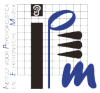
MUSIC RETRIEVAL BY RHYTHMIC SIMILARITY APPLIED ON GREEK AND AFRICAN TRADITIONAL MUSIC

Iasonas Antonopoulos, Aggelos Pikrakis and Sergios Theodoridis

Department Of Informatics & Telecommunications
University of Athens, Greece,
web: http://dsp.di.uoa.gr



Olmo Cornelis, Dirk Moelants and Marc Leman

Institute for Psychoacoustic & Electronic Music,
Department Of Musicology, University of Ghent, Belgium,
web: http://www.ipem.ugent.be

Problem Description

This work addresses the problem of retrieving music recordings by means of rhythmic similarity in the context of Greek and African music. Both corpora exhibit a repetitive nature which is exploited by means of Self Similarity Analysis so as to reveal inherent periodicities. These periodicities are encoded in a sequence of values referred to as "rhythmic signature". As a result, from each recording a single "rhythmic signature" is extracted. To this end, we have investigated the possibility for an optional thumbnailing scheme. *Dynamic Time Warping* has been used to measure similarity between "rhythmic signatures".

Optional Thumbnailing Scheme

- Short-term processing (~186msecs Hamming window, non overlapping): 36 chroma-based MFCC's (with mel Filter Bank Centers coinciding with the chromatic tones) are extracted starting from 110 Hz and moving up to ~6.3 kHz.
- Let $C_{36\times N} = [\underline{c}(1), \underline{c}(2), \ldots, \underline{c}(N)]$, where N is the number of short-term frames.
- Singular value Decomposition (SVD) is applied on the transpose C^T, of C, i.e., $C^T = U \Sigma V$, where $U_{N\times 36}$ and $V_{36\times 36}$ are the projection matrices and $\Sigma_{36\times 36}$ is the matrix of singular values. The first six rows of the transpose U^T of U are selected as the feature sequence.
- The Self Similarity Matrix (SSM) is generated from the six rows of U^T using the Euclidean Distance metric.
- The SSM is correlated with a rectangular window, *w* (size *DxD*). The window has 1's on the main diagonal and 0's elsewhere and *w* is slided toward the main diagonal with unity (lag) step and the sum of the correlation is calculated. *D* defines the thumbnail size.
- Matrix S stores the correlation result

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

- The lowest value in S, i.e., S(I,j) is selected and the position of the thumbnail is indicated by *i*,j indices (measured in lags). The audio is now represented by the selected thumbnail.

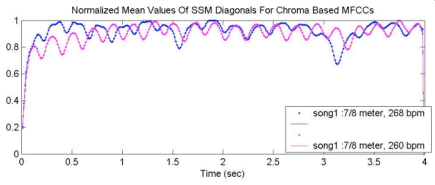
** The proposed thumbnail scheme is optional and D may depend upon the corpus under study. If skipped the whole audio participates in the following analysis.*

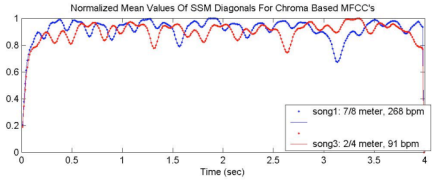
Extracting Rhythmic Signatures

- Short-term processing like above (~93.6msecs Hamming window, ~81.3 msecs overlap)
- Let $C_{36\times N} = [\underline{c}(1), \underline{c}(2), \ldots, \underline{c}(N)]$, where N is the number of short-term frames.
- *C* is long-term segmented with a long-term moving window (window length is 4 secs and step is 1 sec). Let : $C = [c_t(1), c_t(1), \ldots, c_t(M)]$ be the subsequence that corresponds to the *t*-th long-term window, where *M* is the window length measured in frames.
- The *SSM* is calculated for each long-term window using the Euclidean Distance metric and the mean value, $R_t(k)$ of each diagonal in the lower *SSM* triangle is calculated, 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 the diagonal index and $\|\cdot\|$ the Euclidean Distance.

- R_t is treated as a signal and the mean signal R_μ of all signals is computed as: $R_\mu(k) = \frac{1}{T}\sum_{t=1}^{T} R_t(k)$
- R_μ is normalized to unity.
- When plotted against *k*, R_μ exhibits a number of valleys (local minima), corresponding to periodicities inherent in the music signal. We will refer to R_μ as *rhythmic signature*.





Similarity Measure for Signatures

- For the similarity measurement between extracted *rhythmic signatures* standard *Dynamic Time Warping* (DTW) is employed *(Sakoe-Chiba local constraints)*.
- If *L* is the number of music recordings in a corpus, *L rhythmic signatures* are extracted and stored as metadata.
- For a *rhythmic signature* drawn from the corpus, *L*-1 cost values (one against all) are calculated using the adopted DTW technique.
- These cost values are sorted in ascending order with lowest values indicating highest similarity.

Evaluation

Greek Corpus

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

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**	3.2	1.7	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**

- In the evaluation of the Greek Corpus the thumbnail stage was used.
- Only the lowest value was returned.
- Small confusion exists between classes of Greek corpus.

African Corpus

set 1

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

Precision %	Class 1	Class 2	Class 3	Class 4
Class 1	**68.6**	4.3	20	0
Class 2	12.5	**82.6**	12	3.7
Class 3	15.6	13	**64**	3.7
Class 4	3.1	0	4	**92.6**

Recall %	Class 1	Class 2	Class 3	Class 4
Class 1	**78.6**	3.6	17.9	0
Class 2	14.8	**70.4**	11.1	3.7
Class 3	20	12	**64**	4
Class 4	3.7	0	3.7	**92.6**

set 2

class id	# of songs	pattern
1 (quintuple)	10	♪♩♩
2 (sextuple)	14	♪♪♩♪♩♩
3 (triple1)	8	♩♪♪♪
4 (triple2)	5	♪♪♩
5 (duple)	7	♪♪♪

Recall %	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	**85**	5	0	5	5
Class 2	3.6	**96.4**	0	0	0
Class 3	12.5	6.3	**81.3**	0	0
Class 4	20	0	20	**60**	0
Class 5	14.3	21.4	0	21.4	**42.9**

Precision %	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	**70.80**	3.1	0	10	14.3
Class 2	4.2	**84.4**	0	0	0
Class 3	8.3	3.1	**86.7**	0	0
Class 4	8.3	0	13.3	**60**	0
Class 5	8.3	9.4	0	30	**85.7**

- The thumbnailing scheme was not employed since it often tended to contain parts of the songs which were not representative of the recording.
- Most of the confusion between African classes is related to the occurrence of variants of the main pattern within one piece.
- Possible variations are mostly the addition of a percussive event where a rest used to be, or the opposite: omission of a percussive event.

References:

[1] A. Pikrakis, I. Antonopoulos and S. Theodoridis, "Music Meter and Tempo Tracking from raw polyphonic audio", *Proceedings of ISMIR*, Barcelona, Spain,2004.
[2] O. Cornelis et al., "Digitisation of the ethnomusicological Sound Archive of the Royal Museum for Central Africa.", IASA Journal, pp 35-43, 2005.
[3] O. Cornelis et al. "Problems and opportunities of Applying Data & Audio Mining Techniques to Ethnic Music", Journal Of Intangible Heritage, 2007 (in press).