CS229 Autumn 2005
Final Project Report

Music Tempo (Speed) Classification

Yu-Yao Chang, and Yao-Chung Lin
{yychang, yclin79}@stanford.edu

Abstract

Music tempo (speed) is one of the most important features of a song. With successful classification of the tempo of song, content-based music browsing may utilize this feature and search/recommend songs in the same category of tempo. With some paper survey we noticed that prior works were focused on the numerical tempo value instead of the human perceptual speed of song. We proposed Inter-Onset Interval (IOI) reliability feature to present a soft tempo information for SVM. The onset detection algorithm in [4] was implemented to extract the beat onset timing, and the detection result is applied to SVM [6] as training features. The model trained via SVM can classify slow and medium songs with high accuracy; however, the fast songs are not highly correctly classified. We consider improving the feature for fast songs as future work.

Introduction

This project presents a music tempo (speed) classification with machine learning for content-based retrieval and recommendation. In the following of this paper we use the term “tempo” to represent the number of beats per minute described on the music score and “speed” as the human perceptual speed of the song. The tempo can be utilized as a feature for searching/browsing the song database. However, recent works on tempo detection consider all multiples of the target tempo value a “successful” estimate, which is an intolerable error in the sense of speed.

Moreover, the tempo on the score may not correctly reflect the speed of the song. It is possible that a song of 3/4 time signature sounds "faster" than that of 4/4 time signature while they have identical tempo. Although there are the same tempo quantities from these two time signatures, human usually catches the strongest beats which appear more frequently in songs of 3/4 time signature.

Each user may catch different tempo feature. For example, one may catch the bass drum beats, while the other may catch the vocal. In addition, there is no exact boundary for perceptual speed categories. A slow song for one listener may be considered of medium speed for the other. Therefore, personalized classification policies will be applied for different end-users.

From the problems above mentioned, we plan to train a machine that can correctly predict and learn the speed classes of songs for different end-users which have different speed sensitiveness.

Related Works

Many recent works are related to this project. According to our paper survey, musical genre classification, beat tracking, and musical expression will be discussed in this section.

A musical genre classification is presented in [1]. They extract the timbral texture features and rhythmic content features to classify an audio signal belonging to its genre, such as pop, rock, jazz, classic, etc. The timbral texture feature can be used for separate music and speed signal. Rhythmic content features, including temporal envelope and beat histogram, are used for classify musical genre. This work classifies the music into different genre according to its ground truth. In contrast, our work classifies the music into different perceived tempo speed according to the end-user feed back. However, these features are good hints for our tempo classification.

[2], [3], [4], and [4] provide several beat detection and tracking methods. In [2], they adopt filter bank and comb filter to analysis the signal, and detect the beat with their resonator and tempo analysis. On the other hand, [3], [4], and [4] use onset detection methods. In [3], several measures for spectral differencing, including Euclidean distance, Kullback-Liebler distance, and Foote distance, are discussed. They detect the change point by picking the highest peak in the unsmoothed measurement functions. The method in [4] finds the onset times in spectrogram, and detects the bass drum and the snare drum by examining the extracted onset components. Further, they generalized their beat tracking system to detect hierarchical beat structure with model based beat tracking method.

The prior works provide several guidelines for the features related the tempo speed. We are going to adopt them, develop a classification method and analysis the useful features.

Feature Extraction

We extract 8-sec waveform signals from MP3 files with 44.1kHz sampling rate. It is difficult for SVM to learn a model from raw waveform since the raw waveform has highly correlation in time domain. Therefore, some reasonable features should be extracted for machine learning.

The rhythmic content, such as beat information, is considered as the most important feature which is related to the human perceived tempo and speed. Several tempo detection algorithms utilize the onset time detection and do further process. We also extract the feature from onset information. Regarding the onset detection algorithm, we consider a fast onset detection algorithm proposed in [4].

Based on the onset information, we develop a feature, named Inter-Onset Interval (IOI) reliability for SVM input. The IOI reliability represents how probable the tempo of a given song is. The SVM utilize this "soft" tempo information to build the model for classification.

I. Onset detection

The onset time is defined as the beginning time of a beat or note played. To track a beat, we apply spectrogram analysis to raw signals with 4096 window size, and 7/8 overlap. Figure 1 shows a raw wave signal and its spectrogram. The spectrogram represents the power of different frequencies at different time indices. Let *p(t,f)* being the spectrogram of given signal. The degree of onset *d(t,f)* is given by

$$d(t,f) = p(t,f) - pp + \max(0, p(t+1,f) - p(t,f))$$

Where $pp = \max(p(t-1,f), p(t-1,f\pm1), p(t-2,f))$, Finally, the degree of onset is a function of time and given by

$$D(t) = \sum_f d(t,f) \qquad (1)$$

Top plot in Figure 2 shows the degree of onset with respect to frequency and time, and the bottom plot shows the degree of onset, *D(t)* on which we are going to do further process. From the plots, we assume *D(t)* catches most onsets from the raw signal.

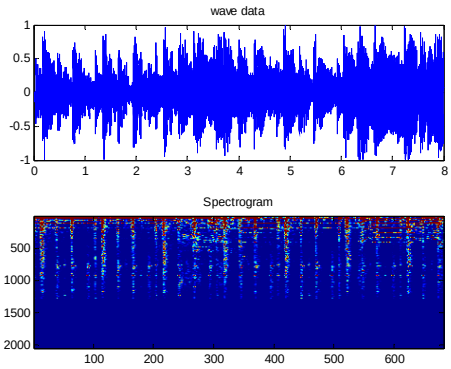

Figure 1: The raw data is the wave form, and its spectrogram.

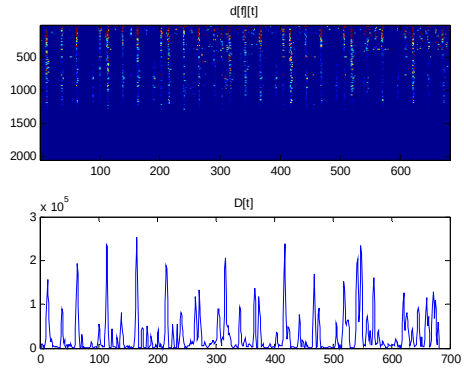

Figure 2: The output of our implemented onset degree

II. Inter-onset interval (IOI) reliability

The degree of onset *D(t)* is phase-dependent and is not a good feature sequence for SVM Learning. We proposed a soft tempo information, which is called Inter-Onset Interval (IOI) reliability, derived from the degree of onset, *D(t)*.

Recalling that we mentioned in introduction, different beat patterns will yield the different perceptual speed. For example, pattern #1 "strong weak weak; strong weak weak" will be slower than pattern #2, "strong strong strong; strong strong strong". Therefore, we proposed a method to evaluate total onset strength as well as the average interval between two adjacent onsets.

For each IOI candidates, say IOI=*k*, we begin at *D(0)* and move forward *k* frames and align to the local maximum of the degree of onset. We repeat this step until we reach the end of *D(t)*. And we average those visited local maximum as our reliability score. We also examine all possible phases and report the one of highest reliability score for IOI=*k*.

The pseudo-code of the algorithm is in the following

```
1     Input: D(t), the onset degree
2     Output: R(T), the IOI reliability score
3
4     For all IOI, T=1..T_max
5         For all possible phase, p < T
6             S(p) = sum D(t+nT + p) over n;
7         End
8         S* = max S(p) over p;
9         N = argmax ( (n+1)T < domain(D))
10        S*(T)=S*/N;
11    End
12
13    S* := S* - min(S*);
14    R = S* / sum(S*);
```

Figure 3: Pseudo code for evaluating IOI reliability

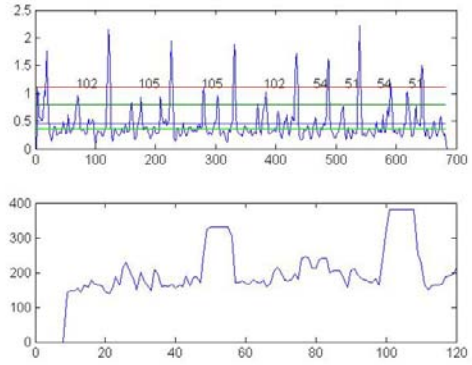

Figure 4: D(t) (top) and corresponding IOI reliability (bottom)

In this algorithm, we loop over possible inter-onset interval from 9 to 120 in unit of the parameter of *D(t)*. Figure 3, line 9~10 assign the score of IOIs to the average peak score. Finally, Figure 3 line 13~14 normalize the score and obtain the IOI reliability score. An example of IOI reliability score is shown in Figure 4.

Amount several features we have tried, the IOI reliability gives the best accuracy.

Experimental Results

Several experiments are conducted using the features in sections described above. First, the data set used for training and testing is described. Then, the experimental setup and results are presented.

We selected 86 popular songs, in which 29 are labeled as fast, 31 are labeled as medium, and 26 are labeled as slow. We excerpt 8-second (1:00.00~1:07.99) waveforms from the selected songs. The waveforms are sampled at sampling rate 44.1 kHz, stereo two channels. We take mean of two channel signals to be our input of feature extractor.

Regarding the learning model, we use the libSVM library in [6] for multi-class classifier. We take leave-one-out-cross-validation (LOOCV) error as well as training error for measurement. Based on our limited size of data set, the closer between training error and LOOCV error the more reasonable result is.

In our experiment, we got 70.93% of training accuracy, and 69.8% LOOCV accuracy. We also analysis the classification results and make a confusion matrix to understanding the misclassification situation.

Table 1 shows the confusion matrix of our experiment. As Table 1 shows, the fast songs are easy to be classified to other two classes and yields lower accuracy, less than 60%. The slow songs are much easy to classified and yield a high accuracy, 85%. The random guess yields 34%, 36%, and 30% accuracy for fast, medium, and slow classes respectively.

Table 1: confusion matrix

Label / Clsfd	Fast	Medium	Slow
Fast	58.6%	29.0%	0. 0%
Medium	20.7%	71.0%	15.4%
Slow	20.7%	0.0%	84.6%

Conclusion

This report has described the human perceived tempo speed problem and its application. The machine learning technique is used for music tempo classification with our proposed feature extraction. We consider three tempo classes for songs, i.e. fast, medium, and slow. One of beat detection methods is used for our feature extraction, which is onset detection proposed in [4]. Base on the onset degree, we have proposed a feature extraction, IOI reliability score, for classifying the human perceived musical tempo speed. The IOI reliability score vector can be considered as soft tempo information over all possible tempo speed, which gives high score if the song has high probability to be in that tempo speed. The SVM learns a model from this feature and yields 85% for slow songs, but for fast songs, it only gets 59% accuracy.

Discussion and Future Work

A bunch of feature candidates were also evaluated: the partial spectrogram of the waveform, degree of onsets *D(t)*, the sequence of IOI, the indices that onsets occurs, and so on. It turns out that the model trained by IOI reliability provides highest classification accuracy.

In general, our accuracies of medium and slow songs are fair. However, we are suffering a high error on classifying fast songs. A possible reason is that we perceive the speed of a song basically according to the vocal (if exists). But in some cases the background beats are intense and dominate the IOI reliability score, making the SVM confused with the authentic fast song IOI pattern. Figure 5 demonstrates a slow song "Rain" by Madonna, with intense beats detected.

Another issue is that the song labeling done by human subject is not robust. A lot of ambiguous labeling decisions were encountered when we labeled the training data. We noticed that one may have different labeling result to the same song in different time. Since SVM relies on the training examples close to the boundaries, it turns out that the ambiguities in our training examples could deteriorate the classification accuracy. In our labeling processing, we tend to believe that there are factors other than the onset strength and pattern that affects human's perception on music speed. The knowledge to a certain genre, for example, could bias the perceptual speed due to past experience.

To extend our work, we plan to separate the vocal and instrumental music and work out the IOI only on the vocal part. We also consider splitting the spectrogram into several frequency bands and extracting significant frequency indices.

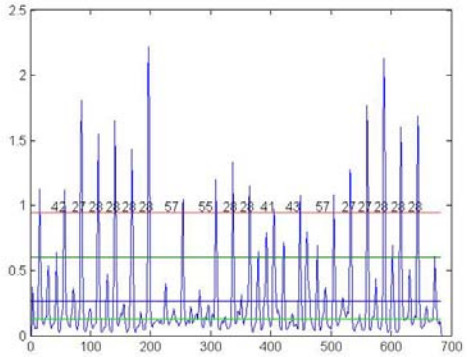

Figure 5: The onset detection of the song "Rain" by Madonna.

Reference

[1] George Tzanetakis, and Perry Cook, "Musical Genre Classification of Audio Signals," IEEE Transcation on Speech and Audio Processing, VOL. 10, NO. 5, July 2002

[2] Eric D. Scheirer, "Tempo and Beat Analysis of Acoustic Musical Signals," Journal of Acoustical Society of America, 1998

[3] Stephen Hainsworth, and Malcolm Macleod, "Onset Detection in Musical Audio Signals," In proceeding of the International Computer Musical Conference, 2003

[4] Masataka Goto, Yoichi Muraoka, "A Beat Tracking System for Acoustic Signals of Music," ACM Multimedia 1994

[5] Masataka Goto, "An Audio-Based Real-time Beat Tracking System for Music With or Without Drum-sounds," Journal of New Music Research, 2001, Vol. 30, No.2, pp.159-171

[6] Chih-Chung Chang, and Chih-Jen Lin, "LIBSVM -- A Library for Support Vector Machines", Available on http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[7] Gerhard Widmer, "The Musical Expression Project: A Challenge for Machine Learning and Knowledge Discovery," In proceeding of the 12th European Conference on Machine Learning (ECML) 2001