MUSIC RETRIEVAL BY RHYTHMIC SIMILARITY APPLIED ON GREEK AND AFRICAN TRADITIONAL MUSIC

Iasonas Antonopoulos, Aggelos Pikrakis and Sergios Theodoridis
Dept. of Informatics & Telecommunications
University Of Athens, Greece

Olmo Cornelis, Dirk Moelants and Marc Leman
IPEM - Dept. of Musicology
Ghent University, Belgium

ABSTRACT

This paper presents a method for retrieving music recordings by means of rhythmic similarity in the context of traditional Greek and African music. To this end, Self Similarity Analysis is applied either on the whole recording or on instances of a music thumbnail that can be extracted from the recording with an optional thumbnailing scheme. This type of analysis permits the extraction of a *rhythmic signature* per music recording. Similarity between signatures is measured with a standard Dynamic Time Warping technique. The proposed method was evaluated on corpora of Greek and African traditional music where human improvisation plays a key role and music recordings exhibit a variety of music meters, tempi and instrumentation.

1 INTRODUCTION

In the context of Music Information Retrieval (*MIR*), finding music recordings with similar rhythmic characteristics is a highly desired task both for the untrained listener and the musicologist.

Over the years, several methods have been proposed in the context of Western music for retrieving music with similar rhythmic characteristics, e.g. tempo, meter and rhythmic patterns. Here a short overview of relevant papers is given: The work in [1] measures the similarity between rhythmic patterns extracted from music recordings and artificially generated percussive sounds. The approach in [2] extracts temporal patterns from the energy envelop of the signal in an attempt to classify music recordings to predefined classes. In [3] a set of classification schemes are proposed that are based on extracting rhythmic patterns from the signal's spectrum. The method proposed in [4] focuses on ballroom dances and is based on features stemming from the histogram of Inter-Onset Intervals. Finally, the work in [5] evolves around self similarity analysis of the music recording. In some of the above methods, the term "rhythmic signature" is used to as a means to encode fundamental rhythmic characteristics of the music recordings.

This paper focuses on rhythmic similarity in non Western music, i.e., Traditional Greek and African music, which have so far received little attention in the field of MIR. Such traditions impose a number of research challenges, mainly due to the complexity of the music meters, the system of music intervals and the highly improvisational attitude of the music performers. The latter gives an additional research challenge as serves the preservation of cultural heritage and also highlights the importance of MIR systems to apply to corpora that fall outside the traditional Western schemes. In an attempt to measure rhythmic similarity in such music corpora, this paper exploits the repetitive nature of the music recordings by means of Self Similarity Analysis. This type of analysis reveals periodicities that are inherent in the music signal. Such periodicities are expressed as a sequence of values to which we also refer by the term *rhythmic signature*. To this end, we investigate the possibility of applying an optional thumbnailing scheme as a preprocessing step to extracting rhythmic signatures. Similarity measurement between signatures is performed by means of a standard *Dynamic Time Warping* technique.

Section 2 presents the proposed audio thumbnailing scheme and Section 3 describes how *rhythmic signatures* are extracted from the music signal. The proposed similarity measure is presented in Section 4. Results and implementation details are given in Section 5 and conclusions are drawn in Section 6.

2 THUMBNAILING SCHEME

The proposed audio thumbnailing scheme is optional and is considered to be a variation of the method proposed in [6], in the sense that a different feature extraction scheme is used in this paper.

2.1 Feature extraction

At a first step, the music recording is short-term processed by means of a moving window technique. The short-term frames are chosen to be $\simeq$ 186 *msecs* long, non-overlapping and are multiplied by a *Hamming* window. Each frame is given as input to a mel-scale filter bank [8] that consists of overlapping triangular filters. The center frequencies of the filters coincide with the frequencies of whole tones on a chromatic scale, starting from $F_0 = 110$ Hz and moving up to $\simeq 6.3$ KHz, resulting into 36 filters

which cover approximately six octaves. In the sequel we will refer to this type of MFCCs as *chroma-based MFCCs* due to the similarities it bears with the "chroma vector" [6]. Further details are given in [7].

To proceed, let $\underline{c}(n)$ be the 36×1 vector of *chroma-based* MFCCs from the n-th frame. The sequence of vectors can be written in matrix notation as

$$\mathbf{C}_{36\times N} = [\underline{c}(1)\ \underline{c}(2)\ \ldots\ \underline{c}(N)],$$

where N is the number of short-term frames.

At a next step, *Singular Value Decomposition* (*SVD*) is applied on the transpose, $\mathbf{C}^T$, of $\mathbf{C}$, i.e., $\mathbf{C}^T = \mathbf{U\Sigma V}$, where $\mathbf{U}_{N\times 36}$ and $V_{36\times 36}$ are the projection matrices and $\mathbf{\Sigma}_{36\times 36}$ is the matrix of singular values. The first six rows of the transpose, $\mathbf{U}^T$, of $\mathbf{U}$, are finally selected as the feature sequence.

2.2 Thumbnail selection

The *SSM* is generated from the first six rows of $\mathbf{U}^T$ using the Euclidean Distance function as metric [5]. By its definition the *SSM* is symmetric around the main diagonal and it therefore suffices to focus on its lower triangle. At a first step, the *SSM* is correlated with a rectangular window, w (size $D \times D$). The window has 1's on the main diagonal and zeros elsewhere. If (i, j) are the position indices of an element of *SSM*, the upper left corner of w is chosen to coincide with (i, j). The correlation result, $S(i, j)$, for $SSM(i, j)$ is therefore computed as follows:

$$\begin{aligned} S(i,j) &= \sum_{d_1=0}^{D-1}\sum_{d_2=0}^{D-1} SSM(i+d_1, j+d_2)w(d_1,d_2) \\ &= \sum_{d=0}^{D-1} SSM(i+d, j+d) \end{aligned} \quad (1)$$

At a second step, let $S(k, m)$ be the lowest value of S. $S(k, m)$ resides on the diagonal with index $k - m$ and elements $\{S(k, m), S(k+1, m+1), \ldots, S(k+D-1, m+D-1)\}$ form a segment on the diagonal that defines the desired thumbnail. The two corresponding feature subsequences, i.e., two instances of the thumbnail, are

$$\{U_k^T, U_{k+1}^T, \ldots U_{k+D-1}^T\}$$

and

$$\{U_m^T, U_{m+1}^T, \ldots, U_{m+D-1}^T\},$$

respectively. Parameter D controls the size of the thumbnail and is user defined, depending on the corpus under study (see Section 5).

It has to be noted that the proposed thumbnailing scheme is optional and depends on the dataset under study. If it is skipped, the *r*hythmic signatures (see next section) will be extracted by taking into account the whole audio recording. This is desirable if one is unsure whether the two instances of the extracted thumbnail are indeed representative of the complete music recording. In the sequel, the term music signal will refer to either the two instances of the selected thumbnail or the complete music recording.

3 EXTRACTING RHYTHMIC SIGNATURES

3.1 Feature extraction

At a first step, the music signal (i.e., the two thumbnail instances or the complete recording) is short-term processed to extract a sequence of *chroma*-based MFCCs, as in Section 2.1. However, this time, shorter, overlapping windows are used (window length is $\simeq 93$ *msecs* and window step is 11.6 *msecs*). Following the notation that was introduced in Section 2.1, let $\mathbf{C} = [\underline{c}(1)\ \underline{c}(2)\ \ldots\ \underline{c}(N)]$, be the new sequence of MFCCs.

At a first step, $\mathbf{C}$ is long-term segmented with a moving long-term window (window length is 4 *secs* and step is 1 *sec*). To simplify notation, let $\mathbf{C_t} = [\underline{c_t}(1)\ \underline{c_t}(2)\ \ldots\ \underline{c_t}(M)]$, be the subsequence that corresponds to the t-th long-term window, where M is the window length measured in number of frames. The *SSM* is then calculated for each long-term window, using the Euclidean Distance metric. For the t-th long-term window, the mean value, $R_t(k)$, of each diagonal in the lower *S*SM triangle is computed, i.e., $R_t(k) = \frac{1}{M-k}\sum_{l=k}^{M}\left\|\underline{c_t}(l), \underline{c_t}(l-k)\right\|$, where k is the diagonal index and $\|.\|$ is the Euclidean distance function. Each R_t is treated as a signal. At a next step, the mean signal, R_μ, of all R_t's is computed, i.e.,

$$R_\mu(k) = \frac{1}{T}\sum_{t=1}^{T} R_t(k),$$

where T is the number of long-term windows. R_μ is then normalized to unity, i.e., $R_\mu(k) = \frac{R_\mu(k)}{max(R_\mu)}$.

As can be seen in Figure 1, R_μ exhibits a number of valleys (local minima). Each valley corresponds to a periodicity that is inherent in the music signal. Such periodicities are related to the rhythmic characteristics of the recording, e.g., music meter and tempo [7]. In what follows, we will refer to R_μ as the *rhythmic signature* of the music recording. The main idea behind this approach, is that, recordings with similar rhythmic characteristics are expected to yield "similar" signatures (as can be seen in the upper part of Figure 1). On the contrary, different rhythmic characteristics will result into "dissimilar" signatures (bottom part of Figure 1). Therefore, the next challenge is to devise a similarity measure for signatures.

4 SIMILARITY MEASURE FOR SIGNATURES

If L is the number of music recordings in a corpus, L rhythmic signatures are first extracted and stored as metadata. In order to measure similarity between signatures, a standard *Dynamic Time Warping* cost has been employed. As is the case with *DTW* techniques [8], a set of local path constraints needs to be first defined. In our study we experimented with two types of constraints, i.e., *Sakoe-Chiba* and *Itakura* and adopted the former.

If a rhythmic signature is drawn from the corpus, its matching cost against the remaining L-1 signatures is calculated using the adopted *D*TW technique. This procedure yields $L-1$ cost values which are sorted in ascending

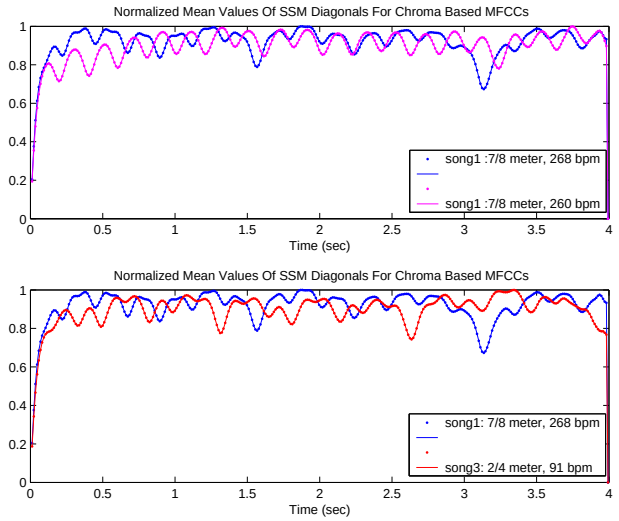


Figure 1. *Top*: *Signatures* from two recordings of music meter $\frac{7}{8}$. *Bottom*: *Signatures* from a recording of meter $\frac{7}{8}$ and a recording of meter $\frac{2}{4}$.

order, with the lowest values indicating highest similarity. The next section focuses on evaluating this matching scheme on two corpora of traditional Greek and African music.

5 EXPERIMENTS AND RESULTS

5.1 Corpus of Greek Traditional Dance music

The first corpus of our study consists of 220 tracks of Greek Traditional Dance Music, which are drawn from various Greek regions. The tracks were manually categorised into four genres as shown in Table 1. These genres exhibit certain variety in terms of instrumentation and rhythm. From the corpus description it can be noticed that the longest music meter duration (approximately 2 *secs*) appears in class 2, for tempo $\cong$ 90 *bpm* and music meter $\frac{3}{4}$. By using a thumbnail which is 10 *secs* long, the longest music meter is repeated up to 5 times in each long-term segment. Our study has revealed that, this expected number of repetitions is sufficient for the extraction of reliable *rhythmic* signatures and justifies our choice for the length of thumbnails. Similarly, the longest meter duration affects the length of the long-term window in Section 3.1. By the definition of self similarity analysis, a periodicity of k lags will manifest itself as a valley of R_μ, if the long-term window is at least $2k$ long. Therefore the length of the long-term window was chosen equal to $4secs$ to capture the periodicity of the longest music meter. Finally, the range of lags of R_μ on which *D*TW is employed starts with the lag corresponding to the fastest tempo value and reaches up to the lag that corresponds to the longest meter-tempo pair.

Table 2 presents the confusion matrix for the Greek corpus, where the leave-one-out method was applied on the complete corpus. Table 2 reveals that, when only the best result (lowest matching cost) was examined, limited confusion occured between the classes 3 and 4 and classes 1 and 2. Further experimentation revealed that, when the two lowest matching costs were taken into account, the confusion matrix remained the same within statistical confidence.

class id	# of songs	meter	tempo range (bpm)
1	53	2/4	91-95
2	63	3/4	93-105
3	62	7/8	250-280
4	42	2/4	150-180

Table 1. Description of Greek Traditional Dance corpus.

Precision %	Class 1	Class 2	Class 3	Class 4
Class 1	**94.3**	3.2	1.7	0
Class 2	3.8	**96.8**	0	0
Class 3	1.9	0	**96.6**	10.9
Class 4	0	0	1.7	**89.1**
Recall %	**Class 1**	**Class 2**	**Class 3**	**Class 4**
Class 1	**94.3**	3.8	1.9	0
Class 2	3.2	**96.8**	0	0
Class 3	1.6	0	**90.3**	8.1
Class 4	0	0	2.4	**97.6**

Table 2. Precision and recall for Greek Traditional corpus.

5.2 Corpus of African music

A collection of 103 pieces was selected from the music archives of the Royal Museum of Central-Africa in Tervuren (Brussels). This institute has one of the most important collections of African music in the world. [1] The current selection contains field recordings of Congo and Rwanda recorded during the second half of the 20^{th} century [9]. Similarly to the Greek music corpus, the rhythmic structure is highly repetitive and contains a wide range of rhythmic structures, including irregular meters that are seldom found in Western music [10].

class id	# of songs	meter
1	27	3/4
2	26	4/4
3	24	5/4
4	26	6/4

Table 3. Description of the 1^{st} set of the African corpus.

The corpus has been manually annotated with a (perceived) ground truth which has been used to evaluate the computer analysis. Two types of classification were made: one according to the meter, the other focusing on a selection of characteristic repetitive rhythmic patterns. The first classification, Table 3, uses four metric classes $\frac{3}{4}$, $\frac{4}{4}$, $\frac{5}{4}$ and

[1] `http://music.africamuseum.be`

Precision %				
Class id	1	2	3	4
1	**68.8**	4.3	20	0
2	12.5	**82.6**	12	3.7
3	15.6	13	**64**	3.7
4	3.1	0	4	**92.6**
Recall %				
Class id	1	2	3	4
1	**78.6**	3.6	17.9	0
2	14.8	**70.4**	11.1	3.7
3	20	12	**64**	4
4	3.7	0	3.7	**92.6**

Table 4. Precision and recall (1^{st} set of African corpus).

$\frac{6}{4}$. Table 5, is restricted to 44 pieces which can be classified as variants of 5 prototypical patterns: short-long-long (quintuple); short-long-long-short-long-long-long (sextuple); long-short-short-long (triple1); short-long (triple2); and short-short-long (duple).

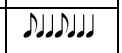
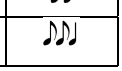
class id	# of songs	pattern
quintuple	10	
sextuple	14	
triple1	8	
triple2	5	
duple	7	

Table 5. Description of the 2^{nd} set of African corpus.

The thumbnail that was selected for each piece by the proposed method, often tended to contain parts of the song where the most percussive events occurred. Since this could lead to a dense regular structure that can easily be confused with patterns in which each beat is articulated the *rhythmic signatures*, R_μ's were extracted from whole audio recordings and the thumbnail scheme was skipped. Tables 4, 6 reveal that the results of both sets of African music are promising. When manually checking the problematic cases, mistakes can mostly be related to the occurrence of variants of the main pattern within one piece. The possible variations are mostly the addition of a percussive event where a rest used to be, or the opposite: omission of a percussive event. The meter and beat stay however the same.

6 CONCLUSIONS

This paper presented a music retrieval method based on rhythmic similarity measurement. The method yielded satisfactory results on coprora of traditional Greek and African music. In future work, more sophisticated *D*TW techniques will be used on larger corpora and the possibility to extract multiple rhythmic signatures per music recording will be investigated.

Precision %					
Class id	1	2	3	4	5
1	**70.80**	3.1	0	10	14.3
2	4.2	**84.4**	0	0	0
3	8.3	3.1	**86.7**	0	0
4	8.3	0	13.3	**60**	0
5	8.3	9.4	0	30	**85.7**
Recall %					
Class id	1	2	3	4	5
1	**85**	5	0	5	5
2	3.6	**96.4**	0	0	0
3	12.5	6.3	**81.3**	0	0
4	20	0	20	**60**	0
5	14.3	21.4	0	21.4	**42.9**

Table 6. Precision and recall (2^{nd} set of African corpus).

7 REFERENCES

[1] J. Paulus and A. Klapuri, "Measuring the similarity of rhythmic patterns.", *Proceedings of ISMIR*, Paris, France, September 2002.

[2] S. Dixon, F. Gouyon and G. Widmer, "Towards characterisation of music via rhythmic patterns.", *Proceedings of ISMIR*, Barcelona, Spain, 2004.

[3] G. Peeters, "Rhythm Classification Using Spectral Rhythm Patterns", *Proceedings of ISMIR 2005*, London, September, 2005.

[4] F. Gouyon and S. Dixon, "Dance music classification: a tempo-based approach.", *Proceedings of ISMIR*, Barcelona, Spain, 2004.

[5] J. Foote, M. Cooper and U. Nam, "Audio Retrieval by Rhythmic Similarity", *Proceedings of ISMIR*, Paris, France, September 2002.

[6] M. Bartsch and G.H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations", *IEEE Transactions on Multimedia*, 7(1), 96-104, 2005.

[7] A. Pikrakis, I. Antonopoulos and S. Theodoridis, "Music Meter and Tempo Tracking from raw polyphonic audio", *Proceedings of ISMIR*, Barcelona, Spain, 2004.

[8] S. Theodoridis and K. Koutroumbas, *Pattern recognition*, Academic Press, 3d Edition, 2006.

[9] O. Cornelis et al., "Digitisation of the ethnomusicological Sound Archive of the Royal Museum for Central Africa.", *IASA Journal*, pp 35-43, 2005.

[10] O. Cornelis et al., "Problems and Opportunities of Applying Data- & Audio-Mining Techniques to Ethnic Music", *Journal of Intangible Heritage*, (in press), 2007.