# Rhythm Tracking Using Multiple Hypotheses

David Rosenthal[†] *and* Masataka Goto[‡] *and* Yoichi Muraoka[‡]

[†]International Media Research Foundation
2-14-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169, JAPAN.
dfr@media.mit.edu

[‡]School of Science and Engineering, Waseda University
3-4-1 Ohkubo Shinjuku-ku, Tokyo 169, JAPAN.
{goto, muraoka}@muraoka.info.waseda.ac.jp

**Abstract**

We briefly describe two rhythm-tracking systems, called, respectively, *Machine Rhythm* and *BTS*.

Given a MIDI stream as input, Machine Rhythm produces an interpretation that is essentially isomorphic to the rhythmic information represented in normal musical notation. The output of the program defines the placement of measures and assigns rhythmic values (half-note beats, quarter-note beats, etc.) to each note. Although Machine Rhythm is not a real-time system, it processes the MIDI information sequentially, paving the way for possible future real-time implementations. The program attempts to handle some of the more sophisticated rhythm-tracking operations of which humans are capable, such as changes from duple to triple meter or changes in tempo.

BTS tracks beats using raw audio signals as input – in general, a much more difficult problem than tracking it from MIDI data. BTS accomplishes this task by leveraging the fact that for a large corpus of music – rock and pop songs – the beat is indicated with some reliability by the bass and snare drums. BTS's non-reliance on MIDI data enables it to handle a broad range of multimedia applications for which MIDI-based beat-tracking programs cannot be used, and the fact that it works in real time enables its application in a variety of live performance situations.

Both Machine Rhythm and BTS use a similar strategy for managing uncertain or noisy input data – namely, the strategy of pursuing multiple hypotheses.

## 1. Rhythm Tracking Issues

Rhythm tracking and related psychoacoustical and psychological issues have been treated by a number of researchers from a variety of disciplines ([Allen and Dannenberg, 1990], [Bamberger, 1980], [Chung, 1989], [Dannenberg and Mont-Reynaud,1987], [Desain and Honing,1989], [Driesse,1991], [Goto and Muraoka,1994], [Katayose et al.,1989], [Lee,1985], [Longuet-Higgens and Lee,1984], [Rosenthal,1992], [Schloss,1985], [Sloboda,1983], and [Vercoe,1985])[1].

One reason that building a computer rhythm-tracker is difficult is that input data is often noisy or ambiguous. At any given point in the rhythm-tracking process, several interpretations may appear plausible; only further on in the processing does the correct interpretation become clear. One way of managing this situation is to maintain a number of hypotheses, which are periodically ranked and selected.

Another problem is that human rhythm-trackers operate in a much more information-rich environment than do computer rhythm-trackers. Beat-tracking and rhythm-parsing, in humans, are part of an array of auditory information-processing methods which interact in ways that we only partly understand. In a reasonable model of human auditory processes, many other processes act on the incoming auditory data. These include: parsing the music into separate events, estimating the power associated with each event, separating the music into streams, noting repeated patterns, parsing the harmonic structure, recognizing instruments, and so on. We assume that these processes interact and inform each other and the rhythm-tracking processes.

The processes of rhythm-tracking itself is also less unary than is assumed by previous simpler models. It appears that humans normally track several levels of rhythmic activity – that is, in a given situation we may track a beat at the measure level, the half-note level, the quarter-note level, and so on. Again, although we don't really understand the degree of cooperation between these processes, a reasonable model is that they are autonomous to some degree, yet informed enough of each other to maintain coordination.

[1]For the sake of brevity, we will assume that the reader is familiar with the general problems of beat-finding, and concentrate on issues which are specific to Machine Rhythm and BTS.

## 2. Machine Rhythm

Machine Rhythm, developed at the MIT Media Laboratory as part of one of the authors' (Rosenthal) Ph. D. dissertation, addresses some of the issues raised in the last section.

First of all, Machine Rhythm deals with ambiguous input by creating a number of conjectures to cover the range of reasonable explanations. Machine Rhythm's strategy amounts to beam search of a hierarchical space of rhythmic hypotheses (see also [Allen and Dannenberg,1990]).

Machine Rhythm also attempts to duplicate the information-rich environment in which human rhythm-tracking apparently takes place, by emulating some of the auditory processing functions that affect rhythm parsing. In particular, Machine Rhythm segregates the MIDI stream into voices, and searches the resulting separated voices for melodic patterns. Detection of such a pattern constitutes evidence that there is a

beat whose period is a multiple of the length of the pattern. Machine Rhythm also groups nearly simultaneous notes into chords, which can then be viewed as unary events. Machine Rhythm's scheme for chord-construction is based on results from psychoacoustical experiments reported in [Bregman,1990].

## 2.1 Overview of Machine Rhythm

The overall operation of Machine Rhythm may be summarized as follows:

1. The system first preprocesses the entire performance. During the preprocessing stage the MIDI information is grouped into chords (notes with nearly simultaneous onset) and voices (e.g., melody and accompaniment).
2. Machine Rhythm then selects an initial segment of the performance – usually 2-3 seconds – called the startup segment, and makes a number of hypotheses about the rhythm of the initial segment.
3. The system then processes the remainder of the performance sequentially, one note (or chord) at a time. Each hypothesis is extended to account for the new note. If the way in which hypotheses should be extended is ambiguous, Machine Rhythm will produce several hypotheses. As a result, the number of hypotheses grows exponentially in the number of notes processed.
4. When the number of hypotheses exceeds a preset limit, the hypotheses are ranked, and the lower-ranked hypotheses are discarded. Hypotheses are ranked according to the following criteria:
   - Stronger beats (such as the beginnings of measures or half-measures) are more likely to occur on chords rather than single notes.
   - Stronger beats are more likely to occur on notes of longer duration, or notes where the time-interval to the next note is longer.
   - It is preferable that the period of a beat should coincide with the period of a detected melodic pattern.
   - Beats which have uniform or slowly changing periods are preferable to those which do not.

   A manager-module checks for informative interactions among these criteria, and makes some context-sensitive decisions as to how to apply them.
5. Machine Rhythm also incorporates a module which detects changes in rhythmic subdivision, the most common example of which is a triplet.

## 2.2 Test Results

We tested Machine Rhythm on a corpus of 92 performances. Of these the largest block was taken from 55 movements from Mozart piano sonatas performed by Mike Hawley of the MIT Media Lab. An additional data set taken from 37 Mozart sonata movements consisted of performances by one of the authors (Rosenthal). We also tested the system, less formally, on a variety of folksongs, national anthems, etc..

Each test consisted of two parts: we first checked whether the startup module could correctly parse the beginning of the piece. If it was successful, we then checked whether the parser could continue without "losing the beat," that is, given that it had parsed a measure correctly, what were the chances that it would parse the following measure correctly. The results were as follows: The startup module succeeded 62% of the time for the Hawley performances and 65% percent of the time for the Rosenthal performances. Given that it had parsed a measure correctly, the program would parse the next measure correctly 95% of the time for the Hawley performances and 98.5% of the time in the Rosenthal performances. More details on the tests can be found in [Rosenthal,1992].

# 3. BTS (A Real-time Beat Tracking System for Musical Acoustic Signals)

BTS, developed at the Muraoka Lab at Waseda University as part of one of the authors' (Goto) M.S. thesis, also addresses some of the issues raised in the first section of this paper, though the approach differs from that of Machine Rhythm.

BTS processes a monaural acoustic signal of music and recognizes temporal positions of beats in real time. Most previous rhythm-trackers were not able to process acoustic signals that contain sounds of various instruments, especially drums. They were able to process only MIDI signals or acoustic signals played on a few instruments in non-real time. BTS deals with commercially distributed popular music such as rock and pop music in which mainly drums maintain the beat.

## 3.1 Specifications of BTS

BTS works on assumptions that fit a large class of popular music. The tempo of an input song is constrained to be between 70 M.M. and 180 M.M. and almost constant; popular songs have less tempo variation than do classical works. The time signature is assumed to be 4/4, this being the most frequent time-signature in the repertoire we are considering.

BTS reports *beat information* (*BI*) that consists of: the temporal position of a beat (*beat time*), its location in a half-measure (*beat type*), and the current tempo. BI corresponding to a quarter note is broadcast to the Ethernet as an RMCP[2] packet synchronized to the music. This enables other computers on the Ethernet to

[2]RMCP stands for remote music control protocol, which is a communication protocol between servers and clients in the RMCP system [Goto and Hashimoto,1993].

use the BI in various ways. For example, a workstation connected to a MIDI instrument may create drum sounds or clapping sounds in time to the input music. A workstation with a graphics engine may also create computer graphics synchronized with music.

Beat type indicates whether a beat is a strong beat or a weak beat – i.e., BTS can track beats at the half-note level. To infer beat type, BTS assumes that a bass drum (*BD*) mainly sounds on a strong beat (the first or third quarter notes in a measure) and a snare drum (*SD*) on a weak beat (the second or fourth). This does not mean that all BD and SD must sound on the strong and weak beats, respectively, but rather that that arrangement should be the most frequent.

## 3.2 Main Issues and Solutions

The principle beat-tracking issues addressed by BTS are as follows:

1. *It is generally impossible to obtain precise onset times from acoustic signals that contain sounds of various instruments.*
   BTS employs sophisticated means of estimating the onset time in the frequency analysis stage. First, BTS finds multiple interpretations corresponding to various time-window widths, one of which is confirmed by subsequent processing. Second, the reliability of an onset time is calculated by a process which takes into account such factors as the rapidity of increase in power, and the power present in nearby time-frequency regions. The higher the reliability of an onset time, the greater its importance in subsequent processing.
2. *BTS should be able to recover the correct tracking even if the current hypothesis becomes incorrect.*
   BTS manages multiple agents that track beats according to different strategies and then examines multiple hypotheses in parallel. Even if some agents lose track of beats, BTS will track correct beats as long as other agents have the correct hypothesis. Each agent interprets onset time and makes his own hypothesis, which consists of next beat time predicted, its beat type, its reliability, and current inter-beat-interval. BTS generates BI on the basis of the most reliable and stable hypothesis.
3. *BTS must acquire the characteristic frequencies of BD and SD dynamically.*
   BTS, like human listeners, utilizes BD and SD as principle clues to the location of strong and weak beats. Because the sounds of BD and SD are not known in advance, BTS automatically acquires the characteristic frequencies of these sounds during the beat-tracking process. Note that BTS cannot simply use the detected BD and SD to track the beats, because this detection process is too noisy. The detected BD and SD are only used to determine the beat type (strong or weak) of an already detected beat.

## 3.3 Overview of BTS

Figure 1 shows the overview of BTS implemented on a distributed memory parallel computer, the Fujitsu AP1000 which consists of 64 processing elements called cells. The number of cells assigned to each process is indicated at the bottom right of rectangles.

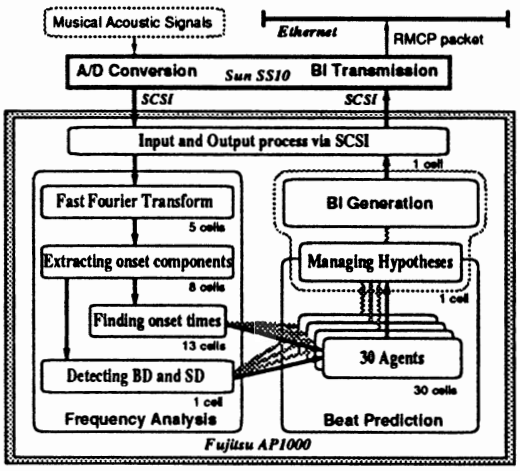


*Figure 1: Overview of BTS*

First, *Frequency Analysis* finds notes' onset times in an input acoustic signal digitized by *A/D Conversion* and also detects BD and SD. Second, multiple agents in *Beat Prediction* interpret the onset times found previously and make parallel hypotheses: each agent first calculates the inter-beat-interval; it then predicts the next beat time, and infers its beat type, and finally evaluates its own reliability. *BI Generation* assembles BI on the basis of the most reliable hypothesis. Finally, *BI Transmission* transmits the BI to other application programs via the Ethernet.

## 3.4 Test Results

We tested BTS for 30 popular songs in the rock and pop music genre. These songs were sampled from commercial compact discs and satisfied the assumptions stated above. Their tempi ranged from 78 M.M. to 167 M.M.

BTS correctly tracked beats in 27 songs out of 30 songs in real time. After the BD and SD had sounded stably for a few measures, the beat type was obtained correctly. The three failures occurred as follows: In two songs, the beat type was reversed as if BD were SD, because BTS could not acquire their characteristic frequencies correctly. In the other song, BTS tracked beats correctly, for the most part, but during about three measures in the middle, the beat type was reversed due to some irregular rhythm in the drums.