# EXTRACTION AND REMIXING OF DRUM TRACKS FROM POLYPHONIC MUSIC SIGNALS

*Olivier Gillet and Gaël Richard*

GET / Télécom Paris, CNRS LTCI. 37 rue Dareau 75014 Paris, France
[olivier.gillet, gael.richard]@enst.fr

## ABSTRACT

This paper presents a novel algorithm to extract the drum track of a polyphonic music signal, based on a harmonic / noise decomposition. This algorithm is causal and does not require prior knowledge or learning. The input signal is split into several frequency bands in which the signal is separated in a deterministic and a stochastic part. The stochastic part can be efficiently used to detect drum events and to resynthesize a drum track. Possible applications include drum transcription, remixing, and independent processing of the rhythmic and melodic components of music signals. Results obtained from real recordings of popular music are presented, as well as a perceptual evaluation of the quality of remixed signals.

## 1. INTRODUCTION

The automatic extraction of drum tracks from polyphonic music signals has many possible applications, both in the field of digital audio effects and audio indexing: beat tracking, automatic transcription of rhythmic parts, style identification, rhythm driven audio effects, remixing or DJing. However, it is not clear how this extraction should be performed, and many solutions have been proposed. The approaches to this problem can be roughly classified in two categories: transcription-based and sound-separation based.

Automatic rhythmic transcription aims at obtaining a detailed transcription (such as a music score) of the drum track of a music piece. The extraction of the drum signal itself is not required for the transcription. However, in order to achieve efficient results, most of the transcription systems try to obtain information about the timbre of each drum instrument used in the music piece - since in many cases an adapted machine learning or template matching algorithm will perform better than a generic one. Thus, by-products of the transcription, such as templates or adapted models, are often available, and can be used for the resynthesis. Zils et al. introduced in [1] an algorithm operating in the time domain that extracts a rhythmic transcription as well as short adapted templates, which can be combined to resynthesize the drum track of the original signal. A similar approach was used by Yoshii in [2] using time-frequency templates. In [3], Sandvold et al. used a machine learning approach to detect occurrences of drum instruments in polyphonic music signals. This study suggests that automatically ranking the detected occurrences according to a confidence score could identify the most representative of them - hence, extract good candidates for template-based synthesis.

Source separation aims at extracting individual sound sources from music recordings, using information gathered from different sensors - for example the two channels of a stereo recording [4] or microphone arrays. A growing number of works recently focused on single source audio separation [5]. While source separation is not geared toward drum track extraction, the separation algorithms generally manage to isolate the drums as one or several distinct sources. It is worth to mention that while it is possible to automatically identify the drums among the different sources proposed by the separation algorithm, many source separation algorithms require manual tuning, and the number of sources has to be known in advance. Hence, they are not always suitable for fully-automated processing. Other separation approaches, like those developed by FitzGerald [6] are using prior knowledge about the timbre of drum instruments to initialize separation algorithms, and can be thus used in an unsupervised way.

In this paper, we introduce a novel approach to this problem, in which a band-wise harmonic/noise decomposition is used to localize and separate drum sources. The proposed system has several characteristics which make it stand apart previous works. Firstly, it does not require a prior learning phase. Secondly, it can process a wide range of music signals with the same set of parameters. Thirdly, it can operate causally - allowing our system to work as an audio effect. Finally, no information is lost in the analysis/synthesis stage, enabling the remixing or substraction of the extracted drum track. This paper is organized as follows: section 2 presents the overall principle of the algorithm, and details each of its components. Following a section 3 presenting the evaluation results, section 4 suggests some conclusions.

## 2. DESCRIPTION OF THE ALGORITHM

The principle of our algorithm is to use a band-wise harmonic/noise decomposition to obtain the stochastic part of the signal in different frequency bands.

Our method is motivated by two observations:

1. In popular music, in which drums are overly present, the stochastic part of the signal is mostly contributed by the drum sounds. Other stochastic components, such as note attacks, can usually be discarded using a simple thresholding approach.
2. Unpitched percussive sounds, for example those produced by the bass drum, the snare drum or the cymbals, have a strong non-harmonic component.

Detection signals for 3 categories of instruments of the drum kit (bass drum, snare drum, and cymbals) can be easily computed from this decomposition. The drum track can thus be reconstructed by reweighting the stochastic signals in each band.

The general overview of the drum track extraction system is presented in figure 1. It consists in:

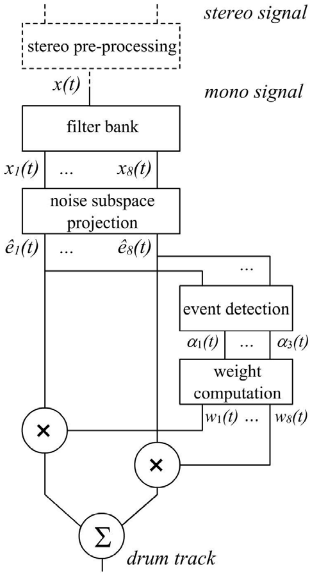

Figure 1: *Overview of the drum track extraction system.*

- **An optional pre-processing stage**, aiming at extracting a monaural signal $x(t)$ with an enhanced rhythmic content from a stereo signal.
- **A filter-bank**, decomposing the signal $x(t)$ into eight non-overlapping subbands $x_k(t)$, $k = 1..8$.
- **A noise subspace projection stage**, extracting the stochastic part of each subband signal $e_k(t)$.
- **A drum event detection stage**, extracting masks $\alpha_l(t)$ for three basic categories of drum instruments ($l = 0..2$).
- **A drum resynthesis stage**, reweighting the stochastic signals in each band with weights $w_k(t)$ computed from the masks $\alpha_l(t)$.

### 2.1. Pre-processing of stereo signals

In the case of signals of popular music, we observed that in a rather large number of cases, a signal with an enhanced percussive content could be obtained by simply mixing the left and right channels of the recording with appropriate gains. This can be accounted by the fact that the so-called "panoramic" mix is widely used in popular music recordings.

Thus, our approach consists in selecting gains $\gamma_1$ and $\gamma_2$ maximizing an impulsiveness measure I of the remixed signal. This impulsiveness measure should favour signals with a sharp and contrasted envelope, which is typical of drum tracks. Assuming that $\gamma_1 \neq 0$, the problem is equivalent to finding $\beta$ maximizing $s_\beta(t) = I(L(t) + \beta R(t))$, where $\beta = \frac{\gamma_2}{\gamma_1}$.

The impulsiveness measure I is computed as follows: firstly, we obtain an envelope signal $s'(t)$ by half-wave rectifying, decimating, low-pass filtering, and differentiating $s(t)$. Then we compute a contrast factor on this envelope:

$$I(s) = \frac{\sum_{t=1}^{T} s'(t)}{T \sqrt[T]{\prod_{t=1}^{T} s'(t)}}$$

The output of our pre-processing stage is finally $x(t) = s_{\beta^*}(t)$ where $\beta^* = \text{argmax} I(s_\beta(t))$.

We also considered using the opposite of the kurtosis of $s(t)$ as an impulsiveness measure. The value of $\beta^*$ obtained by both approaches were comparable.

### 2.2. Filter bank

The use of a filter bank is essential for three reasons. Firstly, each drum instrument has its own characteristic frequency band: it is thus easier to detect drum events in each sub-band signal. Secondly, the noise subspace projection stage performs better when such a decomposition is used, especially with octave-band filter banks. This can be explained by the fact that noise in each narrow frequency band can be considered as white - while it is not always the case on the whole frequency range. Moreover, tracking a fixed number of sinusoids per octave is well adapted to mixtures of harmonic signals. Finally, a polyphase implementation of the filter bank [7], in which the signal in each frequency-band is downsampled, reduces the amount of data to process, and thus the computational cost of the noise subspace projection.

The filter bank used in our system has a dyadic structure, also known as octave-band filter bank, which consists in splitting the source signal in two equal bands (the signal in each band being downsampled by a factor of 2), and then iterating the process on the lower band, until $M = 8$ components are obtained - each resulting frequency band being one octave large. This filter bank was implemented using a 100th order FIR filter as a prototype.

It is worth to mention that this filter bank allows a perfect reconstruction of the signal, a property which is not important in transcription or descriptor extraction applications, but which is essential for the resynthesis of the drum track, and for remixing applications.

### 2.3. Noise subspace projection

The noise subspace projection stage is based on the *Exponentionally Damped Sinusoidal* (EDS) model [8]. According to this model, the signal can be decomposed in a *harmonic* part, modelled as a sum of sinusoids with an exponential decay; and a *noise* part defined as the difference between the original signal and the harmonic part.

Different approaches are possible for the extraction of the sinusoids. We chose a subspace-based technique which overcomes the resolution limit of the Fourier analysis. According to this approach, windows of length $L$ are extracted from the original signal. A vector is defined for each window. The $L$-dimensional space containing this vector is split in a space of dimension $p = 2n$ containing the signal part, and a space of dimension $L - p$ containing the noise; where $n$ is the number of tracked sinusoids. An interesting property of such approaches is that the estimation and substraction of sinusoids is not even required. If $\mathbf{U}^S(t)$ is an orthonormal basis spanning the signal subspace, a noise vector $e(t)$ can be obtained by applying the noise subspace projector $\mathbf{I}_L - \mathbf{U}^S(t)\mathbf{U}^S(t)^T$ to the vector $x(t)$. An entire signal can be processed using an overlap-add method. In our case we use an overlap of $3L/4$ and a Hanning window.

Several algorithms are available to track the signal subspace basis $\mathbf{U}^S(t)$. We used the classical EVD algorithm ([8]) with a FFT implementation of matricial products, the complexity of which is $\mathcal{O}(p^2 L \log L)$ but other subspace trackers can also be used.

The window size used in our system is 46ms. Shorter window sizes resulted in a very instable tracking, as most of the subspace trackers perform better when p is much smaller than L. With longer window sizes, only very stable components are tracked: unwanted information such as glissandi or short notes appear in the noise subspace.

The number of sinusoids in each frequency band was manually adjusted. Two sinusoids are used for $x_1(t)$ (lowest frequency band, in which only the bass is playing), five for $x_2(t)$, ten for $x_3(t)$ and $x_4(t)$; and eight for the other bands. However, it is worth to mention that the number of sinusoids in each band could be automatically selected by appropriate methods [9] but this approach was not tested in this paper.

A noise signal $e_k(t)$ is thus obtained for each sub-band. Because of the multirate implementation of the filter bank, these signals need to be resynchronized in time, by upsamling them and by applying a synthesis filter. This results in eight $\hat{e}_k(t)$ signals.

### 2.4. Drum event detection

A global noise signal can be reconstructed by summing the subband noise signals $e_k(t)$. However, this signal is not a good candidate for a drumtrack, as it contains not only drums and percussive signals, but also the attacks of the notes from other instruments. It can be observed that these attacks have a lower level, and that the percussive signals are localized in well-defined frequency bands. This suggests an approach in which the amplitude of each noise signal is modulated by time-varying masking signals, corresponding to the contributions of various drum instruments.

We define three basic categories of drum instruments, *bass drum*, *snare drum* and *cymbals*. For each of these instruments ($i = 0..2$), a detection signal is computed from the sub-band noise signals:

$$d_i(t) = \sum_k a_{ki}\hat{e}_k(t)$$

The coefficients $a_{ki}$ are chosen so that very characteristic frequency sub-bands will be used for each instrument. For example, for the detection of the bass drum, we combine the noise signal in the first two sub-bands, for the snare drums, the noise signals in the next two sub-bands, and the last sub-band is used for the detection of the cymbals.

A downsampled and half-wave rectified version of $d_i(t)$ is then computed and is noted $d'_i(t)$. Occurrences of each drum instruments are detected by comparing $d'_i(t)$ to a threshold. An occurrence is detected whenever $d'_i(t) > 2\sigma_{d'_i(t)}$, where $\sigma$ is the standard deviation. The duration of each occurrence of a drum instrument is also estimated from $d'_i(t)$. This allows the definition of a mask $\alpha_i(t)$, which is equal to one over the duration of each occurrence of the instrument $i$, and zero otherwise (refer to figure 2 for signal examples).

### 2.5. Synthesis

Subsequently, each sub-band noise signal is modulated according to modulation coefficients $w_k(t)$. These coefficients are computed from the masks $\alpha_i(t)$ with the following rule:

$$w_k(t) = \max_i r_{ki}\alpha_i(t)$$

The contribution coefficients $r_{ki}$ are chosen to reflect the fact that each drum instrument has a characteristic frequency band. For example, when we have detected that only a cymbal is playing, we cancel the sub-band noise signals corresponding to the lowest frequency bands. The coefficients $r_{ki}$ are different from the coefficients $a_{ki}$: while the $a_{ki}$ are only selecting a typical and exclusive frequency band associated to each instrument, in order to provide a precise detection signal, the $r_{ki}$ also includes other frequency band to improve the quality of the synthesis. More precisely, the first 6 frequency bands are used for the bass drum ($r_{k0} = [1, 1, 1, 1, 1, 1, 0, 0]$), all but the first band are used for the snare drum ($r_{k1} = [0, 1, 1, 1, 1, 1, 1, 1]$) and the last four bands are kept for cymbals ($r_{k1} = [0, 0, 0, 0, 1, 1, 1, 1]$). The overlapping of drum instruments is handled by using a max rule.

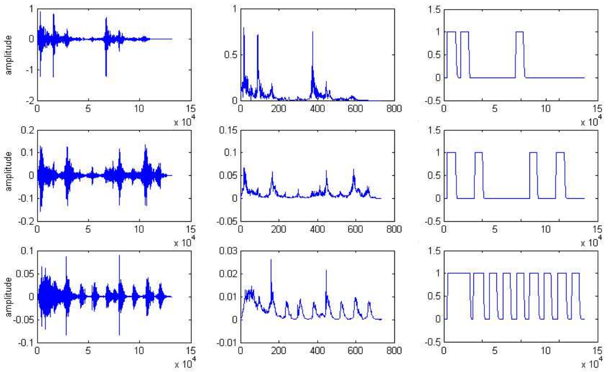

Figure 2: *From left to right: detection signal $d_i(t)$, modified detection signal $d'_i(t)$, and mask $\alpha_i(t)$, computed for bass-drum (first line), snare-drum (second line) and cymbals (third line).*

Finally, we can reconstruct the drum track by modulating each sub-band signal:

$$\text{drums}(t) = \sum_k w_k(t)\hat{e}_k(t)$$

A peculiarity of this whole approach is that contrary to methods based on spectrograms, in which phase information is lost and must be re-estimated, or template-based methods in which synthesis is performed with an averaged template, no phase information is lost during the analysis/synthesis process. This property of our system allows the extracted drum track to be added or subtracted to the original signals to efficiently enhance or attenuate the drums in a music signal.

Audio examples of extracted drum tracks and remixes are available at http://www.tsi.enst.fr/∼gillet/waspaa05_drums_demo.html. All these examples were obtained on mono signals, since the sole stereo processing stage could in some cases directly extract the drum track.

## 3. EVALUATION

While some measures have been suggested for the evaluation of audio source separation in [10], their use is not straightforward when the original monophonic sources are not available. We have chosen to evaluate our drum signal separation approach by assessing the quality of its remixing capabilities - more precisely by assessing the naturalness quality of new remixed signals in which the drum track is either attenuated or amplified by 6dB. This leads us to set up a specific subjective listening test. For this purpose, we have selected 10 monophonic test signals of 15 seconds amongst a database of 55 signals. These test signals gather the best and worst cases for the algorithm as spotted by the authors. For each signal, 4 pairs of stimuli A-B are built:

- $(Orig, mix_{0.5})$ and $(mix_{0.5}, Orig)$ where $Orig$ is the original signal and $mix_{0.5}$ is the remixed signal with an attenuated drum track (-6dB)
- $(Orig, mix_2)$ and $(mix_2, Orig)$ where $Orig$ is the original signal and $mix_2$ is the remixed signal with an amplified drum track (+6dB)

These 40 pairs of stimuli are shuffled and presented sequentially to the subjects. All subjects were asked to listen to the audio signals using high quality headphones and were asked to assess the naturalness of the second signal (signal B) compared to the first signal (signal A) by using a seven grade scale, ranging from -3 to +3. Twelve subjects participated in the experiment.

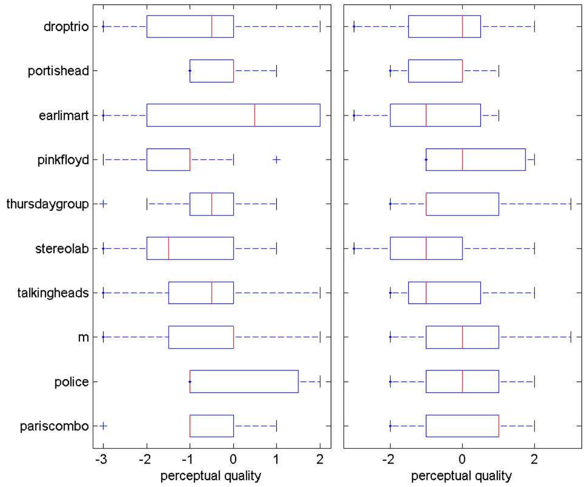

Figure 3: Perceptual quality of test signals for the drum attenuation (left) and amplification (right) tasks

Results for the drum attenuation (-6 dB) and amplification (+6 dB) tasks are summarized in figure 3. The extremities of each box indicate the position of the lower and upper quartiles. The median is marked by a line across the box. It can be seen that the quality of remixes with an amplified drum track is preserved - in one example, the remixed signal is even perceived as more natural, while the algorithm does not succeed in producing natural results for the drum track attenuation task. This can be accounted by the fact that the transient part of some instruments, or the consonants of the vocals are subtracted when they occur at the same time as drum strokes. This results in an effect that most subjects described as similar to mp3 compression artifacts.

## 4. CONCLUSIONS AND FUTURE WORKS

The extraction and the remixing of the drum track of a polyphonic music signal has many real-world applications. Classical rhythm transcription or source-separation approaches are not always suitable for this task. In this paper, we have presented a novel method based on a sub-band harmonic / noise decomposition. The stochastic part of this decomposition is used to efficiently detect drum events and to resynthesize the drum tracks. Resulting signals preserve the rhythmic contents of the original signal, and can be used for high-quality remixes.

While promising results are obtained, several improvements are nevertheless possible. First of all, the stereo pre-processing step presented in 2.1 could be improved by using more advanced separation algorithms. Actually, the extraction of the drum track from a stereo pair could be viewed as an optimization problem - optimizing a quality/percussivity measure on one of the output of a complex source separation algorithm. Secondly, the robustness of the noise subspace projection itself could be improved with order selection methods that would automatically select an optimal number of sinusoids. Finally, efforts have to be made towards a more complete evaluation. This includes the evaluation of the algorithm as a pre-processing stage for a drum transcription task and a subjective evaluation test of the quality of extracted drum signals.

## 5. ACKNOWLEDGEMENTS

This work was partly supported by the MusicDiscover project of the ACI Masse de données.

## 6. REFERENCES

[1] A. Zils, F. Pachet, O. Delerue, and F. Gouyon, “Automatic extraction of drum tracks from polyphonic music signals,” in *Proc. of WEDELMUSIC2002*, Dec. 2002.

[2] K. Yoshii, M. Goto, and H. G. Okuno, “Automatic drum sound description for real-world music using template adaptation and matching methods,” in *Proc. of the 5th Int. Conf. on Music Information Retrieval (ISMIR 2004)*, Oct. 2004.

[3] V. Sandvold, F. Gouyon, and P. Herrera, “Percussion classification in polyphonic audio recordings using localized sound models,” in *Proc. of the 5th Int. Conf. on Music Information Retrieval (ISMIR 2004)*, Oct. 2004.

[4] D. Barry, B. Lawlor, and E. Coyle, “Sound source separation: Azimuth discrimination and resynthesis,” in *Proc. of the 7th Int. Conf. on Digital Audio Effects (DAFX’04)*, Oct. 2004.

[5] E. Vincent and X. Rodet, “Underdetermined source separation with structured source priors,” in *Proc. of the 5th Symposium on Independent Component Analysis and Blind Signal Separation (ICA2004)*, Apr. 2004.

[6] D. FitzGerald, B. Lawlor, and E. Coyle, “Prior subspace analysis for drum transcription,” in *Proc. of the 114th AES Convention*, Mar. 2003.

[7] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[8] R. Badeau, R. Boyer, and B. David, “Eds parametric modeling and tracking of audio signals,” in *Proc. of the Conf. on Digital Audio Effects (DAFX-02)*, Sep. 2002.

[9] R. Badeau, B. David, and G. Richard, “Selecting the modeling order for the esprit high resolution method: an alternative approach,” in *Proc. of the 2004 Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP’04)*, May 2004.

[10] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, “Proposals for performance measurement in source separation,” in *Proc. of the 4th Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Apr. 2003.