MUSIC RHYTHM CHARACTERIZATION
WITH APPLICATION TO WORKOUT-MIX GENERATION

Qian Lin1, 2, Lie Lu1, Christopher Weare3 and Frank Seide1

1Microsoft Research Asia, Beijing, China
2Department of Computer Science, Xiamen University, Xiamen, China
3Microsoft Corporation, Redmond, USA

ABSTRACT

In this paper, we present approaches to musical rhythm pattern extraction, rhythm-based music retrieval, and rhythm-synchronized music mixing. A probabilistic model is used to jointly estimate tempo and time signature as a basis for beat tracking and measure detection. A representative rhythm pattern is then extracted through clustering to characterize the rhythm of a song. Based on this, a probabilistic approach is used for retrieving songs with similar rhythmic patterns. These are then mixed rhythm-synchronously with transitions maintaining continuity and regularity of beats. We apply the presented methods into workout-mix generation, which aims at automatically selecting rhythmically similar music given a seed song and a user-defined tempo profile. Our probabilistic approaches achieve accuracies similar to best published results, but avoid manually tuned parameters and “fudge factors”.

Index Terms—rhythm pattern, rhythm-based retrieval, tempo induction, workout-mix

1. INTRODUCTION

Rhythm is one of the most important features in music perception and characterization. Wikipedia currently defines rhythm as the “variation of the length and accentuation of a series of sounds or other events” [1], including how regular or irregular the melody and harmony is, strength/weakness of the relation among notes, and tempo. Rhythm is an important component in many music applications. The present paper aims specifically at the scenario of workout-mix generation: to automatically select rhythmically similar music to match a user-designed tempo profile defining a user’s workout pace (jogging, spinning, aerobics), and to mix the selected songs with rhythmically smooth transition maintaining continuity and regularity of beats. For this, we need to characterize the rhythm of a song, find music with a similar rhythmic pattern, and mix songs with synchronized rhythm, which in turn requires accurate estimation of tempo, time signature, beat and measures.

Some of the abovementioned techniques have been investigated in the field of music information retrieval. Various algorithms of tempo induction were compared and discussed in [2], [3] presented a novel method formulating the tempo estimation in terms of a nearest neighbor classification problem. [4] induced tempo and time signature through an HMM process based on per-band “accent signals.” [5] utilized signal processing techniques like comb filter analysis to do beat-tracking. [6, 7] proposed an approach to extract the rhythmic pattern from music and to measure their similarity, including tempo induction and beat tracking. [12] presented a method to mix a wider class of songs with dance music (e.g., Mozart with techno) by jointly optimizing the energy alignment of both signals instead of attempting to detect individual beats/tempos. Despite intensive research, many above topics still remain challenging. For example, tempo estimates are prone to errors of detecting halves or multiples, and there is still no systematic way for rhythm-based music retrieval and mixing.

In this paper, we address tempo estimation, rhythm-based music retrieval, and mixing. We use a novel probabilistic approach to jointly estimate tempo and time signature, based on which a music clip is segmented into measure-aligned rhythmic patterns, from which a representative rhythmic pattern is extracted to characterize a song. Furthermore, we present an approach to music retrieval based on a probabilistic model of rhythmic pattern similarity, and a rhythm-synchronized music mixing method. These technologies are finally integrated for use in automatic workout-mix generation.

The remainder of this paper is organized as follows: Section 2 proposes our approach to joint tempo and time signature induction. Section 3 presents the method of rhythmic patterns extraction, rhythm-based music retrieval, and rhythm-synchronized music mixing. Workout-mix generation is described in Section 4, and experimental results are shown in Section 5.

2. TEMPO AND TIME SIGNATURE INDUCTION

A fundamental component of music rhythm characterization is the robust estimation of tempo and time signature. The major feature we use for this task is an autocorrelation curve, derived from amplitude envelopes and onset curves of a set of octave-scale sub-bands [10]. Autocorrelation curves are a common technique for tempo induction, e.g. [2, 8, 9], due to its apparent periodicity. In our approach, we use it as an input to a probabilistic model for jointly estimating music tempo and time signature.

Fig. 1 shows an example autocorrelation curve, where we can see regular peaks indicating tempo and signature information. While previous work on autocorrelation-based tempo induction usually use complex heuristics [8] or template matching [9], we build a probabilistic model to jointly estimate tempo and time-signature, in order to incorporate the strong relationship between tempo and time-signature, and to reduce the double-tempo or half-tempo confusions.

In addition to the autocorrelation feature, we also utilize a supplementary feature—OPM (onsets per minute)—to represent the
relationship between onsets and tempo. Intuitively, OPM is monotonously correlated to tempo. Integrating these features, the posterior probability of a candidate tempo period $T$ can be represented as,

$$P(T | A, O, \lambda_m) \sim P(A | T, \lambda_m) P(O | T) P(T)$$ (1)

where $A$ is the autocorrelation feature, $O$ stands for OPM, and $\lambda_m$ is tempo model under time signature $m$. $P(O | T)$ and prior $P(T)$ are modeled as GMMs in our approach. The time signature dependent tempo model is addressed below. In our implementation, only three time signatures (4/4, 3/4, and 2/4) are considered, and other less frequent combinations are ignored.

Each interval between every two peaks in the autocorrelation curve are considered a candidate tempo period. The optimal tempo and time signature $[T, m]$ can be jointly estimated through an exhaustive search across all possible tempo/time-signature candidates to maximize the posterior as given in Eq. (1):

$$[T, m] = \max_{[T,m]} P(A | T, \lambda_m) P(O | T) P(T)$$ (2)

### 2.1. The Tempo Model

The tempo model is built from the autocorrelation curve based on the relative amplitude of each beat in a measure. Since time signature labels are not available in our training data, we first need to estimate the amplitude and the number of the beats in a measure. We assume that the highest peaks indicate measure boundaries. For example, in Fig. 1, the 4th, 8th, 12th peaks are taken as measure boundaries, and we estimate the measure length to be around 1 second. Thus, if the labeled tempo is 120 BPM (0.5s per beat), we determine there are 2 beats in a measure, and the time signature is 2/4; otherwise if the labeled tempo was 240 BPM (0.25s per beat), there would be 4 beats in a measure, and the time signature would be determined as 4/4.

After determining the time signature, the amplitude of each beat in a measure can also be determined. We assume that the first beat is at the measure boundary and has the largest amplitude, and model the relative amplitudes of other beats with regard to the first beat with a GMM (the relative amplitude of the first beat is always 1). Fig. 2 illustrates time signature dependent tempo models for time signatures 4/4, 3/4, and 2/4.

### 3. RHYTHM ANALYSIS

We assume music rhythm to be regular and repetitive, consisting of a repetition of a fundamental unit. This fundamental unit presents the beat structure of the music. It tells how strong or weak each beat is, and reflects how rough or smooth the music is. In this paper, we choose the measure as the fundamental unit and use term “rhythmic pattern” to describe the rhythm information in each measure, as Fig. 3 shows.

With tempo and time signature estimated as described, beat tracking is now performed to locate beat positions, and the measure boundaries are determined by assuming a measure boundary coincides with the strongest beat. In this way, a piece of music is partitioned into a sequence of measures, from which a representative rhythm pattern is determined, which is the basis for retrieval and mixing. In the following sections, we will address in detail the rhythmic-pattern extraction from measures, rhythm-based music retrieval, and rhythm-synchronized music mixing.

#### 3.1. Rhythmic Pattern Extraction

In [7, 12], the overall amplitude envelope derived from the audio signal is utilized to represent the rhythmic pattern in each measure, yielding the information of beat strength as well as information between beats. Like [7, 12], we also utilize the amplitude envelope to represent rhythm, but, instead of the overall amplitude envelope, we choose to use the percussion amplitude envelope (which is extracted from specific sub-bands), since the music rhythm is usually performed with strong drums or percussions, which is especially true for the music used for workout.

To account for rhythm varying across the entire music and deviations among the rhythmic patterns in measures, we extract a most representative rhythmic pattern to characterize the rhythm of the whole music, by applying a K-means clustering and using the centroid of the biggest cluster to indicate the most representative pattern. Fig. 3 shows a set of measure-based rhythmic patterns (on the left) and the obtained representative rhythmic pattern (on the right) of an example music.

#### 3.2 Rhythm-based Music Retrieval

With the extracted representative rhythmic patterns, we can estimate the rhythm similarity between music clips and retrieve music with similar rhythm. In this section, we will present an approach to
rhythm-based music retrieval with a probabilistic Gaussian model of rhythm similarity.

As tempo may be different between two music clips, alignment between their representative patterns is necessary. We define the rhythm similarity between two patterns as,

$$\text{sim}(D \mid Q) = \max_D P(D \mid A, Q)P(A)$$  \hspace{1cm} (3)

where $Q$ is a query rhythmic pattern, $D$ (document) is a pattern of a song in the dataset, and $A$ is an alignment between $D$ and $Q$ to consider the tempo difference, in our implementation, by linear scaling.

In equation (3), $P(A)$ penalizes the tempo difference. It can be modeled with a Gaussian, or can be ignored in the applications which don’t care about the tempo difference but only care about the shape of the rhythmic pattern. $P(D \mid A, Q)$ is the similarity between $Q$ and the aligned $D$, which is also a Gaussian with aligned $D$ as observation and $Q$ as the expected mean of the observation patterns. It can be written as (in a log format),

$$\log P(D \mid A, Q) = -\frac{1}{2n} \sum_{i=1}^{n} \left( \frac{d_{i} - q_i}{\sigma_i^2} \right)^2$$  \hspace{1cm} (4)

where $n$ is the length of rhythmic pattern, $d_{i}$ and $q_i$ are the points of the aligned pattern $D$ and query pattern $Q$, and $\sigma_i$ is the corresponding variance. The variance can be set flat assuming each point is equal important. In our implementation, the variance is heuristically set based on the deviation of query points from their mean:

$$\frac{1}{\sigma_i^2} = (q_i - \mu_q)^2$$  \hspace{1cm} (5)

where $\mu_q$ is the mean of query $Q$, to reflect the intuition that when measuring the similarity between two rhythmic patterns, peaks and valleys of a pattern are the key points representing rhythm and thus should be emphasized (peaks and valleys that are far from the mean are emphasized with a strong weight). It is noted that the variance is clipped in implementation to avoid zero or infinite variance, and since the variance is determined based on query pattern, the obtained similarity (3) is asymmetric.

3.3 Rhythm-Synchronized Music Mixing

Rhythm-synchronized music mixing is used in workout-mix generation, where two adjacent songs are mixed to produce a smooth rhythmic transition without losing the continuity and regularity of the beats. The necessary beat alignment is based on the timing information obtained in the above procedure. If two adjacent songs have different tempo, the tempo is adjusted gradually: Given two music segments for mixing, with each segment containing several beat periods, we expand or shrink the beat periods (which were originally constant) of each segment to a series of gradually changed period by Synchronized Overlap-Add (SOLA), with their corresponding beats kept aligned. These two scaled rhythm/beat-synchronized segments are mixed together by cross-fading. Subjectively, the smoothness of transitions gives the mix a “professional feel” with rhythm changes sometimes being barely noticeable. Fig. 4 illustrates such a music mixing example, with transition from music 1 (beat period: 50) to music 2 (beat period: 43). 8 beat periods of both music clips are considered for mixing. Finally a mixed segment is generated, with beats aligned and beat period gradually changed from 50 to 43.

4. WORKOUT-MIX GENERATION

In workout exercises like jogging, aerobics, and spinning, tempo profiles are used to define the pace of the exercise. Music playlist (or workout playlist) based on the tempo profile are a common tool: exercisers can simply follow the music beats to drive their pace. The Internet is a great source for workout playlists created and shared out by hobbyists and professionals.

The objective of our workout-mix generation is to take a predefined tempo profile, automatically select rhythmically similar music which matches the profile, and create a continuous music mix from it.

The proposed approaches in Section 3 provide all tools to implement such functionality. A given tempo profile is first used to select music candidates from the user’s music database (their tempo having been automatically determined). Then, while the first song in the playlist can be randomly selected from the candidates with matching tempo, selecting subsequent songs should not only consider the prescribed tempo but also rhythm similarity between adjacent songs. Our implementation also considers other factors of similarity based on our previous work [11], allowing users to provide a seed song to indicate what kind of music they want to listen.

5. EXPERIMENT

5.1 Tempo Induction

We evaluate our method on three datasets: “ccMixter” from ccmixter.org as well as “ballroom” and “songs” from the MIREX benchmark (available online [13] and commonly used [2, 3]). The “ccMixter” set contains 2165 songs with different styles. “Ballroom” contains 698 ballroom-music excerpts of 8 different genres, each around 30 seconds, and “songs” contains 465 excerpts, each around 20 seconds.

Following [2], two evaluation metrics are used in our experiment. One is tempo accuracy with 4% tolerance of the ground-truth tempo (acc1), and the other is the accuracy disregarding the errors of half, double, three times, or one third of the ground-truth tempo (acc2). To be comparable to [2], we apply 10-fold cross validation on each dataset, that is, training on 9 out of 10 disjoint subsets, and test on the other subset.

Table 1 shows the results on the three datasets, comparing the effectiveness of different components in tempo estimation. Fig. 5 shows a comparison of results of our approach and various benchmark systems summarized in [2]. Our approach is at least in the top 2 best systems. For example, in the “ballroom” set, we achieve the
5.2. Rhythm-based Music Retrieval

Since no ground-truth rhythm labels are available, it is difficult to directly evaluate the proposed rhythm-based similarity measure. Instead, in our experiment, we built a system to classify different genres of dancing music in the “ballroom” set, which contains 8 different genres, such as chacha, jive, waltz, etc. Since the rhythmic pattern is generally different among most of the genres (some may still have similar rhythmic pattern, such as between Viennese Waltz and Waltz), evaluation on the classification result can indirectly reflect the performance of the proposed approach to rhythm similarity measure and rhythm-based music retrieval.

In the experiment, we choose the cluster number for representative pattern extraction as 2 (see Section 3.1), since the length of each clip in dataset is only around 30 s. 10-fold cross validation is used for evaluation, and a naive 10-NN (nearest-neighbor) classifier is used for dancing-music genre classification. The results are shown in Table 2, comparing different options in rhythmic pattern representation and rhythm similarity measure. It can be seen that using the percussion energy envelope leads to a 5% improvement over using the overall energy envelope, and the heuristic variance (Eq. (5)) yields another slight improvement. Our approach is slightly better than the rhythm-based classifier presented in [7], which manually determines the first measure boundaries to obtain error-free measures in rhythmic pattern extraction.

5.3. Workout-Mix Generation

Examples are generated based on a number of pre-defined typical tempo profiles, and are downloadable at http://research.microsoft.com/en-us/um/people/llu/WorkoutMix/WorkoutMix.html. For demonstration purposes, only a 5-to-10s segment from each song is chosen in these example workout-mixes and 8 beat periods are used in music mixing. A preliminary small-scale user study was positive, both w.r.t. candidate selection by rhythmic-pattern similarity and the rhythmically smooth transitions. Suggestions for further improvement included considering instrument or vocal similarity in music selection.

6. CONCLUSION AND DISCUSSION

In this paper, we have proposed an approach to characterize music rhythm and applied it to workout-mix generation. For this task, we have proposed a probabilistic method for jointly estimating tempo and time signature, and presented methods of rhythm-based music retrieval and rhythm-synchronized music mixing. For future improvements, we are considering approaches to utilize finer structures of rhythmic pattern, overall structure, and the use of more perception-related features for selecting music that is not only similar in rhythm but w.r.t. other aspects of music similarity.

7. REFERENCES