SELF-ADJUSTING BEAT DETECTION AND PREDICTION IN MUSIC

Robert Harper and M.E. Jernigan

Vision and Image Processing Lab, Systems Design Engineering,
University of Waterloo, Waterloo, Ontario, Canada

ABSTRACT

This paper proposes a new approach to beat detection and prediction in music. Recurrent timing networks are used to detect and predict periodicities in an onset stream and are contained within nodes that compete for selection as the best beat hypothesis. Beat prediction nodes perform period self-adjustment to better represent the detected music beat period. The system is tested using a variety of music from different genres and shows promise, in many cases with high correct beat detection percentages.

1. INTRODUCTION

Music is a complex audio signal essentially containing a cacophony of different sounds, yet, incredibly, the human brain is able to process this signal and extract information such as melody, harmony, and rhythm. Most music listeners, regardless of musical training, have an inherent ability to feel the beat of the music and predict future beat locations. While this ability comes naturally to humans, mimicking it with computational devices poses a significant challenge.

A number of algorithms, both real-time and offline, have been developed to detect the beat in a musical audio signal. Previous approaches use a wide variety of techniques including signal energy periodicities [1], rule-based methods [2], and connectionist models [3], with varying degrees of success. In this paper, we introduce a new, causal model for automatic beat detection and prediction.

2. BEAT DETECTION AND PREDICTION

Beat can be defined as a regularly occurring pulse that is delineated by the onset of notes or sounds within the music. It is at the temporal locations of this pulse or beat that listeners are likely to tap their feet. The beat has both period and phase and it is our intention to be able to determine these parameters and use this information to detect future locations of the beat. This process amounts to detecting the strongest periodicity in the progression of sound onsets in the input music signal. For this purpose, we employ the use of recurrent timing networks, first introduced for the detection of musical beat by Cariani in [4].

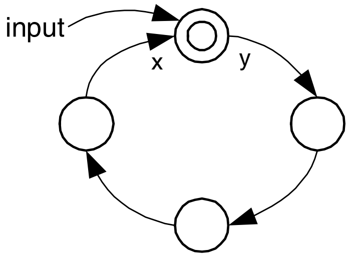

Figure 1: Recurrent timing network, delay of four

Recurrent timing networks are tapped delay loops that allow the input signal to be compared to itself at various instances in the past. Figure 1 shows a recurrent timing net with a delay length of four. When energy in the input signal is coincident with energy that is traveling through the delay loop, reinforcement occurs and the magnitude of the signal component contained within the loop is increased. In this manner, periodic elements in the input signal with period equal to that of the network are reinforced.

Recurrent timing nets are able to detect patterns and periodicities in the input onset signal as well as predict the likelihood of an onset occurring in the input at any time in the future. At each point in time, the first node in the network receives an external input from the stream of sound onsets and an activation level from the timing network indicating the strength of the hypothesis that an input onset will be present. If an onset is present, activation is increased and the hypothesis is strengthened. If an onset is absent, activation is decreased and the hypothesis is weakened. In this way, a prediction of the location of the beat can be made with period equal to the length of the delay in the loop and phase equal to the phase of the largest activation level contained within the loop.

The function used here to grow or decay the activation level in the first node due to the presence or absence of an onset in the input signal differs significantly from that used by Cariani [4]. The function resembles the sigmoid function and is as follows:

$$y = x + I \cdot x \cdot (1 - x)$$

where I is proportional to the strength of the input onset and lies in the interval [-0.5,1], x is the input from network, and y

is the resulting activation level. When no onset exists on the input, *I* assumes a negative value and causes the activation to decay.

3. DESCRIPTION OF THE SYSTEM

The proposed beat detection and prediction system contains three distinct stages. The first stage processes the input audio signal and creates an onset stream in which impulses are produced when a sound onset is detected. The process involved in this stage is not novel to this research and is based directly on the work of Scheirer [1], Klapuri [5], and Duxbury et al [6]. It will not be discussed further here.

The second stage involves the collection of inter-onset interval statistics. The third stage contains competing beat detection nodes representing differing beat period hypotheses. Both stages are discussed below.

3.1. Inter-Onset Interval Statistic Collection

Given that we intend to find a regular pulse occurring within the music onset stream, time intervals that are found to occur commonly between onsets represent valid hypotheses of the period of such a pulse. In order to create beat period hypotheses that can be evaluated using recurrent timing networks of varying length, we must first determine dominant intervals between sound onsets.

Time intervals are calculated between the most recent onset and previous onsets to an arbitrary maximum time delay. These inter-onset intervals are collected in a histogram, which is allowed to decay with the passage of time as seen in [7]. Time intervals corresponding to local maxima in this histogram, calculated with some precision using 2nd-degree polynomial Newtonian interpolation, are selected as potential beat period hypotheses. This information is passed to the final state, the Beat Detection Node Pool.

3.2. Competing Beat Detection Nodes

The third and final stage of the proposed system contains a pool of detection nodes, each representing a beat period hypothesis, that compete for selection as the strongest, most likely, beat prediction. The winning node's beat prediction is used as the beat prediction output for the entire system.

Figure 2 shows the design of a beat detection node. Each node contains a recurrent timing network, used to detect the beat in the incoming onset stream, beat detection logic, a variable rate sampler, used to down-sample the input onset stream, a sample rate controller, used to adjust the down-sample rate, and a node score generator, used to calculate the strength of the current node's beat hypothesis.

The recurrent timing network, as discussed earlier, performs the beat detection and prediction in the incoming note onset stream. The length of the network is calculated upon the creation of the node to contain as many delays as are needed to provide the desired beat period hypothesis. Any required alteration of the node's period hypothesis after creation is achieved through adjusting the down-sampling rate.

The beat detection and prediction logic uses the information within the recurrent timing network to generate a hypothesis of the location of the beat within the network. This amounts to the calculation of the phase of the beat. Very simply, the position in the network with the highest activation energy is selected to represent the timing of the beat. When this activation propagates to the first node, the beat prediction is aligned with the current input sample, and an impulse is sent out as the beat prediction output. If this node is selected as the best beat hypothesis, this beat output impulse stream is used as the beat prediction output for the entire system. A new location within the beat period is selected as the phase of the actual beat if the current guess fails to have the highest activation energy for *N* consecutive periods. We have selected a value of four for *N* in our model.

The incoming onset stream is not fed directly into the recurrent timing network, but is first down-sampled as shown in Figure 2. It is necessary that each node be able to adjust its period hypothesis in an attempt to exactly match the beat period of the input music signal. Since the initial node period is set by the rough period approximation given by the location of a peak in the inter-onset interval histogram, it becomes necessary to fine-tune this estimate. By adjusting the down-sampling frequency of the variable rate sampler on the input to the recurrent timing network, the beat period hypothesis of the current node can be fine-tuned without changing the length of the network.

The controller component that alters the sampling rate does so by comparing the incoming onset stream with the beat prediction output. Onsets found near beat predictions are assumed to correspond to notes that are played on the beat in the input music signal. Deviation is measured between each beat prediction and the closest onset in the input stream and an attempt is made to minimize this deviation for future predictions by altering the input sampling rate. Through this process, both the period and phase of the current beat prediction are synchronized with the period and phase of the beat in the input signal. For example, if the beat predictions are found to be lagging onsets that are most likely on the beat, the sample rate must be decreased to reduce the node period and halt the continuing lag. Moreover, since a phase error has also been introduced in this scenario, the node period must be temporarily decreased to a rate below the actual beat period of the song to realign the beat prediction.

The controller function used to calculate the updated ideal node period after each beat prediction is:

$$p_{ideal}(t_n) = p(t_{n-1}) + G_\phi \cdot e(t_n) + G_p \cdot [e(t_n) - e(t_{n-1})]$$

where $e(t_n)$ is the error (in seconds) between the beat prediction at time t_n and the closest onset, $p_{ideal}(t_n)$ is the ideal

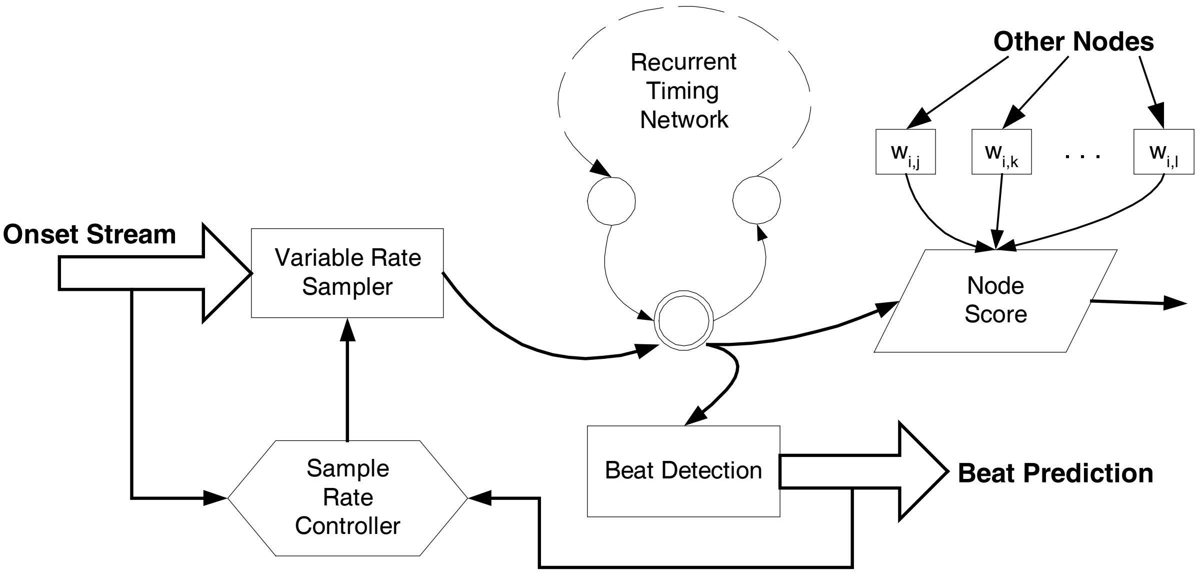

Figure 2: Beat detection node design – internal connections and components

new period, and $p(t_{n-1})$ is the node period used between beat predictions at times t_{n-1} and t_n. G_ϕ is the phase correction gain, and G_p is the period correction gain. In our implementation, we found setting both gains to unity to be ideal.

In the previous formula, the assumption was made that the closest onset to the beat prediction corresponded to a note onset lying on the beat. Unfortunately, this is frequently not the case. It is important to ensure that spurious onsets and onsets not located on the beat affect the determination of the new ideal period as little as possible. To achieve this goal, expectancy curves [8] are centered on each beat prediction and are used to weight the likelihood that an onset in the vicinity of a beat prediction actually corresponds to a note on the beat. We have selected the use of the Gaussian curve for our expectancy curve, such that onsets near the beat are given a large weight (they are expected) and onsets further from the predicted beat are given a small weight. This weight is used to affect the degree to which the current node period is altered to reflect the new "ideal" period.

The final component within each node is the node score calculator. The node score is used to select the best beat detection node for the beat prediction output of the entire system. The score is calculated as a weighted sum of the energies of all of the nodes in the node pool. The node energy is calculated as the RMS of the activation levels within the timing network. This gives an indication of the degree to which strong periodic elements are detected in the timing network. Since beat periods that are harmonics or subharmonics of a given beat period hypothesis represent the same or similar hypotheses, we desire the weights between such nodes to produce appropriate levels of reinforcement. Other period hypothesis ratios that are further from simple integer ratios most likely represent divergent hypotheses and therefore the weights between such nodes should be inhibitory. The weights for node i affected by node j, $w_{i,j}$, are calculated as:

$$w_{i,j} \propto \sqrt{\frac{p_i}{ALCM\left(p_i, p_j\right)}}$$

$$w_{i,i} = C$$

where $ALCM(x,y)$ is the approximate least common multiple of x and y (approximate, meaning common within some small tolerance, ε), p_i is the period of node i, and C is a coupling constant selected to be much larger than the maximum inter-node weight.

4. RESULTS

We have selected four bases upon which to judge the proposed algorithm: percentage of subjectively selected "ground truth" beats correctly predicted by the system, percentage of predicted beats that are incorrect or spurious, the root mean squared error between predicted beats and "ground truth" beats, and the root mean squared error between correctly predicted beats and the closest note onset. Statistics are not collected in the first 8 seconds of the music clip to allow the system to settle.

Six songs are selected for evaluation from a range of music genres. These genres are (1) Rhythm & Blues, (2) Classic Rock, (3) New Rock, (4) Jazz, (5) Classical, and (6) Dance/Techno. Results are as follows:

Song	*Percent Predicted*	*Percent Spurious*	*Beat RMS*	*Onset RMS*
1	100%	27%	31.7 ms	12.8 ms
2	91%	31%	23.8 ms	25.3 ms
3	100%	0%	19.7 ms	14.9 ms
4	46%	36%	23.2 ms	39.8 ms
5	53%	50%	37.2 ms	24.5 ms
6	84%	18%	14.5 ms	9.7 ms

Note that songs 1, 2, 3, and 6 show high correct prediction percentages and reasonably low spurious (incorrect predictions) percentages. In fact, all spurious beat detections in songs 1 and 6 and more than half in song 2 are actually detected off-beats and therefore are essentially valid predictions. The RMS error measurement between the true beat and the predicted beat in song 1 is artificially inflated due to inaccuracies in the determination of the "ground truth" beat. Song 1 has a slight tempo change midway through the clip that the proposed system tracks but the "ground truth" beat measurement does not. This explains the correspondingly low RMS error between the predicted beat and closest onsets. The jazz and classical song beat prediction performance is very poor, showing correct detection percentages near 50% and high spurious beat detections.

5. CONCLUSIONS AND FUTURE DIRECTIONS

A new system for the detection and prediction of the beat within a musical input was introduced in this paper. The proposed system achieves high correct detection percentages in songs from genres that typically have a strong sense of the beat. Selected songs from jazz and classical music genres pose a significant challenge and their performance within the proposed system is poor.

One limitation of the system in its current form is its tendency to jump between harmonics and subharmonics of the true input music tempo. This behavior results in the appearance of large numbers of spurious detected beats that are actually located on the off-beat. A more robust node scoring and competition mechanism or the addition of a slight bias towards favorable tempi (favorable in the sense of human listeners) may help correct this problem.

Another direction that may improve system performance is the addition of stronger cooperation between nodes representing harmonics of the same beat hypothesis. When these nodes experience small adjustments in period, effort could be made to maintain each node's period as an integer multiple of the other related nodes. Decisions regarding the phase of the beat prediction could also be shared between related nodes.

The proposed system for beat detection and prediction shows promising initial results. The causal and predictive capabilities of the system make it an ideal candidate for real-time implementation. Further investigation is warranted into the node competition mechanism and increased node cooperation.

6. REFERENCES

[1] E.D. Scheirer, "Tempo and Beat Analysis of Acoustic Musical Signals," *Journal of the Acoustic Society of America*, 103 (1), pp. 588-601, 1998.

[2] D. Rosenthal, M. Goto, and Y. Muraoka, "Rhythm Tracking Using Multiple Hypotheses," *Proc. of the 1994 Int. Computer Music Conference*, International Computer Music Association, San Francisco, pp. 85-87, 1994.

[3] P. Desain, and H. Honing, "The Quantization of Musical Time: A Connectionist Approach," *Computer Music Journal*, 13 (3), pp. 56-66, 1989.

[4] P. Cariani, "Temporal Codes, Timing Nets, and Music Perception," *Journal of New Music Research*, 30 (2), pp. 1-52, 2001.

[5] A. Klapuri, "Sound Onset Detection by Applying Psychoacoustic Knowledge," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing,* 6, pp. 3089-3092, 1999.

[6] C. Duxbury, M. Sandler, and M.E. Davies, "A Hybrid Approach to Musical Note Onset Detection," *Proc. of the 5th Int. Conf. on Digital Audio Effects*, Hamburg, Germany, pp. 32-38, September 2002.

[7] J. Seppänen, "Tatum Grid Analysis of Musical Signals," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, NY, USA, pp. 131-134, October 2001.

[8] E.W. Large, and C. Palmer, "Perceiving Temporal Regularity in Music," *Cognitive Science,* 26, pp. 1-37, 2002.