Proceedings of the 2nd International Conference on Web Delivering of Music(WedelMusic2002), Darmstadt, Germany, Dec. 9-11, 2002

Automatic Extraction of Drum Tracks from Polyphonic Music Signals

Aymeric Zils, François Pachet, Olivier Delerue, Fabien Gouyon
Sony CSL – Paris, 6, rue Amyot 75005 Paris
E-mail : `{zils, pachet, delerue, gouyon}@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 is able 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 possibility to extract different types of percussive sounds and their occurrences in the audio signal makes it possible to build a drum track representing the essential rhythmic component of a music piece. We present a systematic evaluation of the performance of our algorithm on a database of popular music titles. The system performs well on most of the cases (over 75%). We analyze the reasons for failure on the remaining cases, and propose solutions for yet improving the algorithm. The extracted percussive sounds and drum track serves as a basis for search by rhythmic similarity in the context of the European project Cuidado.

1. Introduction

The recent development of efficient digital audio compression techniques together with the widespread use of Internet is about to create a situation in which millions of users have access to millions of music titles. This situation calls for efficient content-based management techniques, without which the access to large music catalogues remains basically a fantasy.
The development of content-based management techniques has been recently boosted by works on musical metadata. Metadata, seen as descriptors of music are indeed envisaged by many as a crucial ingredient of musical management systems. Many works have been devoted recently to the issues of extracting various types of musical metadata, especially in the context of Mpeg7. For instance, these works focus on the automatic transcription of music by detecting the melodies (see [1]), or on the automatic extraction of the tempo (see [2]).

The main interest of metadata in our context is that metadata forms the basis of sophisticated browsing tools, such as automatic classification or similarity based search.

In this context, Popular music represents the huge majority of electronic music distribution. It is commonplace to note that rhythm plays a very important role in Popular music, both in the way it is perceived and produced. Moreover, we are interested in extracting metadata from music, as it is found, that is in polyphonic recordings. It is therefore natural to look for ways of representing and extracting rhythm in such recordings.

Rhythm is not a fully understood phenomenon, and is therefore difficult to define precisely. The determination of the perceptually relevant dimensions of rhythm is a hard task that has been addressed by many researchers (e.g. [3], [4]). In the computer music field, most of these works focus on the extraction of a beat and of the meter of music titles (see [2], [5], [6], [7]). Although beat and meter are very important features of a music title, they are not sufficient to represent fully the rhythmic dimension of music, and to be used as a basis for content-based search. To give a simple example, a huge majority of the titles of a 10,000 Rock music catalogue will have approximately the same tempo (100) and the same meter (4/4).

If the issue of extracting rhythm information from a polyphonic recording is ill defined in general, we can however, in the context of Popular music, make two important and strong assumptions.
Firstly, the perception of rhythm in Popular music is correlated to the repeated occurrences of accentuated sound along the song. These repetitions occur at different scales: short scale (percussive sounds, successive notes of an instrument), middle scale (sentences, chord patterns), or long scale (verse and chorus). Secondly, rhythm is most often produced by occurrences of percussive sounds, typically a drum set, or any percussive instruments. Many music titles do produce impressions of rhythm although they do not contain any percussive sounds (for example some

folk songs by guitarist-singers like Bob Dylan), but for the vast majority of Popular music titles this is not the case.

In this paper, we propose an approach to extracting rhythmic information, that provides a percussive representation of rhythm, such as the one introduced in [8]. Our method is based on the extraction of the occurrences of percussive sounds in music titles, and exploits fully these two assumptions.

More precisely, the problem we address is the following: given an audio excerpt of polyphonic music, we want to find the occurrences of significant percussive sounds. These occurrences are represented as a set of time series, where each time series represents the occurrences of one particular percussive sound in the signal. Moreover, we also want to identify the percussive sounds themselves, without a priori knowledge. Indeed, since virtually any instrument can produce percussive sounds, it is not realistic to assume that the database of all percussive sounds is available.

2. On the non-stationarity of percussive sounds

Different method can help to detect percussive sounds in audio signals. Firstly, we noticed that the occurrences of important percussive sounds always correspond to local signal energy peaks. This allows to detect the percussive onsets in monophonic signals, but is not sufficient to extract percussive sounds out of polyphonic signals. Indeed, in polyphonic signals, energy peaks correspond not only to percussive onsets, but also to any loud instrument note or sung syllable. An extraction based on the detection of energy peaks would not be precise enough, but the pre-detection of energy peaks in the signal helps to provide a more precise location of the percussive onsets (see 3.6).
Secondly, percussive sounds are short and non-stationary by nature. But non-stationarity does not fit well with traditional spectral analysis and cannot be used as a model of sound. It can only discriminate percussive sounds from the stationary notes of other instruments in monophonic signals. But as polyphonic signals are highly non-stationary by nature, this technique is not efficient for detecting percussive sounds. For instance, experiments on the detection of non-stationary parts in polyphonic signals as transients (see [9]), showed that the method is not precise enough, as it provides not only the percussive sounds, but also the attacks of the notes of all the other instruments.
As traditional signal processing methods are inefficient for the detection of percussive sounds, we needed to build a new approach based on the only hypotheses that we can make, that we are looking for:

- non-stationary sounds: as the exact sounds that we are looking for are not known a priori, the method is based on a very general model of a non-stationary sound,
- short sounds: their length should be less than 100ms,
- repeated sounds: we use the correlation peaks of the signal to detect repetitions of sounds, that are iteratively matched to the audio signal.

However, acoustic drum sounds are known to have a high noise ratio, that makes correlation inefficient. But we can remove a large part of this noise using 2 techniques:

- averaging of the extracted sounds,
- using very short sounds: indeed, the 1st ms of drum sounds correspond to the impulse and the excitation of the 1st vibration modes, that are important compared to the secondary vibration modes, considered as noise.

3. The scheme

The scheme for detecting the occurrences of percussive sounds is based on the progressive identification of the source sound (the percussive sound to be found) during the analysis process. More precisely, it is the following:

- we start by considering a simple, synthetic, percussive sound to be found in the audio signal,
- we look for the occurrences of this sound in the signal, using a correlation technique,
- by applying filters, we determine which occurrences actually denote the “same percussive sounds”: this is an evaluation of the quality of the extraction.

At this point, the system may have found some occurrences of percussive sounds, but it may also have missed some of them, because of the generality of the initial sound.

- then we synthesize a new sound based on an averaging of the percussive occurrences found in the signal.

This scheme is repeated until a fixed point is reached, i.e. the occurrences are the same than in the preceding cycle. At this point, we consider that all the occurrences of percussive sounds have been found, the search is stopped: the system provides the percussive sounds and occurrences found. Let us now review briefly each step of the method.

3.1. Initial synthetic percussive sound

The initial percussive sound I(t) is a basic model of the percussive sound that we want to extract from the audio signal. Typically, we use low-pass filter and band-pass filter impulse responses, that stand for bass-drum-like and snare-drum-like sounds.

3.2. Looking for a percussive sound in the signal

To find the occurrences of the previous percussive sound in the audio signal, we compute the correlation function

$Cor(\partial)$ between the signal S(t), where t belongs to [1, N_S] and the percussive instrument sound I(t), with t belonging to [1, N_I]:

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

The technique is simple and efficient. However, it is very sensitive, by definition, to amplitude. Therefore some peaks in this correlation signal may not correspond to actual similarity between the instrument sound and the signal. Therefore, we introduce a peak quality measure to filter out bad occurrences.

3.3. Assessing peak quality

In order to keep only the most relevant peaks, we assess the quality of correlation peaks using various parameters, by imposing thresholds on the following quality measures:

1. the proximity of the position of the peak with the position of a signal energy peak (see 3.6), that evaluates if the correlation peak corresponds to a percussive peak,
2. the amplitude of the peak in the correlation signal,
3. the relative local energy:

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

Parameters 2 and 3 evaluate how much the signal peak corresponds to the instrument signal.
These various criteria allows us to filter out bad occurrences. However, some good occurrences of the instrument may have been missed, or some false occurrences may have been found, because of the under-specificity of the synthetic percussive sound. To define more precisely the sound to search for, we synthesize a new sound based on the filtered occurrences.

3.4. Synthesizing a new sound

The core idea of our approach is to synthesize a new percussive sound newI(t) 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, and thus synthesis is performed in the temporal domain, by mixing portions of the signal centered around the good occurrences found in 3.3 with the initial sound I(t). An approximation of the synthesis is the following:

$$newI(t) = \frac{1}{2}\left[I(t) + \frac{1}{nbPeaks} \sum_{i=1}^{nbPeaks} S\big(peakPosition(i) + t\big) \right]$$

(This is a simplified formula, that omits the necessary centering and phase synchronization of occurrences)

This new synthetic percussive sound newI(t) is a refinement of the initial sound I(t), made with extracts of the original audio signal, so it can be used to find the missed occurrences.

3.5. Repeating the scheme until fixed point

We now repeat the scheme starting from newI(t) instead of I(t), i.e look for occurrences of newI(t) in the signal, filter out bad peaks, and synthesize a further new sound newnewI(t), and so forth. The process is repeated until a fixed point is reached, or until a maximum number of cycles is reached.
As a result, the system provides a synthetic percussive sound extracted from the audio signal, that is the closest to the initial synthetic sound, together with its time occurrences in the signal.

3.6. Energy peaks preprocessing

We noticed that the occurrences of relevant percussive sounds always correspond to local signal energy peaks.
Therefore, as a preliminary processing in order to reduce the amount of data to look for percussive sounds, we extract the position of the short-term energy peaks in the audio signal, to match them with the positions of the percussive occurrences. Useful to extract the most relevant percussive sounds, that technique is too restrictive to detect secondary percussive sounds in noisy signals.

4. Extension to multiple percussive sounds extraction

In the method we have presented, the resulting percussive sound of depends on the initial synthetic sound given to the system. Thus, running the system with different initial sounds allows to extract different types of percussive sounds out of the audio signal, which can be useful for describing Popular music titles.

4.1. Binary percussive rhythm of Popular music titles

In most of the popular music titles, percussions are drums, and the main drum sounds are the bass drum (low-pitched), and the snare drum (high-pitched). So our idea is to transcribe the drum track of a music title as a sequence of bass-drum-like and snare-drum-like sounds. The rhythm of the title is then described by its 2 most important percussive instruments, and by their respective occurrences.

4.2. Method for full drum track extraction

The method consists in running the system twice, with 2 different initial synthetic percussive sounds: a low-pitched

one, that is the impulse response of a low-pass filter, and a high-pitched one, that is the impulse response of a band-pass filter. In order to avoid difficulties due to simultaneous occurrences of the 2 percussive sounds, priority is given to the bass-drum-like sound.
So the drum track extraction consists in 2 steps:

- a first extraction based on the low-pitched sound provides the occurrences of the most important bass-drum-like sound,
- then a second extraction based on the high-pitched sound provides the occurrences of the most important snare-drum-like sound, that are not conflicting with the previous bass-drum-like occurrences.

4.3. Discriminating between the 2 percussive sounds

We need to introduce a new parameter in order to discriminate between the 2 types of percussive sounds during the extraction. The most relevant parameter for distinguishing between the two main classes of percussive sounds, bass drum-like sounds and snare drum-like sounds, was proven to be the zero-crossing rate, or ZCR (see [10]). So that criterion is introduced in our algorithm, as an additional way of selecting the correlation peaks in the signal (see 3.3.): only the peaks with a correct ZCR are selected.
Finally, as a result, the system provides 2 synthetic percussive sounds extracted from the audio signal, with their time occurrences, and a synthetic audio track representing the drum track of the musical extract.

5. Evaluation

The performances of this process have been evaluated as follows.
We consider a database of 100 musical extracts that are from 10 to 20 seconds of music with percussive rhythm. These extracts are of various genres (rock, pop, dance, jazz, rap), and the percussive sounds are produced by a variety of sources, including drums sounds (bass, snare) but also African percussions (djembe), or electronic percussive sounds (synthetic, claps), etc. We have performed a manual classification of the titles of our database, based on the supposed difficulty to extract their drum tracks:

- 20% of these extracts contain predominant drum sounds mixed with other quiet instruments or voices, like some pop acoustic songs with loud drums. These titles are therefore considered to have an *easily extractable* drum track,
- 60% contain percussive sounds organized in a rhythmic structure, but equally mixed with other instruments or voices. This represents the majority of the Popular music titles. These extracts are considered to have a *possibly extractable* drum track,
- 20% of the titles, for which the percussive sounds and their structure are not obvious or are very quiet compared to the other instruments, like in some jazz, folk or noisy songs, are considered a priori to have a *hardly extractable* drum track.

The test consisted in extracting for each of these extracts a drum track made of 2 different percussive sounds (bass-drum-like and snare-drum-like), and in evaluating how well this drum track fits with the original title.
We assigned to each resulting drum track one among 4 qualitative levels to evaluate the quality of the extraction:

- a *perfect* drum track extraction perfectly provides the correct occurrences of the 2 main percussive sounds, and provides a rhythmic structure that corresponds to the original title,
- an *acceptable* drum track extraction provides a majority of the occurrences of the 2 main percussive sounds, but can miss some of them or add some false occurrences; however, the global rhythmic structure is still obvious, and the result can be used for further analysis,
- a *half acceptable* drum track extraction finds the occurrences of only 1 out of the 2 main percussive sounds (typically the bass-drum-like one), or finds an acceptable drum track with confused sounds,
- a *bad* drum track extraction does not find the 2 main percussive sounds, or is unable to provide a rhythmic structure that is linked to the original music title; the result is unusable for rhythm analysis.

The results are presented in the following table (depending on the supposed difficulty to extract the drum track):

	Quality of the extraction			
	Bad	Half Acceptable	Acceptable	Perfect
Easy (20% of the database)	5%	15%	25%	55%
Possible (60%)	8%	16%	28%	48%
Hard (20%)	50%	10%	10%	30%

Discussion

The performances are very close between easily and possibly extractable drum tracks: about 50% of perfect drum tracks extractions, and more than 25% of acceptable ones. That is to say that our approach provides more than 75% of correct results for usual Popular music titles. For titles containing hardly extractable drum tracks, we obtain 40% of correct results, which is an acceptable performance. In any case, the worst results are due to different phenomena:

- 35%: the occurrences of the percussive sounds of the title are inherently not obvious.

This corresponds to titles that were classified as having hardly extractable drum tracks, for instance jazz extracts with subtle drum parts, or difficult snare drum games. Their extraction would require difficult audio signal preprocessing. However, for these titles, the extraction of all the percussive sounds is probably not the most efficient way to describe their global rhythm.

- 15%: the high-pitched percussive sound found is the voice of the singer.

This confusion appears exclusively between snare-drum-like percussions and women voices. Indeed, the averaging of percussive sounds removes their noisy part, and the result is close to female vocals, that often have a low noise ratio. This could probably be solved by considering another discriminative feature, such as the duration of the sound, which is often longer for sung syllables.

- 10%: the high-pitched percussive sound is confused with the low-pitched one.

This problem sometimes appears when the 2 percussive sounds are hit at the same time, but the priority given to bass-drum-like sounds (see 4.2) should avoid confusions. These confusions are probably due to the use of only 1 discriminative parameter, the zero-crossing rate, which is sufficient for very different-sounding percussive sounds. But another one is probably required, for example when the pitches of the 2 sounds are too close, or for the extraction of more than 2 different percussive sounds.

- 10%: the high-pitched percussive sound is not found because of its specificity.

This problem often appears when the snare-drum-like percussion is a clap sound for example. It is due to the initial percussive signal model (see 3.1), which is probably too general to hook on specific sounds.

- 10%: a high level of noise.

This problem appears in recordings where drums are drown under other instruments, often loud and saturated, and the averaging does not remove enough noise to extract percussive sounds correctly. It would require complex signal processing that would not fit the simple approach developed here.

6. Conclusion

We have presented a new approach to automatically describe the rhythm of percussive Popular music titles, by extracting a drum track representing the occurrences of its 2 most relevant percussive sounds. Our method provides more than 75% of correct extractions for the majority of Popular music titles. The resulting information contains 2 kinds of data: the extracted percussive sounds that represent a percussive audio signature of the title, and the percussive structure of the drum tracks, that represents a rhythmic signature of the title. These two information are useful for musical queries and are currently being used to design rhythmic similarities measures, themselves integrated in a content-based music browser.

7. REFERENCES

[1] Klapuri A, “Qualitative and quantitative aspects in the design of periodicity estimation algorithms”, EUSIPCO Conference Proceedings, 2000.

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

[3] Cooper G.W., Meyer L.B., “The rhythmic structure of music”, University of Chicago Press, 1960.

[4] Gabrielsson A., ”Similarity ratings and dimension analyses of auditory rhythm patterns”, Parts I & II, Scandinavian Journal of Psychology 14, 1973.

[5] Desain P., Honing H., “Computational models of beat induction: The rule-based approach”, Journal of New Music Research, 28(1), 29-42, 1999.

[6] Goto M., Muroaka Y., “A real-time beat tracking system for audio signals”, International Computer Music Conference, 1995.

[7] Brown J., “Determination of the meter of musical scores by autocorrelation”, JASA 94(4), 1993.

[8] Schloss A., “On the automatic transcription of percussive music – From acoustic signals to high-level analysis”, CCRMA internal report, Stanford University, 1985.

[9] Duxbury C., Davies M., Sandler M., “Separation Of Transient Information In Musical Audio Using Multiresolution Analysis Techniques”, Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFX01), Limerick (Irl), December 2001.

[10] Gouyon F., Delerue O., Pachet F., “On the use of zero-crossing rate for an application of classification of percussive sounds”, Third Digital Audio Effect Conference, Verona (Italy), 2000.

Audio examples of the extracted drum tracks can be heard at: http://www.csl.sony.fr/~aymeric/dt