Automatic Extraction of Rhythmic Structure From Music

François Pachet, Olivier Delerue, Fabien Gouyon
Sony CSL – Paris, 6, rue Amyot 75005 Paris
E-mail : {pachet, delerue, fgouyon}@csl.sony.fr

Abstract

We propose an approach for extracting automatically time indexes of occurrences of percussive sounds in an audio signal taken from the popular music repertoire. The scheme allows to detect percussive sounds unknown a priori in a selective fashion. It is based on an analysis by synthesis technique, whereby the sound searched for is gradually synthesized from the signal itself. The extracted information can then be used to build high level representations of the rhythmic structure of the music such as the tempo, rhythm type, or for similarity-based search in large music catalogues.

1. Introduction

We address the issue of automatically extracting musical descriptors from audio signals, to be eventually used in content-based music access applications such as in the context of MPEG 7. Most of the work on audio descriptors focuses on 1) low-level or mid-level descriptors (see e.g. [1]) and 2) small or middle sized audio data, typically sounds. In our project, we focus on high-level descriptors that describe music titles in a global fashion. Such global musical descriptors typically include tempo (see e.g. [2]), global energy, but also musical genre, rhythm type, etc. Moreover, we target applications dealing with popular music, in which rhythm is a predominant feature. This paper addresses precisely the issue of extracting rhythmic information automatically from musical signals.

Although rhythm is acknowledged to be a fundamental dimension of music perception, it is a poorly understood phenomenon –even if some works regarding the automatic transcription of percussive music exist; and in particular there is no reference representation of rhythm that can be used, e.g. for classification. To produce such a representation, we need to extract from the audio signal occurrences of percussive sounds – the reason being that popular music titles generally have a rhythmic dimension that is mainly given by the drum sounds. The problem we address is therefore the following: given an audio excerpt of music title, find the occurrences of significant percussive sounds. Furthermore, we consider rhythm as being produced not only by occurrences of such sounds, but also by the interplay with different sounds (such as bass drum and snare drum). We therefore want to extract a set of time series, each time series representing the occurrences of one particular percussive sound in the signal. There are two main issues to address this problem. First, percussive sounds do not fit well with spectral models of sounds, because of their inherently non stationary nature. Second these sounds are not known a priori. The only hypothesis we can make are therefore:

- we look for short, non stationary sounds in the signal
- these sounds are repeated

2. The Scheme

To detect occurrences of sounds, we have designed a scheme based on the progressive identification of the source sound (the percussive sound to find) during the analysis process. More precisely, the scheme is the following:

We start by taking a simple, synthetic sound, such as a band filter impulse response. We then look for occurrences of this sound in the signal, using a correlation technique. We then apply various sets of filters to determine which occurrences actually denote the "same percussive sounds". At this point, the system may have found some occurrences of percussive sounds, but because the correlation is performed with a general sound, it may have missed a number of occurrences. Finally, we synthesize a new sound from the occurrences found, and repeat the process.

This iterative scheme is repeated until a fixed point is reached, i.e. the occurrences found are the same than in the preceding cycle. Let us briefly review each of these steps

2.1. Looking for a percussive sound in the signal

The first step is to detect all possible occurrences of the percussive sound in the signal. This step is performed by computing the correlation function $Cor(\partial)$ defined as follows:

$$Cor(\partial) = \sum_{t=1}^{N_I} S(t) \times I(t-\partial) \quad \text{with for } \partial \in [1, N_S]$$

for a signal $S(t)$, where t belongs to [1, N_S] with an instrument sound $I(t)$, with t belongs to [1, N_I]. Since we

do not want to keep phase information, we consider only the absolute value of this function.
The function yields a kind of similarity measure between the two signals. Consequently, possible occurrences of the percussive instrument sound $I(t)$ in the original signal $S(t)$ will appear as peaks in this correlation function, as represented in Figure 1. An evaluation of the use of correlation for detecting occurrences of percussive sounds is described in Section XXX.

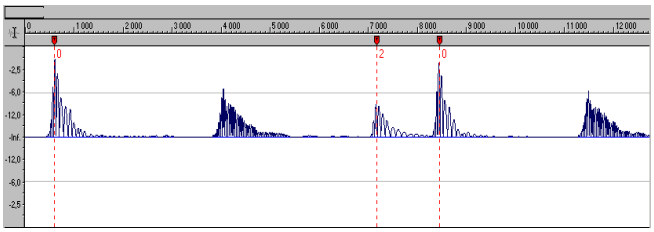

Figure 1. The correlation function applied between a musical excerpt signal and a percussive signal.

However, all the peaks of $\left| Cor(\partial) \right|$ are not necessarily relevant, because of the very nature of the correlation function, when applied to short sounds. For instance, if the percussive sound to detect is a snare drum sound, there may also be peaks due to the occurrence of a louder bass drum sound. To avoid collecting peaks which may not be related to the percussive sound, we perform a series of filters described below. These filters are designed to filter out peaks which do not correspond to the same percussive sound. Furthermore, these filters may be over restrictive, that is, may filter out peaks actually corresponding to "good" occurrences, thanks to the iterative scheme: a good peak which was not selected at some cycle of the scheme may still be in a further cycle, when the percussive sound used for correlation has become more precise.

2.2. Assessing peak quality

There are various ways we can assess the quality of a peak in the correlation signal.
First, we trim away the peaks whose amplitude is less than a certain threshold (60% of the amplitude of the highest peak for instance), assuming that the good peaks have a minimum energy.
Second, we filter out peaks which are too close to each other (with a given threshold of 10 ms). This allows to avoid situations where a series of close peaks have been detected corresponding to a unique occurrence of a sound (typically the case when the signal has been processed with echo or reverberation effects).
Third, we need also to somehow assess the "quality" of the peak, independently of its amplitude. In Figure 1, only peaks 1, 3 and 4 should actually be selected, although peak 3 has a less important amplitude than peak 2. Indeed, the amplitude of a peak in the correlation signal depends on two factors: 1) the actual similarity of the sound I with the signal, and 2) the amplitude of the signal. To take again the example used above, a occurrence of a signal similar to the instrument sound *I* but with low gain will yield the same peak amplitude in the correlation function than an occurrence of a less similar sound but with a higher gain.
To discriminate between these cases, we have studied several quality measures for peaks.
The first quality measure consists in imposing thresholds on the relative on relative local energy of the peak, as follows:

$$Q(Cor,t) = \frac{Cor(t)^2}{\frac{1}{picWidth}\sum_{i=t-\frac{picWidth}{2}}^{t+picWidth} Cor(i)^2}$$

A refinement of this measure consists in taking into account the actual energy in the signal itself at the same time *t*, to compensate for the effect described above:

$$Q_2(Cor,t) = \frac{Q(Cor,t)}{E_{picWidth}(S,t)}$$ where $E_{picWidth}(S,t)$ is the local energy in the musical signal S at time t, in a window of size *picWidth*. This refined measure is under study.

An alternative way of selecting peaks has also been studied, by incorporating knowledge about percussive sounds (see [3]). These studies concluded on the relevance of the zero-crossing rate for distinguishing between two main classes of percussive sounds: bass drum-like sounds and snare drum-like sounds.

These various criteria allow to filter out bad occurrences. In any case, at this point some occurrences of the instrument I may have been missed, because of the under-specificity of I. We now look for a new sound, based on these filtered occurrences.

2.3. Synthesizing a new sound

The core idea of our approach is to synthesize a new sound I', based on the results of the preceding steps. Because of their inherent non stationarity and short duration, percussive sounds do not fit well with spectral synthesis models. Thus we perform synthesis in the temporal domain, by mixing portions of the signal centered around the occurrences, and also from the initial sound *I* itself.
An approximation of the synthesis performed is the following:

$$newInst(t) = \frac{1}{nbPeaks}\sum_{i=1}^{nbPeaks} S\left(peakPosition(i) + t\right)$$

In practice, the synthesis also comprises a step of centering and synchronization of occurrences, to emphasize redundancy between occurrences.
This new sound is therefore made only from ingredients taken from the signal, without any external input but the starting band pass filter impulse response. It is designed so as to be a refinement of the initial sound *I*, refinement in the sense that it may be used to find some missed occurrences.

2.4. Repeating the scheme until fixed point

We now repeat the scheme starting from I' instead of I, i.e; look for occurrences of I' in the signal, filter out bad peaks, and synthesize a further new sound I'', and so forth. At each cycle, the new sound synthesized is, by construction, "closer" to an actual percussive sound occurring in the signal.
The process is repeated until a fixed point is reached, that is, no new occurrences are found. The process also stops when a maximum number of cycles is reached, to avoid infinite loops in cases of diverging synthesis. In practice, the maximum number of cycle has been set to 4.

3. Extracting dual time series

3.1. Using the mechanism to extract rhythmic information

3.2. Influence of starting input signal I

The result of the process depends heavily on several factors, including the nature of the starting instrument signal I. Since we eventually aim at extracting at least two different time series for a given musical signal, we look for two starting sounds corresponding approximately to the two main classes of percussive sounds found in popular music: bass drum sounds and snare drum sounds.
The experiments we conducted, in particular the systematic study of bass drum and snare drum sounds described below, show that the use the of correlation technique does allow to discriminate between families of sounds such as bass drums and snare drums. Therefore, even if we are not able to characterize precisely the relation between the starting sound and the time series obtained, we observe in practice that the synthesized sounds remain in the same family as the starting sound.
Therefore, starting with a low-pass filter impulse response tends to yield occurrences of bass drum sounds. High-pass filter impulse response tend to yield occurrences of other sounds such as snare drums or hi hats.

4. Evaluation

We have conducted several types of evaluation of our scheme: 1) evaluation of the correlation method to yield occurrences of percussive sounds and 2) evaluation of the quality of the extracted time series.

4.1. Evaluation of the correlation method for extracting percussive sounds

4.1.1.1 Expériences

Pour tester la validité de l'utilisation de la fonction de corrélation, nous avons effectué la progression d'expériences suivante :

1. Corrélation d'une séquence de son monophoniques et non-bruités issus d'un synthétiseur (banque General MIDI du Korg05RW) avec les sons de cette séquence
2. Corrélation d'une séquence, transformée par ajout d'effets, de son de synthétiseur monophoniques et non-bruités avec les sons de la séquence non-transformée
3. Corrélation d'une séquence de sons réels (donc bruités par un environnement polyphonique) avec ces mêmes sons
4. Corrélation d'une séquence de sons réels transformée (donc bruités par un environnement polyphonique) avec les sons non-transformés

4.1.1.1.1 Expérience 1

On génère une séquence de 44 sons avec des sons percussifs monophoniques et non-bruités.

NB : Les sons 1 à 5 correspondent à des grosses caisses, les sons 6 à 13 sont des caisses claires, 14 à 18 sont des charlestons, il y a ensuite des toms, des cymbales, des cloches et diverses percussions.

On calcule la valeur absolue de la corrélation de cette séquence avec chacun des sons la composant, et on mesure les hauteurs des pics de corrélation.
On note $c_k(l)=R_{xy_k}(n_l)$, $c_k(l)$ est la hauteur de la fonction de corrélation entre les signal $x(n)$ et le signal $y_k(n)$, à l'indice temporel correspondant au son numéro l.
NB : n est un indice temporel ($\in[1,N]$) ; et k, l, i et j sont des indices de numéro de sons ($\in[1,44]$).
On définit une mesure normalisée (entre 0 et 1) de proximité entre sons à partir des hauteurs des pics de corrélation, cette mesure est la suivante :

$$d_k(l)=\frac{c_k(l)}{\sqrt{c_k(k)}\times\sqrt{c_l(l)}}$$

Équation 4.1—1

Elle a été construite afin de vérifier les propriétés suivantes :

- $d_k(k)>d_k(j) \quad \forall j,k\in[1,44]$
- $d_k(j)=d_j(k) \quad \forall j,k\in[1,44]$
- $d_k(k)>d_i(j)\geq 0 \quad \forall k,(i,j \text{ tel que } i\neq j)\in[1,44]$
- $d_k(k)=d_j(j)=1 \quad \forall j,k\in[1,44]$

On dispose donc d'une mesure nous permettant de comparer les proximités des sons entre eux.
Les mesures peuvent se résumer en une matrice carrée symétrique, se prêtant bien à la représentation, comme le montre le graphique suivant.

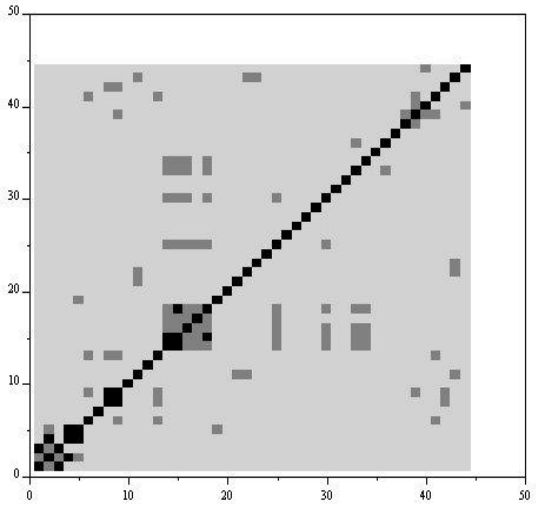

Figure 4.1—1 : Mesure de proximités entre sons de la banque GM du Korg05RW

Pour visualiser cette notion de proximité, nous fixons deux seuils en amplitudes (e.g. 0.8 et 0.35), et attribuons une couleur pour les pics dépassant chaque seuil : les carrés noirs correspondent à une proximité $d_k(j)$ entre les sons d'indices k et j supérieure à 0.8, ceux gris foncés à une proximité comprise entre 0.35 et 0.8, et enfin, les gris clairs à une proximité comprise entre 0 et 0.35.
En observant la diagonale sur ce graphique, on se rend bien compte que la corrélation entre un signal et un son de référence nous permet de retrouver ce son dans le signal. De plus, il semble que, en adaptant judicieusement le nombre et les hauteurs des seuils, apparaisse un regroupement des sons par famille de percussions (e.g. les grosses caisses correspondent à des mesures toutes comprises entre 0.8 et 1).

4.1.1.1.2 Expérience 2

La même séquence est transformée violemment par ajout de distorsion et de réverbération.
La même mesure de proximité entre sons est effectuée, à partir de la valeur absolue de la corrélation de la séquence transformée et des sons originaux.
Une représentation graphique du résultat est la suivante.

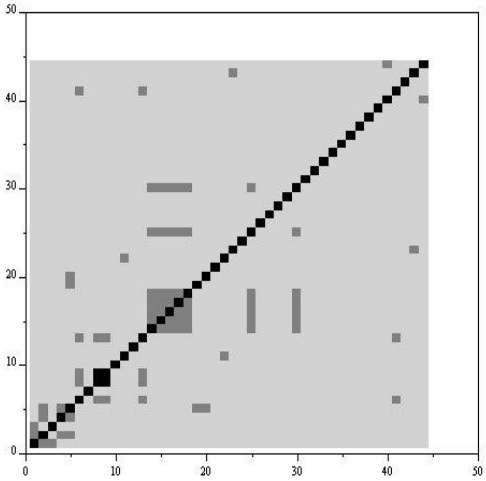

Figure 4.1—2 : Mesure de proximités entre sons de la banque GM du Korg05RW originaux et transformés

Le phénomène de regroupement apparaît un peu moins que dans le cas où aucune transformation n'est appliquée, mais la détection du son de référence est toujours effectuée.

4.1.1.1.3 Expérience 3

On s'intéresse maintenant aux sons réels, i.e. extraits directement de morceaux de musique. Les sons de percussions peuvent toujours être considérés comme étant monophoniques, mais ils sont plongés dans un environnement polyphonique qui, de notre point de vue, peut être considéré comme du bruit.
Dans l'exemple suivant, il s'agit de la mise en séquence de 8 occurrences de caisse claire et de 10 occurrences de grosse caisse directement extraites du morceau "High times".

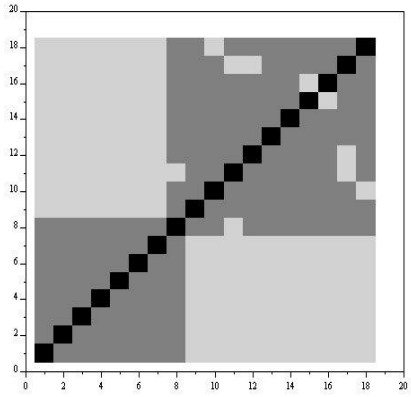

Figure 4.1—3 : Mesure de proximités entre diverses occurrences de grosse caisse et caisse claire dans "High times"

On s'aperçoit que la détection de chaque son est effective – le contraire serait inquiétant car les sons de la séquence et ceux de références sont exactement les mêmes. De plus, le phénomène de regroupement en famille de son est là encore présent : les grosses caisses correspondent toutes à une proximité élevée, c'est la même chose pour les caisses claires.

4.1.1.1.4 Expérience 4

On s'intéresse toujours aux sons réels.
Dans l'exemple suivant, il s'agit de la mise en séquence de 2 occurrences de caisse claire et de 3 occurrences de grosse caisse directement extraites du morceau "By and Bye".
NB : les grosses caisses correspondent aux indices 2, 3 et 4.

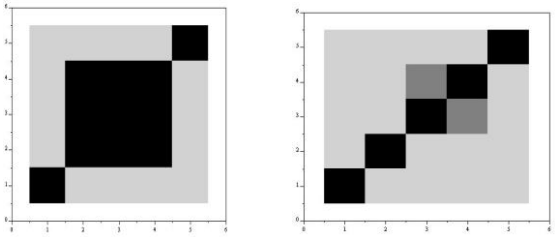

Figure 4.1—4 : Mesure de proximités entre diverses occurrences de grosse caisse et caisse claire dans "By and Bye", sans puis avec transformation de la séquence

Dans l'exemple sans transformations, les grosses caisses donnent toutes des mesures de proximité proche de 1. On voit bien que la proximité diminue entre les grosses caisses lorsqu'une transformation violente est appliquée à la séquence.

4.1.1.2 Performances d'une mesure de proximité basée sur la corrélation

Les expériences précédentes montrent que prendre comme critère de proximité la hauteur des pics de corrélation entre une séquence et les sons la composant est une méthode efficace de détection d'un son dans une séquence.
Plus important, la cohérence de cette mesure est robuste à des transformations (même importantes) des sons de la séquence.
Cependant, le problème important qui se pose est celui de la détermination de seuils ad hoc pour pouvoir déterminer le regroupement de divers sons.
De plus, dans le cadre réel d'application de notre algorithme de détection, nous ne disposons pas (au moins lors des premières itérations) d'un son assez proche de celui recherché dans le signal, et qui permettrait de se contenter d'utiliser simplement une mesure de hauteur de pics de corrélation. Les expériences précédentes suffisent uniquement à nous conforter dans l'idée que l'utilisation de la fonction de corrélation est sensée.
Dans la réalité de notre algorithme, la première corrélation est effectuée entre le signal musical et un son de référence, ce dernier étant étranger au signal.
L'exemple suivant montre la valeur absolue de la fonction de corrélation normalisée d'un extrait musical avec un son de référence générique.

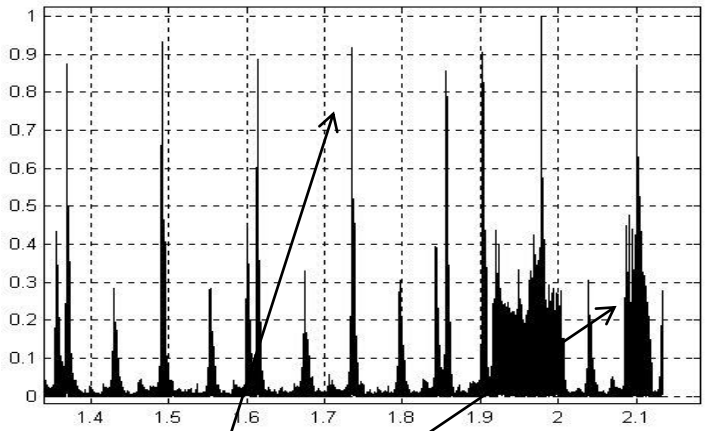

Figure 4.1—5 : Extrait de la valeur absolue de la corrélation entre "A night to remember" et le son de référence "filter1"

Sons de grosse caisse

Son de caisse claire

On voit bien que dans les cas ou l'on ne regarde que la hauteur des pics de corrélation, on court le risque soit de ne pas détecter des occurrences de grosse caisse (si on fixe un seuil trop élevé), soit de commettre des erreurs de détection (si le seuil est bas) et prendre en compte des sons de caisse claire, mais aussi d'autre sons perturbateurs.

Finalement, un critère basé sur la hauteur des pics de corrélation entre le signal et un son de référence à caractère percussif nous permet de déterminer une série d'indice temporels auxquels sont présents quelques sons percussifs du signal. Il faut fixer un seuil en amplitude pour déterminer ces indices. Selon le son de référence initial, la détection sera orientée vers tel ou tel type de son. Avec une valeur de seuil fixée, il faut prendre en compte les deux problèmes suivants[1] : il existe un risque de ne pas détecter des occurrences du son recherché ; et il est probable que des occurrences de sons parasites ne soient pas "filtrées" par la corrélation. Le premier problème est abordé au paragraphe **Erreur ! Source du renvoi introuvable.**, et le deuxième dans le prochain paragraphe.

4.2. Evaluation of the quality of extracted time series

We have applied our scheme to a database of 312 excerpts of popular music. Each excerpt is 20 second long on average, with a sampling frequency of 11025 hz. The excerpts are from various musical styles, ranging from pop, country, funk, disco, jazz, as well as some classical music.
First we have performed an auditory evaluation of the extracted time series: for each title, we have built a so-called "drum track" signal, made by sequencing the two synthesized sounds from the couple of peak series. We

[1] Respectivement l'un ou l'autre selon que le seuil est élevé ou faible.

have then played the drum track with the original signal and asked listeners to judge whether the drum track would retain the rhythmic characteristic of the original signal. The result is that 245/312 = 78% of the titles are “correctly” analyzed. Most of the titles (66%) which are not correctly analyzed are titles with no significant rhythmical structure (at least in the sense we have followed, that is without occurrence of percussive sounds). The remaining 34 % (of the 22% of badly analyzed titles) contain rhythmic information which is either very weak, or buried within intense noise or reverberation.

Among the titles with a correct peak extraction, we have then computed the tempo from the series extracted. This computation is based on an auto-correlation analysis of the peak time series, following [4]. The result is compared to “real tempo” obtained by measuring manual hand clapping. The results are: 196/245, i.e. 80% of correct tempo.

These results are good, considering the extreme generality of the approach taken, i.e. without any external knowledge a priori on percussive sounds.

Current work now focuses on the implementation of a distance measure based on the extracted peak series, for similarity based search. Additionally, application of the approach to automatic segmentation and identification of speech without a priori knowledge is investigated.

5. References

[1] Herrera P., Serra X., Peeters G. (1999): "Audio Descriptors and Descriptors Schemes in the Context of MPEG-7". Proc. ICMC 1999

[2] Scheirer E. (1998): "Tempo and beat analysis of acoustic signals", JASA, 103(1)

[3] Gouyon F., Delerue O., Pachet F. (2000) "Classifiying percussive sounds: a matter of zero-crossing rate ?". Third Digital Audio Effect Conference, Verona (Italy).

[4] Brown J. (1993) "Determination of the meter of musical scores by autocorrelation". JASA 94(4)