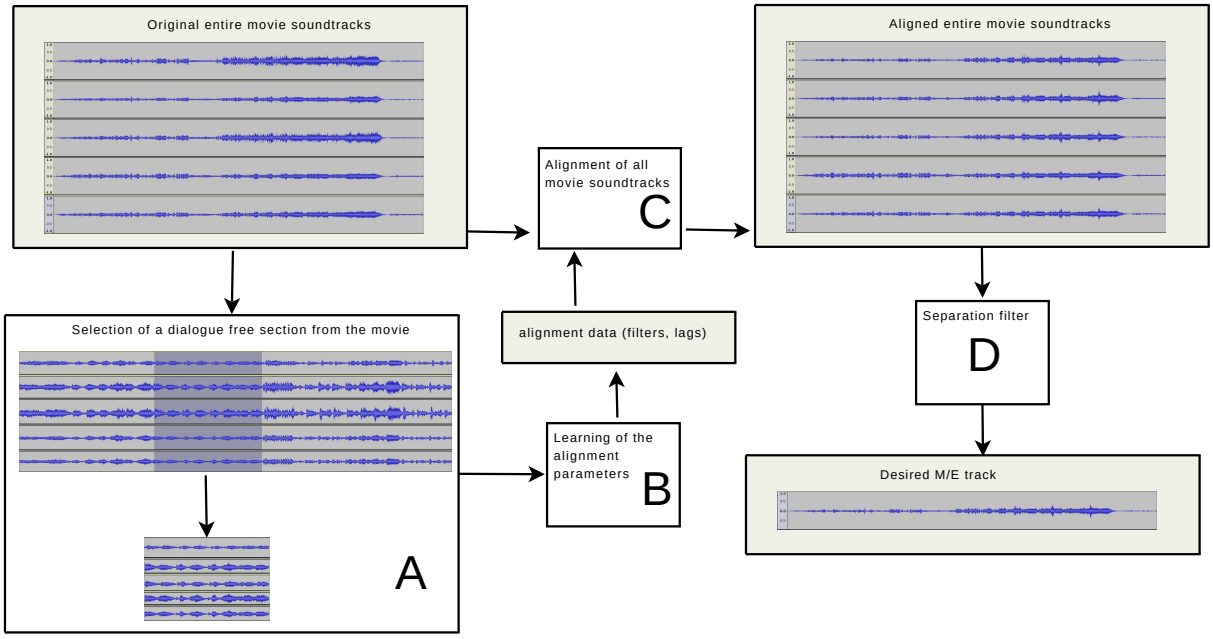

Fig. 2: General structure of the proposed separation method.

extraction system that can very rapidly extract the ME track with decent performance. We will first give an overview of the proposed system and will then linger on the main two problems it has to solve : alignment and statistical filtering.

3.1. Overview

The general structure of the proposed separation method is outlined in Figure 2. The extraction procedure involves four important steps that will be reviewed now :

- Step A : First, some user needs to identify manually some section of the movie soundtrack on which there is no voice and on which there is a significant spectral richness of the ME track. Such a configuration is typically very easily found on the musical parts of a soundtrack. In our experiments, a minute-long excerpt was fine to process one whole movie. It would of course be possible to automatically detect such an excerpt, but we have not lingered on this task here.
- Step B : Given some isolated excerpt on which we know that no voice signal is present, we can estimate the *alignment parameters* that consist in the inverse filters h_i^{-1} and delays τ_i from model (4).
- Step C : With these alignment parameters, we will be able to easily transform the data so that the naive model (2) is adequate. This is explained in section 3.2.
- Step D : The naive model (2) now being realistic on the transformed data, we will be able to derive a simple statistical filtering scheme (as detailed in 2.5) in order to estimate the desired ME track from the mixtures. This is explained in section 3.3

3.2. Alignment

In section 2, we saw that the different available international versions, called *mixtures*, can be modelled in several ways. The most simple model simply stated in (2) that each mixture $m_i(t)$ was the sum of a common background signal of interest, $b(t)$, with

a voice signal $v_i(t)$. We saw that for several reasons, this model can in no way be called realistic. On the contrary, we derived in (4) and (5) another model which accounts for most obvious discrepancies observed between the ME tracks of real international versions of the same movie, namely a time delay and a different equalization.
Supposing that we knew all the mixing filters h_i, that they were invertible and that we knew all the delays τ_i, we would have :

$$\forall i \in [1:N], h^{-1} * m_i(t+\tau_i) = b(t) + h^{-1} * v_i(t+\tau_i) \tag{7}$$

Which can be recast in the spectral domain as follows : $\forall i \in [1:N]$

$$\mathcal{F}\{h_i^{-1}\}.\mathcal{F}\{d^{-\tau_i} m_i\}_n = \mathcal{F}\{b(t)\}_n + \mathcal{F}\{h_i^{-1}\}.\mathcal{F}\{d^{-\tau_i} v_i\}_n \tag{8}$$

These expressions mean that we could process the data so as to be able to consider the naive model (2) as valid on this processed data.
As we will see in 3.3, the naive model (2) along with the sparsity assumption found in 2.4 will allow us to derive very efficient extraction method. This is why estimating the delays and filters will be of special interest to us.
We saw in 3.1 that a first step of the system consists in selecting an excerpt of the soundtrack on which there is no voice signals, which can easily be found in practical situations. On such an *instrument excerpt* $\mathbb{T} = [t_1 : t_2]$ we have :

$$\forall t \in \mathbb{T}, \begin{cases} \forall i \neq i_0, & m_i(t) = \sum_u h_i(u) b(t - \tau_i - u) \\ & m_{i_0}(t) = b(t) \end{cases} \tag{9}$$

We will first review methods to estimate delays in 3.2.1 and then methods to estimate *alignment filters* h_i^{-1} in 3.2.2.

3.2.1. Temporal alignment

We noted in 2.2 that delays between tracks can actually be quite large and we observed values up to almost half a second in practical situations. As was suggested in 2.3, some mixture may be taken as a reference and if we have identified an *instrumental excerpt* $\mathbb{T}$ on which no additive voice signal is present, relations (9) hold.
Let R_{bb} be the autocorrelation of b, i.e :

$$R_{bb}(\tau) = \mathbb{E}_{t\in\mathbb{T}}[b(t)b(t+\tau)]$$

Let us furthermore define $R_{i_0 i}$ as the correlation on $\mathbb{T}$ between the target track i_0 and track i, i.e :

$$R_{i_0 i}(\tau) = \mathbb{E}_{t\in\mathbb{T}}[m_{i_0}(t) m_i(t+\tau)]$$

We readily see that $R_{i_0 i}$ can be expressed slightly differently as :

$$R_{i_0 i}(\tau) = (h_i * R_{bb})(\tau - \tau_i) \tag{10}$$

Thus, we see that the correlation between the target track and another international version on some instrumental excerpt $\mathbb{T}$ corresponds to the delayed convolution of the autocorrelation of b by the equalizing filter h_i.
In theory, if the autocorrelation of the background signal on the instrumental excerpt $\mathbb{T}$, R_{bb}, is very peaky (this is the case for wide band signals), i.e $R_{bb}(\tau) \approx \delta_\tau$ and if the impulse response of the mixing filter h_i is centered on zero, the maximum of $R_{i_0 i}$ will be observed for $\tau = \tau_i$, which gives a very simple procedure to estimate τ_i.
Annoying perturbations were nevertheless observed in cases where the mixing filters were complicated and when the autocorrelation R_{bb} of the background signal was very different from the idealized case. Indeed, both of these effects can interact and translate the maximum of $R_{i_0 i}$ far from the desired delay.
In order to address this issue, different strategies were tried and a practical and efficient solution was found that consists in computing the correlations on the squared signals instead of the raw data. Indeed, the autocorrelation $R2_{bb}$ of $b^2(t)$ computed on an instrumental excerpt shows much less perturbations than R_{bb} as can be seen in Figure 3.

It was thus decided to compute the correlations between the squared signals in order to estimate the delays between the different versions. The expression of the correlation of these squared signals is indeed given by :

$$R2_{i_0 i}(\tau) = \mathbb{E}_{t\in\mathbb{T}}[m_{i_0}^2(t) m_i^2(t+\tau)] \tag{11}$$

Which is readily shown to be :

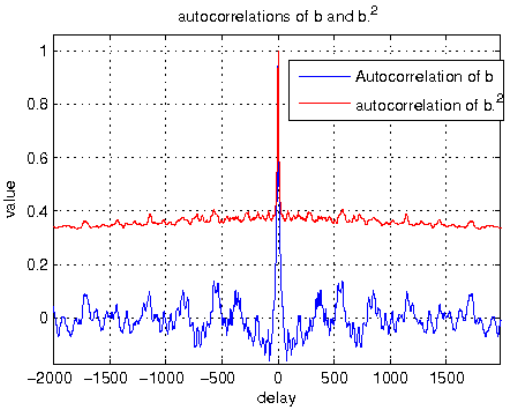

Fig. 3: Normal and energy autocorrelations for the ME track computed on a 30s excerpt.

$$R2_{i_0 i}(\tau) = \left(h_i.^2 * R2_{bb}\right)(\tau - \tau_i) \qquad (12)$$

This expression shows that the correlation between $m_{i_0}^2$ and m_i^2 can also be interpreted as the convolution between a filter and a correlation function. Still, the correlation function $R2_{bb}$ is much smoother that R_{bb} and h_i^2 is centered on nearly the same point as h_i anyway.

Figure 4 shows the differences between $R_{i_0 i}$ and $R2_{i_0 i}$ whereas Figure 5 shows the delays computed using these two different methods on a realistic example. We can first see that the correlation function computed on squared signals is indeed much smoother and then that the delays computed this way are much more adequate than those computed using plain correlations functions. Intuitively, this correlation scheme may be far less sensitive to phase shifts and may thus concentrate on energy correlations, which seems to be a good feature for our particular purpose.

3.2.2. Spectral alignment : equalization

When the delays between the different versions have been estimated, we can readily translate accordingly

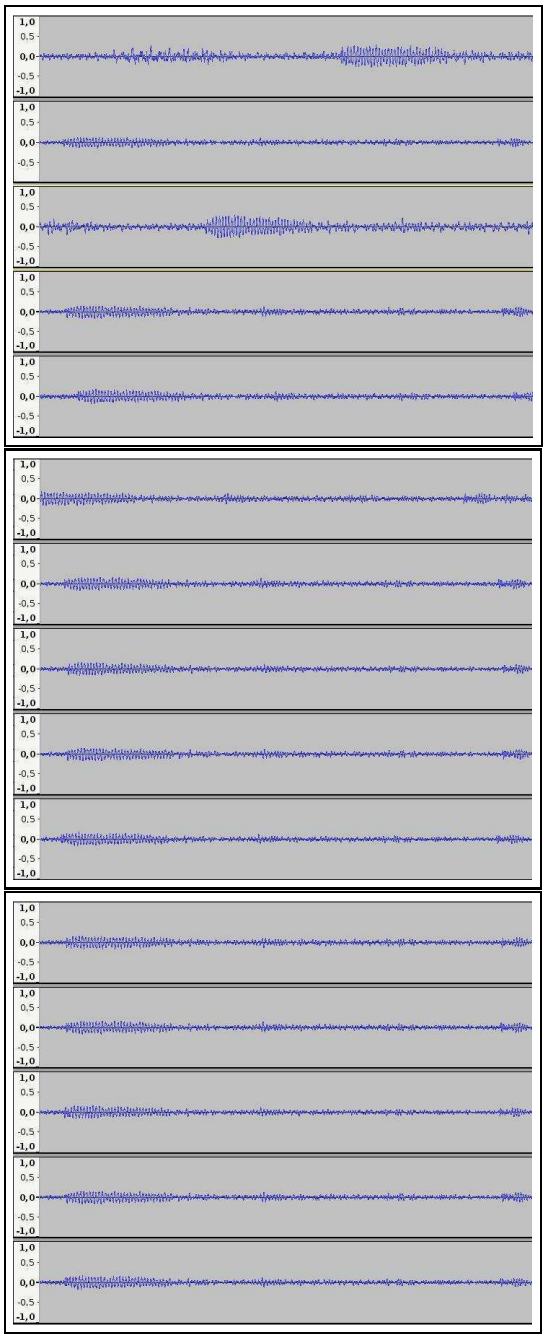

Fig. 5: Zoom for temporal alignment. Original tracks (above). Alignment with correlation (middle). Alignment with energy correlation (below).

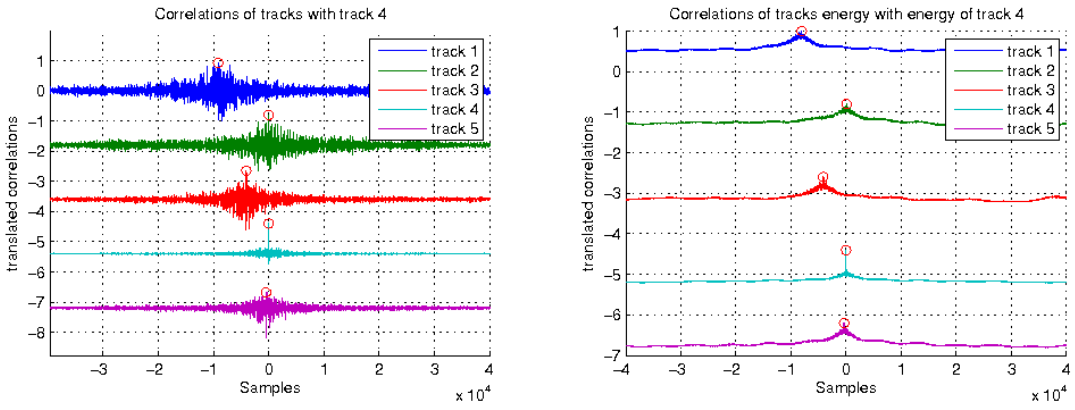

Fig. 4: Two different strategies for temporal alignment. Left : correlation. Right : energy correlation. See text in 3.2.1.

the different versions on the instrumental excerpt so as to obtain :

$$\forall t \in \mathbb{T}, \forall i \in [1:N], m_i(t+\tau_i) = (h_i * b)(t)$$

Which can be cast into the spectral domain using some STFT $\mathcal{F}$ on $\mathbb{T}$ to get:

$$\mathcal{F}\{d^{-\tau_i} m_i\}_n = \mathcal{F}\{b\}_n . \mathcal{F}\{h_i\} \tag{13}$$

Where n is a frame number. If we suppose as was suggested in 2.3 that filters h_i are invertible, we can write (13) as :

$$\frac{1}{\mathcal{F}\{h_i\}} = \frac{\mathcal{F}\{b\}_n}{\mathcal{F}\{d^{-\tau_i} m_i\}_n}$$

Where $\frac{a}{b}$ denotes element-wise division. This is equivalent to :

$$\mathcal{F}\{h_i^{-1}\} = \frac{\mathcal{F}\{b\}_n}{\mathcal{F}\{d^{-\tau_i} m_i\}_n} \tag{14}$$

Theoretically, this expression allows us to compute $\mathcal{F}\{h_i^{-1}\}$ which can then be injected in (8) in order to be able to consider the naive model (3) as valid on the processed data $\mathcal{F}\{h_i^{-1}\}.\mathcal{F}\{d^{-\tau_i} m_i\}$. Still, in practice we cannot assume that expressions such as (14) are indeed perfectly verified for real data in the sense that the ratio $\frac{\mathcal{F}\{b\}_n}{\mathcal{F}\{d^{-\tau_i} m_i\}_n}$ is never observed as being constant over the frames n, which can be explained stating that real mixing filters h_i may be non-linear or even non-invertible.

However, we can go on and nevertheless estimate $\mathcal{F}\{h_i^{-1}\}$ from the observation of the different $\frac{\mathcal{F}\{b\}_n}{\mathcal{F}\{d^{-\tau_i} m_i\}_n}$, i.e one for each frame of the *instrumental excerpt*. Several estimation schemes were tried, some of which are listed now :

- The first naive idea would be to simply take $\mathcal{F}\{h_i^{-1}\}$ as the expectation of the absolute value of these ratios over n, i.e :

$$\mathcal{F}\{h_i^{-1}\} = \mathbb{E}_n\left[\frac{||\mathcal{F}\{b\}_n||}{||\mathcal{F}\{d^{-\tau_i} m_i\}_n||}\right] \tag{15}$$

  This strategy was called *the non weighted real scheme* for the resulting filters are considered to be gains only and because every time-frequency bin is taken into account the same way. An example of filters computed this way is given in Figure 6 upper-left.

- Another strategy was to try to weight the different time-frequency bins according to their value, i.e :

$$\mathcal{F}\{h_i^{-1}\} = \frac{\sum_n ||\mathcal{F}\{b\}_n||.||\mathcal{F}\{d^{-\tau_i} m_i\}_n||.\frac{\mathcal{F}\{b\}_n}{\mathcal{F}\{d^{-\tau_i} m_i\}_n}}{\sum_n ||\mathcal{F}\{b\}_n||.||\mathcal{F}\{d^{-\tau_i} m_i\}_n||} \tag{16}$$

  This strategy was called *the weighted complex scheme* because the resulting filters are complex and because every time-frequency bin is taken into account accordingly to its amplitude. An

example of filters computed this way is given in Figure 6 upper-right.

- We also tried to weight each time-frequency ratio $\frac{||\mathcal{F}\{b\}_n||}{||\mathcal{F}\{d^{-\tau_i}m_i\}_n||}$ accordingle to the *magnitude squared coherence* C_{bm} between the two signals b and $d^{-\tau_i}m_i$. This quantity is given as

$$[C_{bm}]_n = \frac{|[P_{bm}]_n|^2}{[P_{bb}]_n[P_{mm}]_n}$$

Where $[P_{bb}]_n$ and $[P_{mm}]_n$ are the power spectral densities of b and $d^{-\tau_i}m_i$ for frame n respectively and $[P_{bm}]_n$ is the cross power spectral density of these two signals. Inverse filters were then computed by :

$$\mathcal{F}\{h_i^{-1}\} = \frac{\sum_n [C_{bm}]_n \cdot \frac{\mathcal{F}\{b\}_n}{\mathcal{F}\{d^{-\tau_i}m_i\}_n}}{\sum_n [C_{bm}]_n} \quad (17)$$

This strategy was called was called the *coherence scheme*. Examples of filters computed this way can be found in Figure 6 lower-left.

- Finally, a last strategy was to try to use the same idea as in the weighted complex scheme but computing real filters only, i.e :

$$\mathcal{F}\{h_i^{-1}\} = ...$$
$$\frac{\sum_n ||\mathcal{F}\{b\}_n||.||\mathcal{F}\{d^{-\tau_i}m_i\}_n||.\frac{||\mathcal{F}\{b\}_n||}{||\mathcal{F}\{d^{-\tau_i}m_i\}_n||}}{\sum_n ||\mathcal{F}\{b\}_n||.||\mathcal{F}\{d^{-\tau_i}m_i\}_n||.}$$

Which can be simplified to :

$$\mathcal{F}\{h_i^{-1}\} = \frac{\sum_n ||\mathcal{F}\{b\}_n||^2}{\sum_n ||\mathcal{F}\{b\}_n||.||\mathcal{F}\{d^{-\tau_i}m_i\}_n||.} \quad (18)$$

This strategy was called *the weighted real scheme* because the resulting filters are real and because every time-frequency bin is taken into account accordingly to its amplitude. An example of filters computed this way is given in Figure 6 lower-right.

A first thing to notice here is that weighted schemes (16) and (18) allow to handle the non-stability issues that might occur in (15) and that were raised in 2.3 as caused because of some bins in $||\mathcal{F}\{d^{-\tau_i}m_i\}_n||$ that may be very close to zero. Indeed, if such is the case, a weighted scheme will end up not taking such damaged bins into account thus preventing non realistic inverse filters to be estimated. Indeed, it can be seen in 6 that the non weighted (15) and coherence (17) schemes both are very unrealistic as the estimated filters involve amplification gains reaching 80dB, which is totally unacceptable. The weighted complex scheme (16), even if theoretically could estimate a more realistic filter, is actually lost for high frequencies due to phase incoherence between the different versions, that could not properly be modelled as a simple linear effect.
On the contrary, the weighted linear scheme (18) seems to produce filters that do not seem to show all these weaknesses and that indeed look very smooth and realistic as to the range of their amplification gains. It was hence decided to proceed to spectral alignment using the weighted real scheme (18).

3.2.3. Alignment results

Figure 7 gives qualitative results of the overall alignment procedure. An instrumental excerpt of 30s was selected from a real movie soundtrack for which $N = 5$ to be used as the *instrumental excerpt* $\mathbb{T}$. Then, track 4 was chosen arbitrarily as a target track and the alignment parameters were computed. Delays were estimated using the squared correlation strategy (11) and alignment filters were estimated using the weighted real scheme (18).
After this alignment estimation, and for the sake of illustration, *another* instrumental excerpt $\mathbb{T}_2$ was chosen from the original soundtrack, that is displayed *as is* in the upper part of Figure 7. Alignment parameters estimated on $\mathbb{T}$ were then used to compute the $h_i^{-1} * d^{-\tau_i}m_i$ signals on $\mathbb{T}_2$, and these signals are displayed in the lower part of the Figure. As can be seen, the dynamics of the different versions look much more similar that before alignment. The proposed system is then able to provide a descent estimation of the alignment parameters τ_i and h_i^{-1} for further processing.

3.3. Statistical filtering

In the previous section 3.2 about alignment, we saw that we may estimate some alignment parameters, namely delays τ_i and alignment filters h_i^{-1} that allow us to consider the naive model 3 to be valid on some

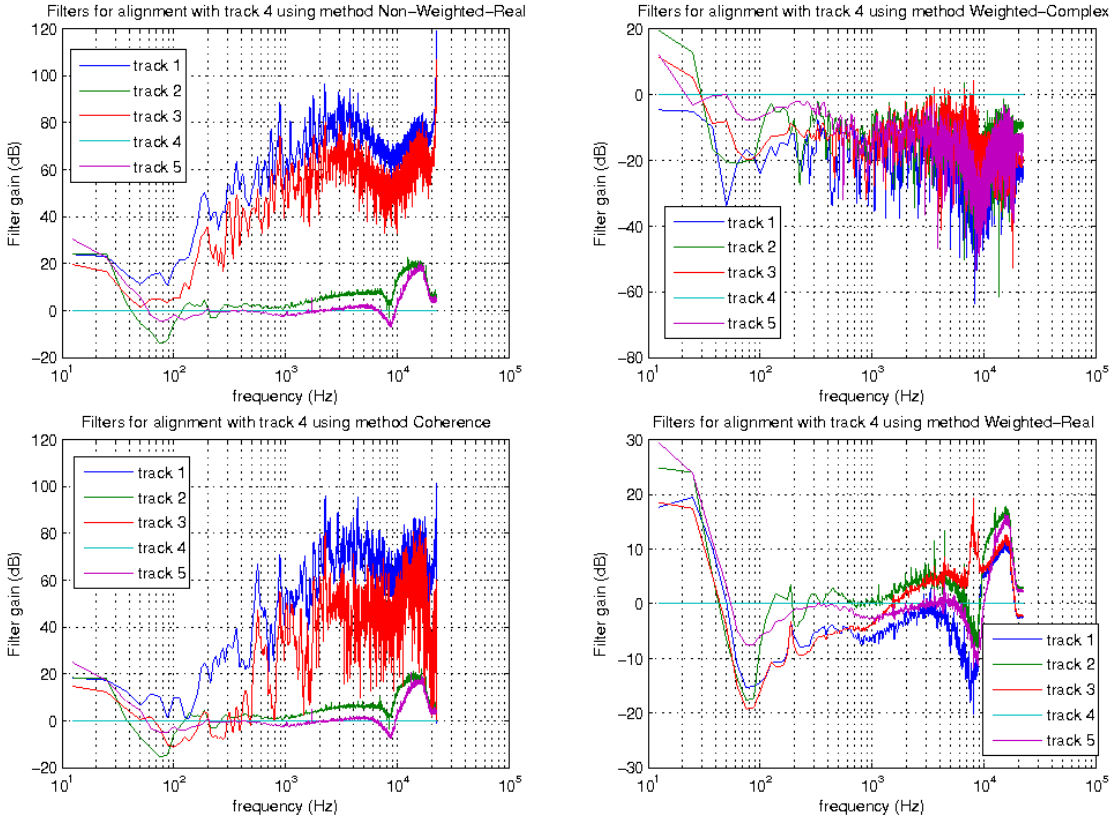

Fig. 6: Different spectral alignment designs'.

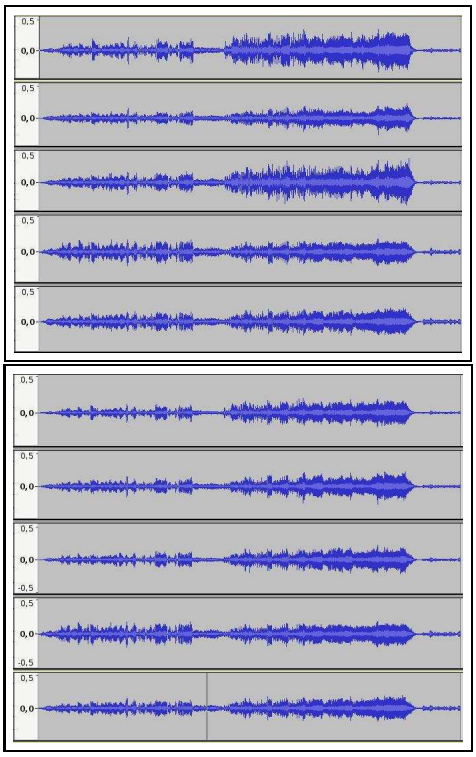

Fig. 7: Waveforms of non aligned (above) and aligned (below) versions.

processed data m'_i :

$$\forall i \in [1:N], m'_i(t) = b_i(t) + v'_i(t)$$

with :

$$\mathcal{F}\{m'_i\}_n = \mathcal{F}\{h_i^{-1}\}.\mathcal{F}\{d^{-\tau_i} m_i\}_n$$

and

$$\mathcal{F}\{v'_i\}_n = \mathcal{F}\{h_i^{-1}\}.\mathcal{F}\{d^{-\tau_i} v_i\}_n$$

Thus, the different extraction methods described in 2.5 can be applied on the transformed signals to obtain the desired ME track. Several quantitative results will be given in the following section, but in order to test the exprimental validity of the method quickly, Figure 8 shows the distribution of the ratio between the amplitude of the STFT points of the estimated ME track amplitudes and the amplitude of the STFT points of the target ME track using the median filter described in 2.5.1 on aligned real data. The Figure shows a very pronounced peak on 1, which shows that the estimator is unbiased.

4. RESULTS

4.1. Performance metrics

The metric used to evaluate the source separation is the Source to Distortion Ratio (SDR), defined in the BSS_EVAL toolbox [4]. Among all the measurements proposed in this toolbox, SDR is the most correlated to the human assessment [5].

We recall here the definition of this measurement:

The Signal-to-Distortion Ratio is:

$$\mathrm{SDR} = 10 \log_{10} \frac{||s_{\mathrm{target}}||^2}{||s_{\mathrm{target}} - s_{\mathrm{estimated}}||^2} \qquad (19)$$

where s_{target} is the target source signal, $s_{\mathrm{estimated}}$ is the estimated source signal.

4.2. Synthetic experiment

In order to validate the extraction method, a first experiment was conducted in which perfect alignment was simulated. In order to do this, we chose an excerpt of the movie on which no voice signal was present and chose arbitrarily one of the versions that was then simply duplicated. Then, an excerpt of the same length was chosen on which no ME signal was

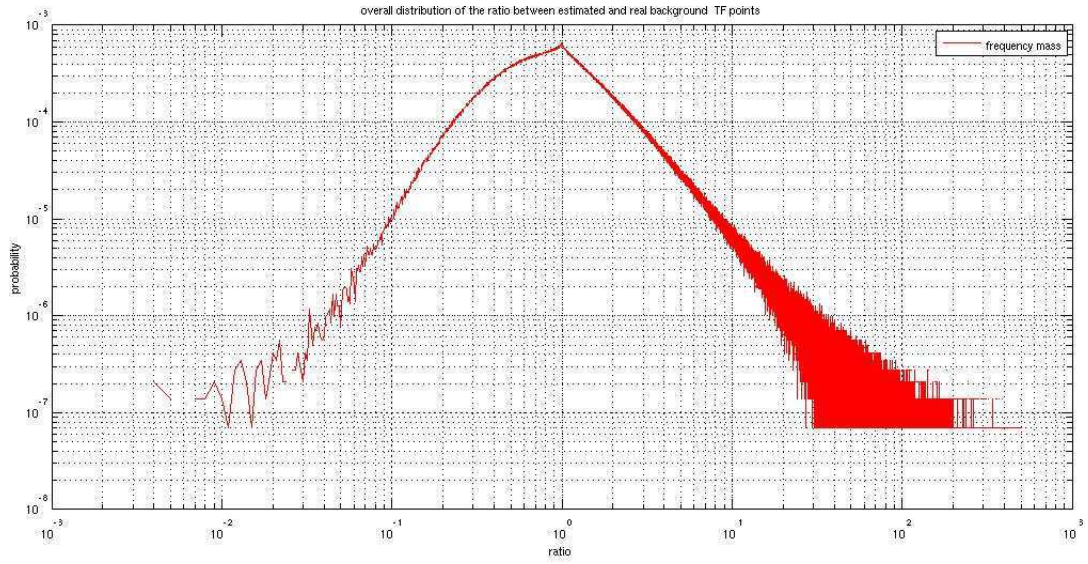

Fig. 8: Distribution of the ratio between estimated and original ME track with the MEDIAN filter.

present. These signals were then mixed together so as to obtain data on which real voice signals from different international versions were separately mixed down on exactly the same ME signal. The naive model (2) is thus known to be perfectly appropriate for this synthetic data. Median filtering was then performed along with min filtering between the result of the median filter with one of the mixtures to obtain the estimated ME track. Perceptually, the separation was extremely good. Quantitative results are given in table 1 for 5 different versions and confirm that performance are very good when the naive model is adequate (SDR and SAR around 12dB, SIR around 30dB). The audible noise on some resulting signals are often related to the voice parts that are non sparse, namely the noise present in consonants when they occur simultaneously in the versions.

4.3. Realistic experiments

In the following experiments, alignment data was estimated on an excerpt from a movie soundtrack on which no voice signal was present, and the mixtures artificially produced from two excerpts of the same length, the first of which only contains ME signal and the second of which only contains dialogues. We tried to make the mix realistic, in the sense that voice signals were slightly overmixed compared to the ME track, in order for them to be very clearly understandable on the mixtures.

Targetted ME	SDR ME (dB)
1	12.3153
2	12.2399
3	11.9619
4	11.6382
5	11.9122

Table 1: Averaged ME Extraction results for an idealized test case using the MEDIAN filter for 5 available international versions. Experimental setup detailed in 4.2

4.3.1. Performance of the MEAN filter

In order to verify what was explained in 2.5.3 concerning the poor expected performance of the *mean* filter, an experiment was conducted using this filter. Performance is shown in Figure 2 and are not good. Perceptually, the resulting ME track did sound really damaged or *blurred* compared to the original version.

4.3.2. Performance of the MEDIAN filter

Performance of the method was of course tested us-

Targetted ME	SDR ME (dB)
1	0.7269
2	0.6455
3	1.0951
4	0.3257
5	0.5211

Table 2: ME Extraction results for a real world test case using the MEAN filter for 5 available international versions. Experimental setup detailed in 4.3.1

ing the median strategy, which was presented in 2.5.1. As was hinted in 2.5.2, the result of the median filter was further processed through the min filter in order to guarantee that the resulting signal did not have higher energy than the target version on any time-frequency bin.
The extracted ME track did sound much better than with the MEAN filter, even if it was not as good as in the idealized case. Still, the resulting tracks did perceptually sound good nonetheless, and were anyways a good basis for building another international version of the movie using some further audio engineering. In any cases, the resulting estimated ME track did not contain understandable voice signals.
The quantitative results given in tables and show that the extraction method does give very reasonable performance, when considering its extreme efficiency : one whole movie could be processed in less than 15 minutes, which is remarkable considering the computing time of state-of-the-art systems (around 10 times real-time). Performance actually seems quite similar with 3 or 5 international versions. However, experience showed that the more international versions, the better in general.

4.3.3. Performance of the MIN filter

A last experiment was conducted for which only two international versions were provided. As suggested in 2.5.2, the chosen strategy in this case was the MIN filtering whose results are given in table 5. These results are actually quite good and show that the system can already provide some interesting results in cases where only two international version of the movie are available.

Targetted ME	SDR ME (dB)
1	3.9355
2	1.5180
3	2.7493
4	0.4616
5	2.1307

Table 3: ME Extraction results for a real world test case using the MEDIAN filter for 5 available international versions. Experimental setup detailed in 4.3.2

Targetted ME	SDR ME (dB)
1	3.3700
2	0.9421
3	3.1034

Table 4: ME Extraction results for a real world test case using the MEDIAN filter for 3 available international versions. Experimental setup detailed in 4.3.2

Targetted ME	SDR ME (dB)
1	2.2242
2	-0.0801

Table 5: ME Extraction results for a real world test case using the MIN filter for 2 available international versions. Experimental setup detailed in 4.3.3

5. CONCLUSION

In this paper, we proposed a method to address the problem of the extraction of a common signal between several tracks, with an application on the extraction of the Music and Effects track from several localized versions. Preprocessing strategies have been proposed to synchronize the localized versions, and to equalize the music and effect tracks. Results have been computed on artificially mixed localized versions when the ME track is known to be exactly the same in the different versions. In this case, the results have a good quality. When the test signals are not aligned nor equalized, the imperfection of the preprocessing step yields lower performances. However, they can be used in most realistic cases, and they use far less processing time than other source separation techniques. The perspective of this work is to investigate multi-resolution analysis to improve the sparseness of the sources to separate. Indeed, using a sparser sound representation improves the source separation performance in this kind of context, as for stereo source separation [6].

6. REFERENCES

[1] R. McAulay and T. Quatieri. Speech analysis/Synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing [see also IEEE Trans. on Signal Processing],*, 34(4):744–754, 1986.

[2] Jean-Louis Durrieu, Gal Richard, Bertrand David, and Cdric Fvotte. Source/filter model for main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3), March 2010.

[3] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Processing*, 52(7):1830–1847, 2004.

[4] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.

[5] B. Fox, A. Sabin, B. Pardo, and A. Zopf. Modeling perceptual similarity of audio signals for blind source separation evaluation. In *Proc. of Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA)*, volume 4666, page 454. Springer, 2007.

[6] R. Gribonval. Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, 2002.