QUERY-BY-TAPPING

Pierre Hanna, Julien Allali, Pascal Ferraro and Matthias Robine
LaBRI - University of Bordeaux
351, cours de la Libération
33405 TALENCE Cedex - FRANCE
firstname.name@labri.fr

ABSTRACT

We propose a new retrieval system based on musical structure using symbolic structural queries. The aim is to compare musical form in audio files without extracting explicitly the underlying audio structure. From a given or arbitrary segmentation, an audio file is segmented. Irrespective of the audio feature choice, we then compute a self-similarity matrix whose coefficients correspond to the estimation of the similarity between entire parts, obtained by local alignment. Finally, we compute a binary matrix from the symbolic structural query and compare it to the audio segmented matrix, which provides a structural similarity score. We perform experiments using large databases of audio files, and prove robustness to possible imprecisions in the structural query.

1. INTRODUCTION

The number of digital audio documents available online is increasing. New interfaces for browsing have to be proposed to users in order to allow retrieval based on musical properties instead of only text. Content-based music retrieval research area investigates the development of retrieval tools that permit users to sing or whistle an excerpt of the musical piece searched [1]. Even if these query-by-humming/singing/whistling systems become more and more efficient and precise, the automatic comparison between two musical pieces is still an open problem, especially in the case of large database, even monophonic [2,3].

Experiments with existing systems for automatically estimating melodic similarity between musical pieces clearly show that adding the information about note duration improves their accuracy [4]. Both information about melody and rhythm are thus combined. But approaches focusing exclusively on rhythmic properties have already been proposed [5], leading to Query-By-Tapping (QBT) systems [6].

QBT systems only consider the rhythm of the song's melody. No pitch information is taken into account. Users enter a sequence of taps or claps and the system has to retrieve the corresponding melody in a given database. MIDI interfaces such as e-Drum or MIDI keyboards have been experimented [6], but one can also think about PC or mobile phone keyboards. A few QBT systems considering audio signals have been recently presented. They record with a microphone (for example with a mobile phone) the user tapping or clapping the rhythm of the melody requested [7].

QBT systems rely on the automatic estimation of the similarity between two rhythmic patterns. By considering these patterns as strings, adaptations of string matching techniques, such as N-grams, have been proposed in [8]. Other systems compute a similarity measure based on dynamic programming [7]. More efficient algorithms for comparison have been proposed in [6]. More recently, algorithms dedicated to the geometric representation of music have been developed and experimented [9]. The main difficulty for these systems is to be able to retrieve the pieces that are similar, even if not necessarily identical.

In this paper, we propose and experiment a new QBT system, based on a new algorithm for rhythmic event detection and on an adaptation of alignment algorithm, successfully applied for the estimation of the melodic similarity [4]. In Section 2, we discuss the existing onset or transient detection method and propose a new one, based on the fourth statistical moment. Then, in Section 3, we detail the representation chosen for the analyzed rhythmic patterns, and the alignment algorithm proposed. Results of the MIREX are proposed in Section 4.

2. ANALYSIS OF RHYTHMIC EVENTS

The first step of a QBT system is the analysis of the input audio query in order to extract the rhythmic information necessary for the comparison with the musical pieces of the database. This analysis is one of the most important part of the global system, since errors in the analysis will result in errors in all the retrieval system. However, it is important to note that spurious errors are generally unavoidable.

Concerning a QBT system, a complete transcription of the audio query is not required. Only the rhythmic information has to be extracted. Information about pitch has no interest. The algorithms considered are based on techniques for note onset detection. Literature about such methods is profuse [10, 11], and several accurate methods have been proposed. However, the audio signal analyzed by QBT systems are generally not musical, in the sense that no pitch can be heard. Therefore some of these methods cannot be as accurate with such audio signal than with purely musical signals.

Many methods are based on the analysis of the variations of the energy of the signals. Improvements by weighting the energy variations in the high frequencies allow the detection of percussive onsets [12]. In [11] are described the methods relying on phase vocoder, that is on the Short Time Fourier Transform (STFT) of the signal analyzed. The spectral difference method calculates the difference between the spectral magnitudes of two successive frames. Variations on the phase can also indicate the presence of onsets. Considering differences in the complex domain permits to take into account both magnitude or phase variations. Applying Kullback Leibler distance helps to highlight high variations and to ignore small ones.

All these approaches lead to the definition of onset detection function. Problems come with the selection of the correct onsets. A peak-picking algorithm generally consider local maxima which are higher than a arbitrarily chosen threshold [10]. The choice of this threshold has a large impact on the results.

One important point to take into account is that queries for QBT systems are generally noisy. Moreover, no note is played. Techniques have to detect transients instead of note onsets. Existing methods may thus be limited, since they have generally been developed for music transcription. Therefore, we propose to experiment a new method dedicated to the noisy context and to the percussive characteristics of the musical events considered. This new approach relies on the analysis of the probability density function (PDF) of the audio signal samples.

2.1 Transient Detection Based on Kurtosis Variation

The transient detection method we proposed is based on the fourth statistical moment, assuming the analyzed signal x as a random signal X. Kurtosis, denoted K, is defined from the fourth moment:

$$K = \frac{\mathcal{E}(X^4)}{\sigma_X^4} \quad (1)$$

where $\mathcal{E}(X)$ is the expected value of X and σ_X represents its standard deviation.

Kurtosis characterizes the general shape of a PDF and more particularly its flatness. The higher the kurtosis, the sharper peak the PDF has. At the opposite, a low kurtosis is related to a PDF with a more rounded peak. Since a natural random signal is generally assumed as Gaussian, the kurtosis value associated is generally expected to be 3. Higher kurtosis value indicates a higher probability (than a normally distributed variable) of values near the mean. It may characterize the presence of transients in noisy sounds.

Therefore, the transient detection function proposed relies on the assumption that the presence of a transient may significantly increase the value of the kurtosis. In order to detect variations in kurtosis, the signal is analyzed during successive overlapping short frames ($N = 512$ samples, sampling rate 8000Hz, overlap rate 0.875%). It is important to note that a very low-level noise may be added to the analyzed signal to avoid null samples which may false this method. One kurtosis value K_r is associated to each frame r following the equation:

$$K_r = \frac{\sum_{n=n_r}^{n_r+N} x[n]^4}{\frac{1}{N}(\sum_{n=n_r}^{n_r+N} x[n]^2)^2} \quad (2)$$

where r represents the index of the current frame, and by assuming the mean of the signal samples to be null.

Since we focus on the variations of kurtosis, we propose to compute the derivative dK_r of the kurtosis, and a half-wave rectifier is applied to this derivative to calculate the transient detection function (DF):

$$\mathrm{dK}_r = K_r - K_{r-1} \quad (3)$$

$$DF_r = \frac{\mathrm{dK}_r + |\mathrm{dK}_r|}{2} \quad (4)$$

The detection function DF_r takes values in the range $[0; \infty)$. Figure 1 illustrates the detection method proposed: the audio (noisy) query is represented on the top figure, the variations of kurtosis on the middle, and the derivative on the bottom. A threshold can be applied since high values may characterize the presence of transients. Experiments show that a value of 4.5 for the threshold leads to good results (see section 4). Furthermore, a filter can be applied to avoid the detection of two proximate transients. If two transients are detected within 512 samples, only the higher derivative value is preserved. This choice is justified by the physical difficulty for general public users to produce too fast taps.

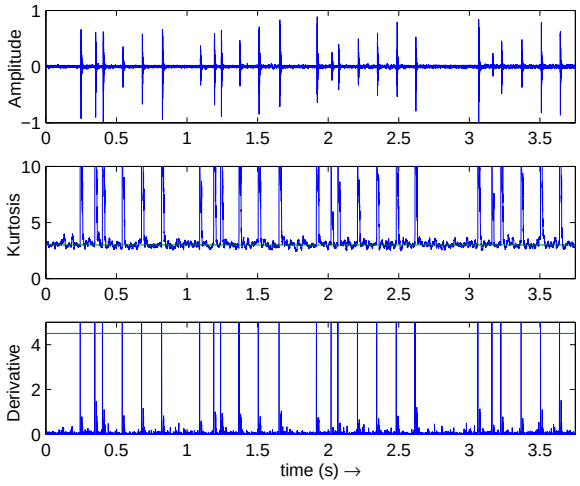

Figure 1. Illustration of the transient detection method applied to an audio query : waveform (top), kurtosis computed on 512 samples frame (middle) and derivative of the kurtosis with chosen threshold (4.5) (bottom).

3. ALIGNMENT OF SEQUENCES

Robust QBT systems have to take into account potential spurious errors in tapping [6]. However, experiments show that the similarity measures which perform well with synthetic queries (no error) do not necessarily yield good results with manually entered queries [8]. The QBT presented here is dedicated to the retrieval of music from manual queries, eventually with errors. This assumption leads to the consideration of approximate string matching techniques. One of these techniques is local alignment [13]. The adaptation of this method is presented in this section.

3.1 Representation of Rhythmic Patterns

The first choice for applying string matching algorithms is the choice of the representation of the rhythmic patterns analyzed from the audio query. The transient or onset detection estimates the times of rhythmic events. These events are due to note onsets or taps/claps. The representation chosen is a sequence of Inter-Onset Intervals (IOI). These IOIs represent the time intervals between two successive rhythmic events. For example, Figure 2 represents an excerpt of the musical piece *Happy Birthday*. The associated IOI sequence is: 90, 30, 120, 120, 120, 240, 90, 30, 120, 120, 120, 240, 90, 30, 120, 120, 120, 120 (with tick MIDI as time unit).

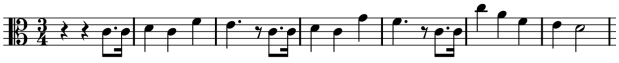

Figure 2. Musical score of an excerpt of *Happy Birthday*.

A few representations for these IOIs can be considered. For example, IOIs can be represented as absolute values in time units (seconds for example). Nevertheless, it is important to note that users may propose an audio query similar to the musical piece requested, with no respect for the tempo. In this case, the two sequences (query and musical piece tested) are totally different whereas they are musically very similar. Therefore, we propose to encode the IOIs as relative values. Each rhythmic event is represented by the ratio of the current absolute IOI by the previous one. For example, the excerpt shown by Figure 2 can be represented by the sequence: 1, 0.33, 4, 1, 1, 2, 0.375, 0.33, 4, 1, 1, 2, 0.375, 0.33, 4, 1, 1, 1. This way, a same sequence represents a musical piece played at a different tempo.

3.2 Adaptations of Local Alignment Algorithm

As previously seen, users tapping or clapping rhythmic patterns may not always be musicians, and may thus generate a query with spurious errors. Furthermore, onset or transient detection algorithms may also estimate wrongly the rhythmic patterns requested. All these potential errors may have a large impact on the quality of a music retrieval system such as QBT. Therefore we propose to adapt approximate string matching methods that have been experimented as very precise and efficient for the estimation of the melodic similarity [4, 14].

Among several existing methods, Smith and Waterman's approach [13] consists in detecting local similar areas between two sequences. This *local alignment* or *local similarity* algorithm finds and extracts a pair of regions, one from each of the two given strings, that exhibit high similarity. A similarity score is calculated by considering elementary operations that transform one string into the other. The operations between sequences include deletion, insertion of a symbol, and substitution of a symbol by another. This similarity measure makes use of the dynamic programming principle to achieve an algorithm with quadratic complexity.

Adaptation to the specific problem of estimation of similarity between rhythmic patterns requires the definition of the elementary operations. The costs associated to the insertion and deletion are fixed and are the same. Concerning the substitution cost $S(x, y)$ between IOIs relative values x and y, a positive score is associated to a match ($x = y$) and a penalty (negative score) is given in case of mismatch. This score is related to the ratio between the values of the relative IOI sequences. However, since mismatch due to errors in sound analysis or in the query may occur, we propose to match two IOI relative values even if they are not exactly the same. A threshold $T_m > 1$ is fixed. If the IOI ratio is lesser than T_m, the substitution cost is positive, but lesser than the cost associated to perfect match (denoted S_{match}), depending on the IOI ratio. Then, if the IOI ratio is greater than T_m, a fixed penalty score, denoted S_{mismatch} is given:

$$\begin{aligned}
S(x,y) &= S_{\text{match}} && \text{if} && x = y \\
S(x,y) &= S_{\text{match}} + 1 - \frac{x}{y} && \text{if} && 1 < \frac{x}{y} < T_m \\
S(x,y) &= S_{\text{match}} + 1 - \frac{y}{x} && \text{if} && 1 < \frac{y}{x} < T_m \\
S(x,y) &= S_{\text{mismatch}} && \text{else} &&
\end{aligned} \tag{5}$$

4. EXPERIMENTS

TODO

5. ACKNOWLEDGEMENT

6. REFERENCES

[1] Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith. Query by Humming: Musical Information Retrieval in an Audio Database. In *Proceedings of the ACM Multimedia*, pages 231–236, 1995.

[2] Bryan Pardo and William Birmingham. Query by Humming: How Good Can It Get ? In *Proceedings of the Workshop on Music Information Retrieval, SIGIR*, Toronto, Canada, 2003.

[3] Roger B. Dannenberg, William P. Birmingham, Bryan Pardo, Ning Hu, Colin Meek, and George Tzanetakis. A Comparative Evaluation of Search Techniques for Query-by-Humming Using the MUSART Testbed. *Journal of the American Society for Information Science and Technology (JASIST)*, 58(5):687–701, 2007.

[4] Pierre Hanna, Pascal Ferraro, and Matthias Robine. On Optimizing the Editing Algorithms for Evaluating Similarity Between Monophonic Musical Sequences. *Journal of New Music Research*, 36(4):267–279, 2007.

[5] Jouni Paulus and Anssi Klapuri. Measuring the Similarity of Rhythmic Patterns. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, Oct. 2002.

[6] Gunnar Eisenberg, Jan-Mark Batke, and Thomas Sikora. Beatbank - an MPEG-7 compliant query by tapping system. In *116th Convention of the Audio Engineering Society*, Berlin, Germany, May 2004.

[7] Jyh-Shing Roger Jang and Hong-Ru Lee. Hierarchical Filtering Method for Content-Based Music Retrieval via Acoustic Input. In *Proceedings of the ninth ACM International Conference on Multimedia*, pages 401–410, Ottawa, Canada, 2001.

[8] Alexandra L. Uitdenbogerd. *Music Information Retrieval Technology*. PhD thesis, RMIT University, Melbourne, Australia, July 2002.

[9] Rainer Typke and Agatha Walczak-Typke. A tunneling-vantage indexing method for non-metrics. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 351–352, Philadelphia, USA, Sept. 2008.

[10] Simon Dixon. Onset Detection Revisited. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 133–137, Montreal, Canada, September 2006.

[11] Paul M. Brossier. *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, Centre for Digital Music, Queen Mary College, University of London, UK, 2006.

[12] P. Masri and A. Bateman. Improved modelling of attack transients in music analysis-resynthesis. *Proceedings of International Computer Music Conference (ICMC'96), Hong-Kong, China*, pages 100–103, 1996.

[13] T.F. Smith and M.S. Waterman. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[14] Marcel Mongeau and David Sankoff. Comparison of Musical Sequences. *Computers and the Humanities*, 24(3):161–175, 1990.