EXTRACTION OF DRUM PATTERNS AND THEIR DESCRIPTION WITHIN THE MPEG-7 HIGH-LEVEL-FRAMEWORK

Matthias Gruhne ghe@idmt.fraunhofer.de
Christian Uhle uhle@idmt.fraunhofer.de
Christian Dittmar dmr@idmt.fraunhofer.de
Markus Cremer cre@idmt.fraunhofer.de

Fraunhofer IDMT
Langewiesener Str. 22
98693 Ilmenau, Germany

ABSTRACT

A number of metadata standards have been published in recent years due to the increasing availability of multimedia content and the resulting issue of sorting and retrieving this content. One of the most recent efforts for a well defined metadata description is the ISO/IEC MPEG-7 standard, which takes a very broad approach towards the definition of metadata. Herein, not merely hand annotated textual information can be transported and stored, but also more signal specific data that can in most cases be automatically retrieved from the multimedia content itself.

In this publication an algorithm for the automated transcription of rhythmic (percussive) accompaniment in modern day popular music is described. However, the emphasis here is not a precise transcription, but on capturing the “rhythmic gist” of the piece of music in order to allow a more abstract comparison of musical pieces by their dominant rhythmic patterns. A small-scale evaluation of the algorithm is presented along with an example representation of the thus gained semantically meaningful metadata using description methods currently discussed within MPEG-7.

1. INTRODUCTION

Stimulated by the ever-growing availability of musical material to the user via new media and content distribution methods an increasing need to automatically categorize audio data has emerged.

Descriptive information about audio data which is delivered together with the actual content represents one way to facilitate this search immensely. The aims of so-called metadata (”data about data”) are to e.g. detect the genre of a song, specify music similarity, perform a segmentation on a song, or simply recognize a song by scanning a database for similar metadata.

There have been a number of publications describing an approach to achieve these aims using features that belong to a lower semantic hierarchy order (“Low-Level-Tools”) [1]. These features are extracted directly from the signal itself in a computationally efficient manner, but carry little meaning for the human listener. The usage of high level semantic information relates to the human perception of music.

The rhythmic elements of music, determined by the drum and percussion instruments, play an important role especially in contemporary popular music. Therefore, the performance of advanced music retrieval applications will benefit from using mechanisms that allow the search for rhythmic styles or particular rhythmic features.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

© 2004 Universitat Pompeu Fabra.

1.1. The Metadata Standard MPEG-7

One example of a number of upcoming standards for the specification of metadata for audiovisual data is the MPEG-7 standard, which was finalized in late 2001.

The first version of MPEG-7 Audio (ISO/IEC 15938-4) does not, however, cover high level features in a significant way. Therefore the standardization committee agreed to extend this part of the standard. The work of contributing high level tools is currently being assembled in MPEG-7 Audio Amendment 2 (ISO/IEC 15938-4 AMD2). One of its features is *RhythmicPatternDS*. The internal structure of its representation depends on the underlying rhythmic structure of the considered pattern. The main advantage consists in the fact that for every pattern the most compact representation can be provided, resulting in an efficient comparison of the patterns and minimal memory needed for storage.

The system presented in this paper has been designed to extract an MPEG-7 Audio AMD2 conformant *RhythmicPatternDS* out of a musical audio signal. The system's complexity is low enough to allow real time operation on today's personal computers.

2. SYSTEM OVERVIEW

The system presented in this paper consists of three different parts. At the first processing stage occurrences of un-pitched percussive instruments are detected and classified. Based on the resulting drum transcription actual drum patterns are extracted. Finally, an MPEG-7 Audio conformant XML description is created.

2.1. Transcription of Percussive Instruments

This section gives an overview on the method for detection and classification of un-pitched percussive instruments, described in detail in [2]. The detection and classification of percussive events is carried out using a spectrogram-representation $\mathbf{X}$ of the audio signal. Differentiation and halfway-rectification of $\mathbf{X}$ yield a non-negative difference spectrogram $\hat{\mathbf{X}}$, from which the times of occurrence $\mathbf{t}$ and the spectral slices $\hat{\mathbf{X}}_t$ related to percussive events are deduced. Principal Component Analysis (PCA) is applied to $\hat{\mathbf{X}}_t$ according to (1).

$$\tilde{\mathbf{X}} = \hat{\mathbf{X}}_t \cdot \mathbf{W} \qquad (1)$$

Thereby, transformation matrix $\mathbf{W}$ reduces the number of slices to d components unifying decorrelation and variance normalization. The principal components $\tilde{\mathbf{X}}$ are subjected to Non-Negative Independent Component Analysis (NNICA) [3], which attempts to find un-mixing matrix $\mathbf{A}$ by optimizing a cost function describing the non-negativity of the components.

The spectral characteristics of un-pitched percussive instruments, especially the invariance of spectra of different notes compared to pitched instruments allows a separation of $\tilde{\mathbf{X}}$ using un-mixing matrix $\mathbf{A}$ into spectral profiles $\mathbf{F}$ according to equation (2).

$$\mathbf{F} = \mathbf{A} \cdot \tilde{\mathbf{X}} \qquad (2)$$

The spectral profiles can be used to extract the spectrogram's amplitude basis, from here forward referred to as amplitude envelopes E according to (3).

$$\mathbf{E} = \mathbf{F} \cdot \mathbf{X} \qquad (3)$$

This procedure is closely related to the principle of Prior Subspace Analysis (PSA) [4], modified to estimate the spectral profiles from the analyzed audio signal itself. In further contrast to the original procedure introduced in [4] no further ICA-computation is carried out on the amplitude envelopes.

The extracted components are classified using a set of spectral-based and time-based features.

The classification shall provide two sources of information. Firstly, components should be excluded from the rest of the process which are clearly harmonically sustained. Secondly, the remaining dissonant percussive components should be assigned to pre-defined instrument classes.

A suitable measure for the distinction of the amplitude envelopes is represented by the percussiveness, which is introduced in [5]. A slightly modified version is employed to distinguish the amplitude envelopes related to percussive instruments from the ones related to sustained sounds.

A spectral-based measure is constituted by introducing the spectral dissonance, earlier described in [5], [6]. A slightly modified version is employed to distinguish the spectra of harmonic sustained sounds from dissonant ones related to percussive sounds.

The assignment of spectral profiles to a priori trained classes of percussive instruments is provided by a k-nearest neighbour classifier with spectral profiles of single instruments from a training database. To verify the classification in cases of low reliability or several occurrences of the same instruments, additional features describing the shape of the spectral profile, e.g. centroid, spread, and skewness, are extracted. Other features are the center frequencies of the most prominent local partials, their intensity, spread, and skewness.

Drum-like onsets are detected in the amplitude envelopes using conventional peak picking methods. The intensity of the onset candidate is estimated from the magnitude of the envelope signal. Onsets with intensities exceeding a predetermined dynamic threshold are accepted. This procedure reduces crosstalk influences of harmonic sustained instruments as well as concurrent percussive instruments.

2.2. Extraction of Drum Pattern

The extraction of recurring pattern (drum pattern) from a list of automatically detected events (see section 2.1.) is illustrated in Figure 1.

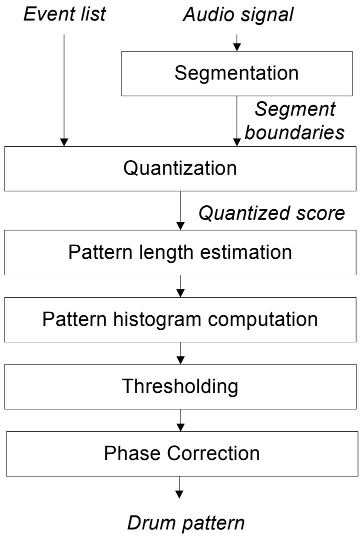

Figure 1: Block diagram of the second stage of drum pattern extraction

At first, the audio signal is segmented into similar and characteristic regions using a self-similarity method initially proposed by Foote [7]. The segmentation is motivated by the assumption, that within each region not more than one representative drum pattern occurs, and that the rhythmic features are nearly invariant.

Subsequently, the temporal positions of the events are quantized on a tatum grid. The term tatum grid refers to the pulse series on the lowest metric level [8]. Tatum period and phase is computed by means of a two-way mismatch error procedure, originally proposed for the estimation of the fundamental frequency in [9] and applied to tatum estimation before in [10]. An additional note onset detection process, finding note onsets in the audio signal, complements the list of note onsets from the percussive un-pitched instruments.

The pattern length is estimated by searching for the prominent periodicity in the quantized score with periods equaling an integer multiple of the bar length. The periodicity function is obtained by calculating a similarity measure between the signal and its time shifted version. The similarity between two score representations is calculated as weighted sum of the number of simultaneously occurring notes and rests in the score.

An estimate of the bar length is obtained by comparing the derived periodicity function to a number of so-called metric models, each of them corresponding to a bar length. A metric model is defined here as a vector describing the degree of periodicity per integer multiple of the tatum period, and is illustrated as a number of pulses, where the height of the pulse corresponds to the degree of periodicity. The best match between the periodicity function derived from the input data and pre-defined metric models is computed by means of their correlation coefficient. A periodicity function and two exemplary metric models are illustrated in Figure 2.

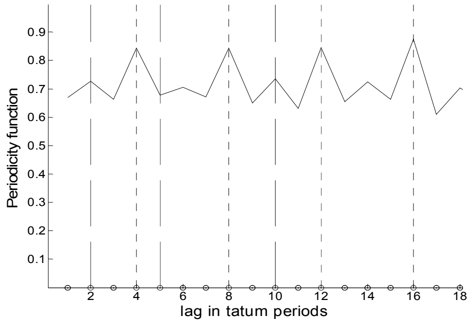

Figure 2: Periodicity function and two examples of metric models corresponding to four-four time (dotted line) and five-four time (dashed line)

A histogram-like representation $\mathbf{H}_{i,j}$ of the score $\mathbf{T}_{i,l}$ is obtained by measuring the frequency of occurrence of events per instrument and metric position according (4).

$$\mathbf{H}_{i,j} = \sum_{k=1}^{r} \mathbf{T}_{i,j+(k-1)b} \tag{4}$$

where $i=1...n$ and $j=1...b$, n represents the number of instruments, b is the bar length and r equals the number of bars.

The drum patterns are extracted by choosing the positions whose occurrence exceeds a threshold q_i (5).

$$\mathbf{D}_{i,j} = \begin{cases} \mathbf{H}_{i,j} & (\mathbf{H}_{i,j} > q_i), \\ 0 & (otherwise) \end{cases} \tag{5}$$

The final processing step estimates the start position of the pattern. It is assumed that the start of the pattern corresponds to the position featuring the strongest occurrence of kick drum notes. A further strategy is to identify common playing styles and to compare the extracted pattern to various exemplary patterns.

2.3. MPEG-7 AudioRhythmicPattern

The *AudioRhythmicPattern* descriptor uses a non-linear indexing of the velocity values with help of a so-called *PrimeIndex*, derived from prime factorization of the grid indices. The *PrimeIndex* indicates the rhythmic significance (rhythmic level) within the pattern. In general, velocity values that occur on a beat will be indicated by a *PrimeIndex* with a lower integer value than velocity values occurring between two beats (offbeat). Depending on meter and micro time different levels of rhythmic hierarchy will result.

Part of the bar	**1**	1+	**2**	2+	**3**	3+	**4**	4+
Rhythmic level	***	*	**	*	***	*	**	*
Grid position	1	2	3	4	5	6	7	8
Prime index	1	5	3	6	2	7	4	8
Velocity	100	0	112	0	150	68	120	0

Table 1: Example of a Rhythmic Pattern

The term micro time defines as the (close to) integer ratio between the beat period and the tatum period.

An example for a rhythmic pattern is given at Table 1. The meter is 4/4 and the micro time equals 2. This results in a total pattern size of 8.

Rhythmic level	***	***	**	**	*
Prime index	1	2	3	4	7
Velocity	100	150	112	120	68

Table 2: Rhythmic Pattern after deleting zeros and re-ordering

All velocity values equal to zero and their corresponding prime indices (elements) are deleted. According to the ascending order of the prime indices the elements will be rearranged, resulting in the final representation (see Table 2).

```
…
    <Audio xsi:type="AudioSegmentType">
     <AudioDescriptionScheme
      xsi:type="AudioPatternType">
      <Meter>
        <Numerator>4</Numerator>
        <Denominator>4</Denominator>
      </Meter>
      <TimePoint>PT00N1000F</TimePoint>
      <Pattern>
        <BarNum>1</BarNum>
        <InstrumentID>36</InstrumentID>
        <Microtime>2</Microtime>
        <PrimeIndex>1 2 3 4 7 </PrimeIndex>
        <Velocity>110 150 112 120 68 </Velocity>
      </Pattern>
     </AudioDescriptionScheme>
    </Audio>
 …
```

Figure 3: section of an example rhythmic pattern XML description

3. TEST RESULTS

An informal listening test has been conducted in order to quantify the abilities of the presented system. Nine human listeners (a mixture of lab members and students with varying degree of musical training ranging from non-musicians to skilled performers) were confronted with 92 excerpts of 40 test songs. Each excerpt features a minimum duration of six seconds. The songs include a wide range of musical genres where the appearance of drum patterns is common, e.g. Rock, Pop, Latin, Soul and House. The human listeners were instructed to compare the original excerpts to the synthetic rendition of the extracted patterns. The rating ranges from five (for a perfect extracted pattern) to one (for an unrecognizable pattern). Details on the results of the listening test are displayed in Figure 4. The solid line shows the mean score value per test item in descending order. The dashed line shows the corresponding standard deviation arranged as a tolerance interval around the mean value. It can be seen that almost 70 percent of the test items have been assigned a score equal or greater three. Another interesting observation is the fact, that the standard deviation does not diverge strongly amongst the test subjects. The presumption that a rating of pattern quality could be a very subjective task is invalidated by a small average standard deviation of 0.74 score points.

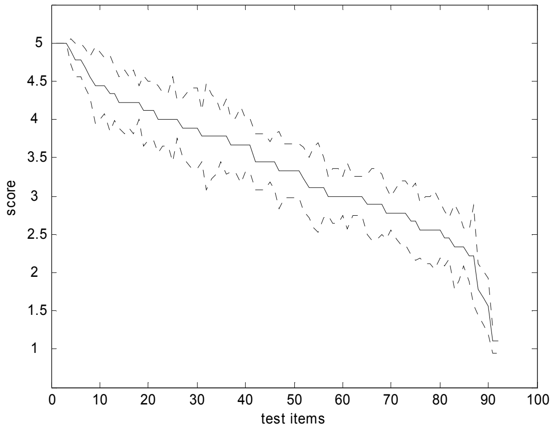

Figure 4: Results of Listening Tests

4. CONCLUSIONS

In this publication an algorithm for the automated extraction of rhythmic patterns from popular music items has been introduced and evaluated. The presented test results verify that the algorithm produces viable results, and can indeed be deployed for the comparison of rhythmic aspects of different music items. The extracted rhythmic patterns are represented using the latest description methods under preparation within the international standard ISO/IEC MPEG-7.

Thus, with the tools and methods presented above the realization of a complete, fully automated system for efficient rhythmic characterization and comparison of contemporary popular music tunes is possible.

5. REFERENCES

[1] Allamanche, E. Herre, J. Hellmuth, O. "Content-based Identification of Audio Material Using MPEG-7 Low Level Description", *Proceedings of the 2nd Annual Symposium on Music Information Retrieval,* Bloomington, USA, 2001.

[2] Dittmar, C. Uhle, C. "Further Steps towards Drum Transcription of Polyphonic Music", *Proceedings of the AES 116th Convention*, Berlin, Germany, 2004.

[3] Plumbley, M. "Algorithms for Non-Negative Independent Component Analysis", *Proceedings of the IEEE Transactions on Neural Networks*, 14 (3), pp 534- 543, 2003.

[4] Fitzgerald, D. Lawlor, B. and Coyle, E. "Prior Subspace Analysis for Drum Transcription", *Prcoeedings of the 114th AES Convention*, Amsterdam, Netherlands, 2003.

[5] Uhle, C. Dittmar, C. and Sporer, T. "Extraction of Drum Tracks from polyphonic Music using Independent Subspace Analysis", *Proceedings of the Fourth International Symposium on Independent Component Analysis,* Nara, Japan, 2003.

[6] Sethares, W. "Local Consonance and the Relationship between Timbre and Scale", *Journal of the Acoustical Society of America*, 94 (3), pt. 1, 1993.

[7] Foote, J. "Automatic Audio Segmentation Using a Measure of Audio Novelty", *Proceedings of the IEEE Int. Conf. on Multimedia and Expo*, vol. 1, pp. 452-455, 2000.

[8] Bilmes, J.A. "Timing is of Essence", *MSc Thesis*, Massachusetts Institute of Technology, 1993.

[9] Maher, R. Beauchamp, J. "Fundamental Frequency Estimation of Musical Signals Using a Two-Way Mismatch Procedure", *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254-2263, 1994.

[10] Gouyon, F. Herrera, P. Cano, P. "Pulse-Dependent Analysis of Percussive Music", *Proceedings of the AES 22nd Int. Conference on Virtual, Synthetic and Entertainment Audio,* 2002.