DOWNBEAT PREDICTION BY LISTENING AND LEARNING

Tristan Jehan

Media Laboratory
Massachusetts Institute of Technology
20 Ames Street, Cambridge, MA 02139
tristan@media.mit.edu

ABSTRACT

The perception of downbeat is not yet very well understood, although it generally "feels" natural and intuitive to music listeners. We propose an *unbiased* and *predictive* modeling strategy of downbeat by combining psychoacoustic models of music *listening* with time-lag embedded segment *learning*. The model is causal, tempo independent, and could inform in a top-down manner a bottom-up beat tracker. Results with particularly complex music examples are presented.

1. INTRODUCTION

One of the most fundamental and recognizable properties of music is its *metrical structure*. Somewhat hierarchically organized, rhythms, harmonies, melodies and timbres generally fall into place within a common metric that arranges and schedules sounds in time. A significant evidence of metrical structure is the *beat*, a regularly spaced perceptual pulse, often expressed by musicians and listeners through the act of foot tapping along to the music. The rate of at least two repeating beats gives the *tempo*, and is defined as a number of beats per minute (BPM).

Many previous works have dealt with the specific task of tracking beat and tempo in audio signals [1]. Most generic approaches do not assume any prior knowledge about the music, and process audio signals in a *bottom-up* fashion, sometimes implementing listening models. They, however, all tend to fail with syncopated, polyrhythmic, or generally more difficult rhythms from a western music standpoint. Indeed, it may be naive to expect that systems that only use signal processing techniques could perform well with certain types of Indian, Latin, or African music, without any kind of assumption or prior exposure to these rhythms.

This is generally the concern of *rule-based* approaches that impose *top-down* processes as a way to model a form of persistent mental framework, which aims at guiding the perception of new incoming material [2]. These models can generally be criticized for being specific and biased, as well as for needing a symbolic and quantized representation of the music, therefore being effective only as a second stage to a beat tracking system.

As listeners intuitively group musical events contained in the acoustic stream, they start constructing a perceptually-grounded internal representation of *patterns*. If finding the beat may be a difficult task, then finding the *downbeat* is even more difficult. We define downbeat as the *first* beat in a musical pattern. Although generally equivalent to the first beat in a *measure* as defined by the western music notation (as the conductor's arm moves downward), we are more concerned with finding a perceptually valid representation of downbeat: a common and recurrent reference point for musicians, dancers, and possibly listeners to follow.

Together with the notions of beat and downbeat comes the notion of *time signature*, a description of *meter*: the number of beats per measure, or in the general case, per pattern. There have been only few attempts at analyzing the meter in music audio, generally based on autocorrelation or cross-correlation of successive pattern candidates [3]. Even fewer considered the question of *downbeat detection*. Goto and Muraoka used a complex agent-based system for analyzing chord changes simultaneously with beat tracking. They estimated downbeat by assuming that chord changes most likely coincide with it [4], a generally fair assumption for western pop music.

More recently, Klapuri et al. proposed a full analysis of musical meter by looking simultaneously into three different time scales, including the so-called atomic *tatum* level, the *tactus* (beat) level, and the measure level. A probabilistic model jointly estimates the length and phase of each level, by taking into account the temporal dependencies between successive estimates. It was found that determining the phase of the measure pulse was difficult and required some kind of rhythmic pattern matching technique. Two binary reference patterns (of two and four beats) were then constructed through trial and error until they best generalized the database [5]. A similar "template" strategy was also used by Ellis and Arroyo in a drum-pattern classification task [6].

It is not surprising that downbeat turns out to be the most difficult metrical measure since there is no well-defined heuristics regarding how humans actually manage to detect it. In this paper we claim that downbeat estimation requires some fair amount of prior knowledge, which may be acquired either consciously through active *learning*, or unconsciously by experience with *listening*: a cultural bias. We demonstrate that not only *training* is necessary in the algorithm, but the technique can also improve the robustness of bottom-up approaches to beat induction by providing the beat tracker with top-down feedback control.

2. EVENT-SYNCHRONOUS TIMBRE LISTENING

Our first task consists of building a perceptually meaningful and *low-rate* audio surface description, or *fingerprint*. Our strategy includes segmenting the musical signal into small units of sounds. We define a sound as perceptually meaningful if it is timbrally consistent, i.e., it does not contain any noticeable abrupt changes. We base our segmentation on an *auditory model*. Its goal is to remove the information that is the least critical to our hearing sensation, while retaining the most important parts.

2.1. Auditory spectrogram

We first apply a running STFT, and warp the spectra into a Bark scale. We model the non-linear frequency response of the outer and middle ear, and apply both frequency and temporal masking [7], turning the outcome into a "what-you-see-is-what-you-hear" type of spectrogram (top pane of Figure 1). We then lower the frequency resolution to 25 critical bands, preserving essential characteristics of timbre while reducing the complexity.

2.2. Segmentation

We convert the auditory spectrogram into an *event detection function* by calculating the first-order difference function for each spectral band, and by summing across channels. Transients are localized by *peaks*, which we smooth by convolving the function with a 150-ms Hanning window to combine perceptually fused peaks together, i.e., two events separated in time by less than about 50 ms. Desired onsets can be found by extracting the local maxima in the resulting function and by searching the closest local minima in the loudness curve, i.e., the quietest and ideal time to cut (middle pane of Figure 1).

We finally reduce dimensionality in the time domain by averaging the spectral content on a segment basis, synchronizing 25-coefficient feature vectors to musical events (bottom pane of Figure 1). Another good option would consist of first calculating the musical *tatum* as described in [8], and then averaging spectral energies at the tatum rate.

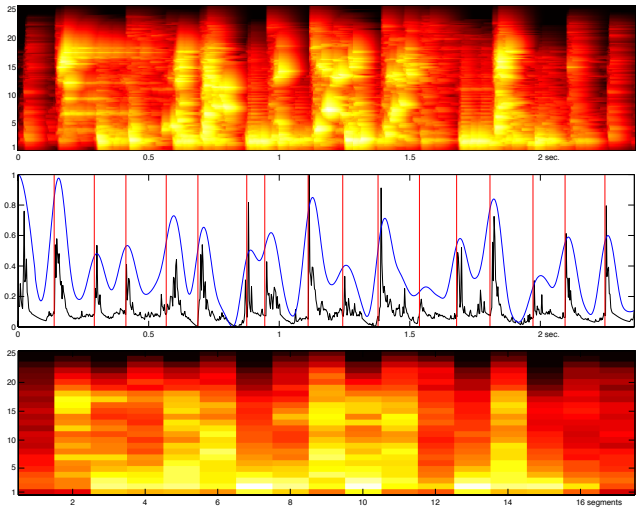

Figure 1: *Short audio excerpt of James Brown's "sex machine."* Top: *auditory spectrogram.* Middle: *segmentation, including raw detection function (black), smooth detection function (blue), and onset markers (red).* Bottom: *event-synchronous timbral feature vectors.*

2.3. Beat tracking

A beat tracker based on the front-end auditory spectrogram and Scheirer's bank-of-resonators approach [1] is developed. It assumes no prior knowledge, and outputs a *confidence* value. Yet, one of the main drawbacks of the model is its unreliable tempo-peak selection mechanism. As depicted in Figure 2, a few peaks could give a plausible answer, and choosing the highest is not necessarily the best, or most stable strategy. Because it does not rely on tempo, our downbeat detector can help improve that reliability.

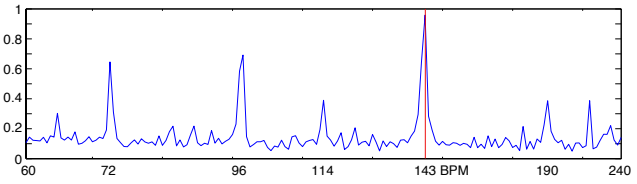

Figure 2: *Tempo spectrum of a fusion piece at roughly 143 BPM. A peak at the sub octave 72 BPM is visible, as well as other harmonics of the beat.*

Meanwhile, we use the current beat tracker in a supervised fashion—with manual selection of tempo and phase—as a way to accurately and consistently label our *training* segments with a metrical time constant, therefore defining their relative location in the measure. If the tempo does not vary too quickly, then beat markers get usually extracted accurately throughout the whole song. The constant is a float value $p \in [0, M[$, where M is the number of beats in the measure, and where 0 represents the downbeat and is typically selected by hand. In this article, we have constrained our problem to constant 4/4 time signatures, i.e., $M = 4$. The resulting phase signal looks like a sawtooth. Taking the absolute value of the derivative of this phase signal returns a *ground-truth* downbeat prediction signal much like is displayed in the left pane of Figure 4.

3. LEARNING A METRICAL STRUCTURE

The previous section has represented the musical surface as a compact, yet perceptually meaningful causal fingerprint. We also constructed a simple signal that carries the information of periodic downbeat. In this section, we are interested in modeling the *correlation* between the musical surface and that signal.

3.1. State-space reconstruction

Linear systems have two particularly desirable features: they are well characterized and are straightforward to model. One of the most standard ways of fitting a linear model to a given time series consists of minimizing squared errors given a set of coefficients in an ARMA (autoregressive moving average) model. This idea of forecasting future values by using immediately preceding ones (a tapped delay line) was first proposed by Yule. It turns out that the underlying dynamics of non-linear systems can be *understood* as well, and their behavior inferred from observing delay vectors in a time series, where no or *a priori* information is available about its origin [9]. This general principle is known as *state-space reconstruction* [10], and inspires our method. We look into the trajectory of the time lagged space to infer an underlying metrical behavior, and predict a correlated one-dimensional signal, which carries the information of downbeat. Projecting a high-dimensional space onto a single dimension, and *generalizing* the space reduction, given a limited training data set is a task generally referred to as *learning*.

3.2. Support-Vector Machine

Support Vector Machines or SVM [11] rely on preprocessing data as a way to represent patterns in a *high* dimension—typically much higher than the original feature space. With appropriate non-linear mapping into the new space (through a *basis function*, or *kernel*, such as gaussian, polynomial, sigmoid functions) data can always be regressed (and classified) with a *hyperplane*. Support vector machines differ radically from comparable approaches as SVM training always converges to a *global* minimum, i.e., the search corresponds to a *convex* quadratic programming problem. While obviously not the only *machine learning* solution, this is the one we chose for the present experiments. Other solutions may include for instance mixture of gaussian models or artificial neural networks.

3.3. Downbeat training

Training is a semi-automatic task that requires only little human supervision. The listening stage, including auditory spectrogram and segmentation is entirely unsupervised. So is the construction of the time-lagged feature vector. It is built by simply appending a certain number of preceding 25-dimensional vectors. Best results were found by using 6 to 12 past segments, corresponding to nearly the length of a measure. We model short-term memory *fading* by linearly scaling down older segments, therefore increasing the weight of most recent segments.

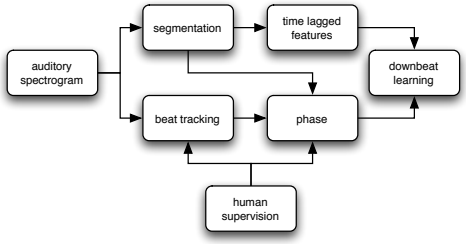

Figure 3: *Supervised learning schematic.*

Training a support vector machine at predicting the downbeat corresponds to a regression task of several dozens of feature dimensions (e.g., $9\,past\ segments \times 26\,features\,per\ segments = 234\,features$) into one single dimension (the corresponding phase of the following segment). We expect that the resulting model is not only able to predict the downbeat of our training data set, but generalizes well enough (as opposed to memorizing) to predict the downbeat of new input data—denominated *test* data in the experiments of the next section. An overall schematic of the training method is depicted in Figure 3.

4. EXPERIMENTS

Although downbeat may often be interpreted through harmonic shift [4], or a generic "template" pattern [5], sometimes none of these assumptions apply. This is the case of the following example.

4.1. The James Brown case

James Brown's music is often characterized by its repetitive single chord and syncopated rhythmic pattern. The typical assumptions mentioned earlier would not hold. There may not even be any energy in the signal at the perceived downbeat. Figure 4 shows the results of a simple learning test with a 30-second excerpt of "I got the feelin'." After listening and training with only 30 seconds of music, the model demonstrates good signs of learning (left pane), and is already able to predict reasonably well certain downbeats in the next 30-second excerpt of the same piece (right pane).

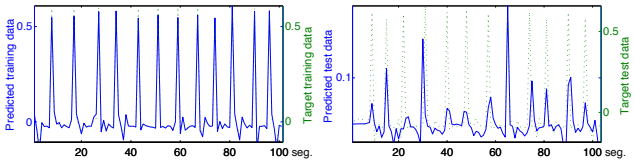

Figure 4: *Downbeat prediction for James Brown's "I got the feelin'." Left: training data set, including a dozen measures. Right: test data set: the dotted green line represents the ground truth; the blue line our prediction. Note that there can be a variable number of segments per measure.*

Note that for these experiments: 1) no periodicity is assumed; 2) the system does not require a beat tracker and is actually tempo independent; 3) the predictor is causal and does, in fact, predict one segment into the future, i.e., about 100-300 ms. The prediction schematic is given in Figure 5.

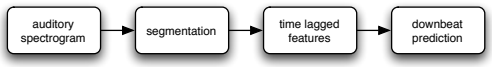

Figure 5: *Causal downbeat prediction schematic.*

4.2. Inter-song generalization

Our second experiment deals with an even more complex rhythm from the North East of Brazil called *Maracatu*. One of its most basic patterns is shown in standard notation in Figure 6. The bass-drum sounds are circled by dotted lines. Note that two out of three are syncopated. A listening test was given to several musically trained western subjects, none of whom could find the downbeat. Our beat tracker also performed very badly, and tended to lock onto syncopated accents.

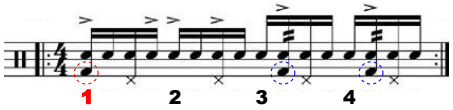

Figure 6: *Typical Maracatu rhythm score notation.*

We trained our model with 6 Maracatu pieces from the band Maracatu Estrela Brilhante. Those pieces include a large number of drum sounds, singing voices, choirs, lyrics, and a variety of complex rhythmic patterns at different tempi. Best results were found using an embedded 9-segment feature vector, and a gaussian

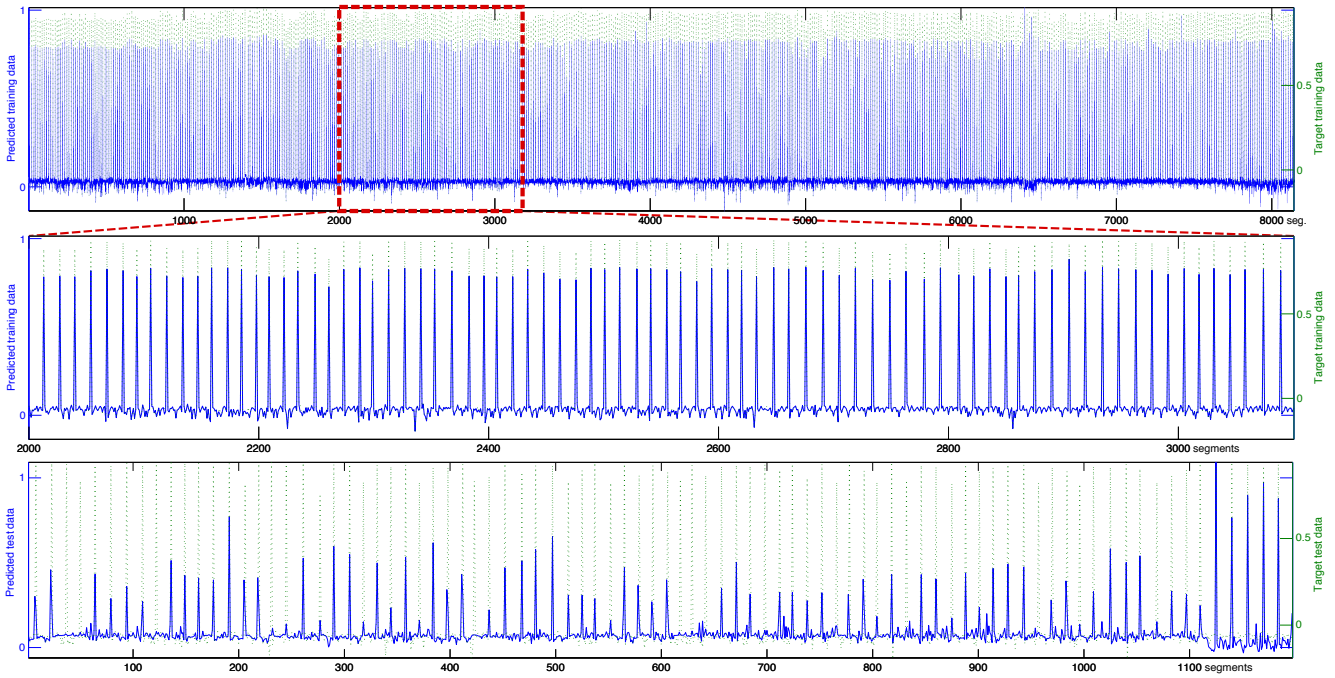

Figure 7: *Downbeat prediction results for the Maracatu expert model.* Top: *full training data set (6 songs) including training downbeat (dotted green) and predicted downbeat (blue). Note that learning is consistent across all data.* Middle: *zoom (red section) in the training data to show phase prediction accuracy.* Bottom: *results for the new test data (1 song). Note that the last few measures are particularly good: most songs tend to end in the same way.*

kernel for the SVM. We verified our Maracatu "expert" model both with the training data set (8100 data points), and with a new piece from the same album (1200 data points). The model performs very well on the training data, and does well on the new untrained data (see Figure 7). Total computation cost (including listening and modeling) was found to be somewhat significant at training stage (about 15 minutes on a dual-2.5 GHz Mac G5 for the equivalent of 20 minutes of music), but minimal at prediction stage (about 15 seconds for a 4-minute song).

5. CONCLUSION

We have demonstrated the workability of an unbiased downbeat predictor based on surface listening, and time-lag embedded learning. The model is causal, and tempo independent: it does not require beat tracking. Instead, it could very well be used as a phase-lock mechanism for bottom-up beat detectors, which typically run in an open loop fashion.

6. REFERENCES

[1] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *Journal of the Acoustic Society of America*, vol. 103, no. 1, January 1998.

[2] P. Desain and H. Honing, "Computational models of beat induction: the rule-based approach," *Journal of New Music Research*, vol. 28, no. 1, pp. 29–42, 1999.

[3] C. Uhle and J. Herre, "Estimation of tempo, micro time and time signature from percussive music," in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, UK, September 2003.

[4] M. Goto and Y. Muraoka, "Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions," *Journal of Speech Communication*, vol. 27, pp. 311–335, 1999.

[5] A. P. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transaction on Speech and Audio Processing (in Press)*, 2005.

[6] D. P. Ellis and J. Arroyo, "Eigenrhythms: Drum pattern basis sets for classification and generation," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.

[7] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, 2nd ed. Berlin: Springer Verlag, 1999.

[8] J. Seppänen, "Tatum grid analysis of musical signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, October 2001.

[9] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*, ser. Lecture Notes in Mathematics, D. Rand and L. Young, Eds., vol. 898. New York: Springer-Verlag, 1981, pp. 366–381.

[10] N. A. Gershenfeld and A. S. Weigend, "The future of time series: Learning and understanding," in *Time Series Prediction: Forecasting the Future and Understanding the Past*, A. S. Weigend and N. A. Gershenfeld, Eds. Reading, MA: Addison–Wesley, 1993, pp. 1–63.

[11] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data [in Russian]*. Moscow: Nauka, 1979, (English translation: Springer-Verlag, New York, 1982).