Beat Induction

Peerapong Dhangwatnotai, Rajendra Shinde, Pawin Vongmasa

15th December 2006

Abstract

Previous works [1,3] have focused on finding the tempo and the beats in music given the entire signal as input. The goal of this project is to find an online algorithm for real time beat detection. Our approach is to discretize the audio signal into short frames and make a prediction on whether the current frame is a beat or not depending on the current and previous frames. We propose a novel two-part classification system for beat detection. The first part of the prediction uses a classification algorithm such as logistic regression and support vector machines to predict the probability of the current frame being a beat. The second part involves the use of the temporal location of the previous beats to predict the probability of the current frame being a beat. Thus while the prediction of the first part is based on the current features, the prediction of the second part is based on the onsets of beats in the past.

1 Introduction

When we listen to music, we never find any difficulty in tapping along the beat, but how exactly do we do that? Is it possible to make the computer do the same task? For a couple of years, the problem of automated beat induction has been studied as a step toward better understanding of human perception. Many models have been proposed for beat perception, none of which are perfect, and there is a general consensus that it is impossible to find a perfect model because the formal definition of beats does not exist. Because of this intrinsic ambiguity in the notion of beats, it is fair to expect a beat induction algorithm to work on a set of specific assumptions regarding beat period or specific inputs. Possible applications include music classification based on rhythm, synchronizing machine accompanists with human performers, and in automated transcription systems, visualizations to music.

2 Background and Related Works

2.0.1 Beat Induction Algorithms for Symbolic Input

Many researchers have tackled the problem of beat induction from different levels and with different approaches. Longuet-Higgins and Lee([1]) were the first to initiate the idea of beat induction on symbolic input (beat onsets have been annotated). After this there have been many algorithms based on input in MIDI format which use onset information and Inter-Onset Intervals (IOIs, meaning the time intervals between onsets) and other features to build up hypotheses of beats. Desain ([2]) gave good comparisons and discussions on some prominent algorithms by Longuet-Higgins and Lee ([1, 3]).

2.0.2 The Role of Machine Learning

Recently, Gouyon et al. ([4]) viewed the problem differently, and showed that the best indicator for identifying whether a beat occurs in the given time-framed signal or not are energy features. They got this result from running ten-fold cross validation to choose the most relevant features. (Note that this is a classification problem where the given query is a short signal and the expected result is whether this signal indicates a beat or not.) This shows us not only that we can use energy features and get a good prediction, but also that it is not necessary to depend on onsets or IOIs we can determine

the beat by seeking periodicity in other features as well. However, we will still need to take IOIs into account in order to get the most probable tempo among its multiples.

3 Database

We used the database for the Mirex 2006, which is a contest for music information and retrieval. The database can be obtained at (http://www.music-ir.org/mirex2006/index.php/Main_Page). This database consists of a set of 20 music files, each having a sampling frequency of 44.1kHz and each being of 30 sec duration. The temporal beat locations derived from 40 different listeners.for each music file have been specified in text files.

4 The Online Beat Induction Problem

The following is our formulation of the online beat induction problem.

The signal of a piece of music, represented as a discrete sampling of the original performance, is grouped into overlapping frames. Each frame contains 1024 samples, which is about 0.0929 seconds. Given frame 1 up to frame t, determine whether frame t is a beat.

5 Difficulty of Beat Induction

The beat induction problem is different from conventional classification problem in many aspects. First, there is no single right answer. Humans perceive beats at different metrical levels. One may perceive a beat with 1 second interval. Another may perceive a beat twice as fast. This is evident in the MIREX 2006 dataset we use in this project. For each piece of music, there is 30-60 sets of labels, each from a different person. For some set of labels, there are twice as many beats as another set. It is impossible to train a classification algorithm with all of the labels because obviously some are contradictory. This is different from having noisy labels because the problem inherently has many sets of correct answers.

The way we get around this is by creating a "standard" set of labels. Let b_{ij} denotes whether person i thinks frame j is a beat. $b_{ij} = 1$ if person i thinks frame j is a beat and $b_{ij} = 0$ otherwise. The sum $\sum_i b_{ij} = \alpha_j$ counts the number of people who think frame j is a beat. Then in our standard set of labels, frame j is a beat iff $\alpha_j > max(\alpha_j)/2$. This should capture the beats at the most salient metrical level (according to the consensus).

Another property of beat induction problem is that the correct set of beats must have a constant inter-beat interval. This is another constraint that conventional classification problem does not have. Below, we suggest an idea about this but do not have time to fully explore this aspect.

6 Our Approach

We break this problem into two parts. First, we assume that it is possible to tell (with high probability) whether the current frame is a beat given the current frame and the last two frames. We want to compute p(beat | the last three frames). This is just a classification problem and we have tried applying SVM and logistic regression to it.

However, it is not always possible to tell whether the current frame is a beat given just the last three frames (or 10 or 20 frames). For example, a silent frame (a frame with almost no sound) can either be a beat or not a beat depending on the previous beats. Human perception of beats continue even after the signal disappears. This can be translated into the constraint that inter-beat interval is constant. Therefore, we also need to model p(beat | onsets of previous beats). Then combine the two parts together.

In this report, we will show how we tackle the first part and propose models for the second part.

6.1 The First Part

Features

We design our features using subband energies which have been reported as being a good feature for beat detection ([4]). More specifically, each frame is converted to the frequency domain using fast fourier transform. The energy of each frequency is the norm of its coefficient. The frequencies are group into 12 subbands and the energy of each subband is the average energy of the frequencies in that subband. Then we compute the ratio of each subband energy to the moving average. Let e_i^t be the energy of subband i at time t. The energy ratio of subband i at time t is

$$r_i^t = \frac{k \cdot e_i^t}{\sum_{j=1}^{k} e_i^{t-j}} \tag{1}$$

where k is the size of the period that we average over. In the experiment, we use $k = 32$. When t is small (we don't have that much data in past), k is $t - 1$.

The feature set for each frame is the subband energy ratios of the last three frames. Formally, it is $(r_1^t, r_2^t, ..., r_n^t, r_1^{t-1}, r_2^{t-1}, ..., r_n^{t-1}, r_1^{t-2}, r_2^{t-2}, ..., r_n^{t-2})$.

Classification Algorithm

We tried SVM and logistic regression as the classification algorithm. We did not get a good result from SVM. We have tried SVM with a polynomial kernel of degree up to 3 and the results are unsatisfactory. SVM almost never classifies a frame as a beat. For this reason and the training time consideration, we decided to continue working with logistic regression.

We trained logistic regression using batch gradient descent and the features described above. The data is obtained from MIREX 2006. The signals are down-sampled to 11025 Hz. 75% of the data is used for training and 25% is used for testing.

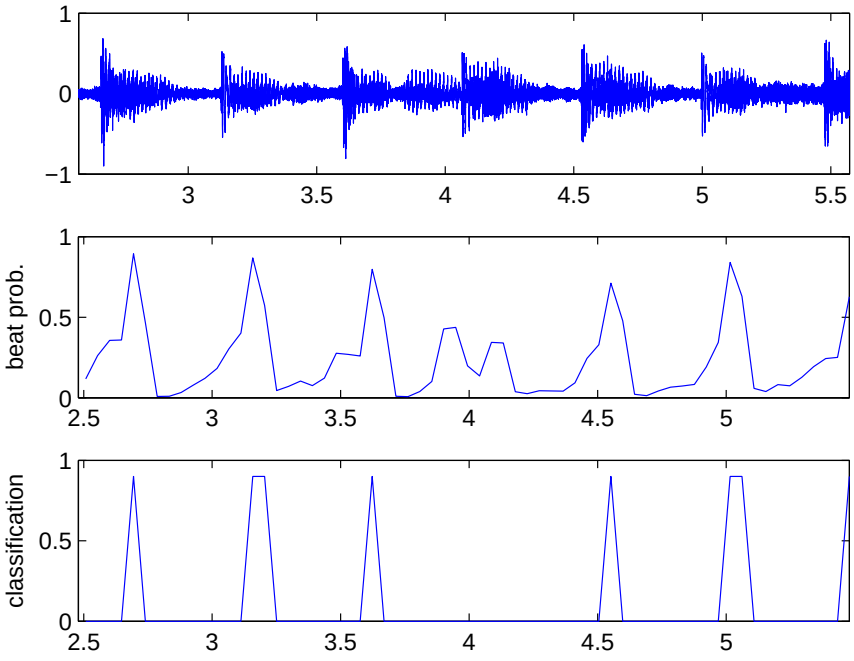

Figure 1: A sample run of the beat detection algorithm. The top plot is the audio signal. The middle plot is the beat probability. The bottom plot is the classification (whether the probability is greater than 0.5)

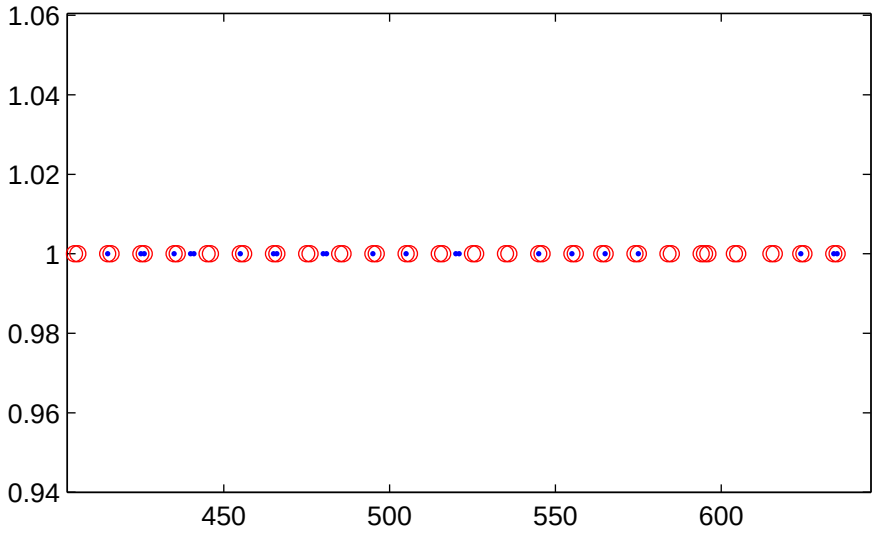

Figure 2: Comparison between the program's output and the label. The x-axis is time. The circles are the beats labeled by human. The dots are the beats output by the program.

6.2 The Second Part

After we have got the probability of the current frame for being a beat or not, we will incorporate this with our guesses made in the past to make a better guess according to the long periodic nature of beats. We models dependency of the present and past data by the following equations:

$$S(t,T) = 1 - \left[\frac{\sum_{\tau=t-\alpha T_2+T-1}^{t-1} \left| BP(\tau) - BP(\tau - T) \right|^p}{\alpha T_2 - T + 1} \right]^{(1/p)} \tag{2}$$

$$T_0(t) = \underset{T}{\operatorname{argmax}}\, S(t,T) \tag{3}$$

$$BP(t) = \gamma S(t, T_0(t)) BP(t - T_0(t)) + \big(1 - \gamma S(t, T_0(t))\big) h(x(t)) \tag{4}$$

where T_1, T_2, α, γ and p are parameters. Their meanings are as follows:

- T_1 and T_2 are the minimum and the maximum number of frames corresponding to the fastest and the lowest tempi respectively. (We choose $T_1 = 4$ and $T_2 = 30$ to indicate the tempo between 40 and 300 beats mean minute.)
- α determines how far in the past we will consider. It must be a positive integer. (We use $\alpha = 2$ in our experiments.)
- γ determines how much we will depend on the past data. It must be a number between 0 and 1. (We use $\gamma = 0.8$ in our experiments.)
- p indicates the way we measure the vector. It is analogous to p-norm of the vector. We used $p = 2$ in our experiments.

Equation (2) defines the tempo probability T at the time t. It can be seen that if BP repeats itself with period T, S will have a high value. Equation (3) picks the most probable period T_0. Finally, equation (4) computes the beat probability by putting the weight between the predicted value of the current frame $h(x(t))$ and the past data $BP(t - T_0(t))$.

7 Results

For the first part, logistic regression's accuracy over the test set (which includes 20 pieces of music) is 0.8537. A sample run of the beat detection program on a piece of music is shown in figure 1. It is more fun and more intuitive to listen to the music while looking at the program's output. Unfortunately, we cannot embed audio into this report. The algorithm can pick out salient acoustic cues of the beats but fails on the more subtle ones. When used with the second part, we got the result as shown in figure 2.

8 Discussion and Future Works

The subband energy ratio is an informative feature for beat detection. While the logistic regression with subband energy ratios achieves the accuracy of 0.8537, it still misses many of the beats as shown in figure 2. This is expected because the algorithm is local, i.e. it only looks at signal around the current frame (about 0.5 seconds), but beats have long range influence on one another (5-10 seconds). An algorithm that uses the information from previous beats such as the one proposed in the second part should improve the accuracy. This can be an item for future work.

References

[1] H. C. Longuet-Higgins and C. S. Lee, Perception of musical rhythms, Perception, 1982.

[2] P. Desain and H. Honing, Computational models of beat induction: The rule-based approach,Journal of New Music Research, 1999.

[3] [3] C. S. Lee, The rhythmic interpretation of simple musical sequences: towards a perceptual model, Musical Structure and Cognition, 1985.

[4] [4] F. Gouyon, G. Widmer, X. Serra, and A. Flexer, Acoustic cues to beat induction: A machine

learning perspective, 2006.