Performance Measures for Tempo Analysis

William A. Sethares and Robin D. Morris

William A Sethares Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706-1691 USA. 608-262-5669 sethares@ece.wisc.edu

Robin Morris RIACS, NASA Ames Research Center, MS 269-2, Moffett Field, CA 94035-1000 rdm@email.arc.nasa.gov

Abstract

This paper builds on a method of tracking the beat in musical performances that preprocesses raw audio into a collection of low level features called “rhythm tracks” and then combines the information using a Bayesian decision framework to choose a set of parameters that represent the beat. An early version of the method used a set of four rhythm tracks that were based on easily understood measures (energy, phase discontinuities, spectral center and dispersion) that have a clear intuitive relevance to the onset and offset of beats. For certain kinds of music, especially those with a steady pulse such as popular, jazz, and dance styles, these four were often adequate to successfully track the beat. Detailed examination of pieces for which the beat tracking was unsuccessful suggested that the rhythm tracks were failing to provide a clear indicator of the beat times. This paper presents new sets of features (new methods of generating rhythm tracks) and a way of quantifying their relevance to a particular corpus of musical pieces. These measures involve the time waveform, spectral characteristics, the cepstrum, and sub-band decompositions, as well as tests using (empirical) probability density and distribution functions. Each measure is tested with a variety of “distance” metrics. Results are presented in summary form for all the proposed rhythm tracks, and the best are highlighted and discussed in terms of their “meaning” in the beat tracking problem.

EDICS: 2-MUSI Permission to publish this abstract separately is given.

I. INTRODUCTION

A common human response to music is to "tap the foot" to the beat, to sway to the pulse, to wave the hands in time with the music. Underlying such common motions is an act of cognition that is not easily reproduced in a computer program or automated by machine. The beat tracking problem is important as a step in understanding how people process temporal information and has applications in the editing of audio/video data, in synchronization of visuals with audio, in audio information retrieval [23] and in audio segmentation [25]. Several methods of finding the pulse directly from the audio have been proposed in the literature, including [5], [21], [22] and [16] and our own work in [11] (recently expanded in [19]). An overview of the beat tracking problem and a taxonomy of beat tracking methods can be found in [17], and a review of computational approaches for the modeling of rhythm is given in [7].

Despite this attention (and a significant amount of related work in the cognitive sciences ([8],[10]) and music theory ([3], [9])) there is no agreed upon formal definition of "beat." This leaves the field in the curious position of attempting to track something (the beat) without knowing exactly what it is that one wishes to track. The next section defines the *auditory boundary* which is used to define *beat* in a way that is useful in constructing algorithms for the detection and tracking of musical pulse and tempo. This may be viewed as a preliminary step towards an automated identification of musical rhythm. Section III describes a variety of possible measures that may be used to identify auditory boundaries. Section IV describes the underlying probabilistic rhythm track model and provides an experimental method for determining the quality of a candidate rhythm track. Experimental results comparing a large number of different rhythm tracks on a large variety of music are presented in section V, and section VI clusters the results to find that there are only six distinct features among the most successful rhythm tracks. Those rhythm tracks which perform best are then discussed in greater detail in section VII in an attempt to correlate the mathematical measures with corresponding aspects of perception.

II. WHAT IS A "BEAT"?

The verb "to beat" means "to strike repeatedly," while "the beat" of a musical piece is "a steady succession of rhythmic units." Such dictionary definitions are far from technical definitions that might be useful in the design of algorithms for the processing of audio data. To concretely define the "beat," observe that humans are capable of perceiving a wide range of auditory phenomenon, and of distinguishing many kinds of auditory events. The definition hinges on the idea of an auditory boundary, which effectively pinpoints times at which changes are perceived in the audio stream.

Definition 1: An *auditory boundary* occurs at a time t when the sound stimulus in the interval $[t-\epsilon, t]$ is perceptibly different from the sound stimulus in the interval $[t, t+\epsilon]$.

Auditory boundaries are quite general, and they may occur on different time scales depending on the size of ϵ. For example, long scale auditory boundaries (with ϵ on the order of tens of seconds) occur when a piece of music on the radio is interrupted by an announcer, when a car engine starts, and when a carillon begins playing. Short scale auditory boundaries (with ϵ on the order of tenths of a second) occur when instruments change notes, when a speaker changes from one syllable to the next in connected speech, and each time a hammer strikes a nail. At yet smaller values of ϵ (on the order of milliseconds) the "grains of sound" [14] merge into a perception of continuity so that boundaries do not occur.

Perhaps the most common example of short time scale auditory boundaries involve changes in amplitude (or power) such as occur when striking a drum. Before the strike, there is silence. At the time of the strike (and for a short period afterwards) the amplitude rises sharply, causing a qualitative change in the perception (from silence to sound). Shortly afterwards, the sound decays, and a second, weaker boundary is perceived (from sound into silence). Of course, other aspects of sound may also cause boundaries. For example, pitch (or frequency) changes are readily perceptible. An auditory boundary might occur when a sound changes pitch, as for example when a violin moves from an A note to an $A\sharp$. Before the boundary, the perception is dominated by the violin at fundamental $A = 440$ Hz while after the boundary the perception is dominated by the sound of the violin playing the $A\sharp = 466$ Hz. On the other hand, boundaries do not necessarily occur at all pitch changes. Consider the example of a pitch glide (say an oscillator sweeping from 100 Hz to 1000 Hz over a span of thirty seconds). While the pitch changes continuously, the primary perception is of the "glide" and so boundaries are not perceived except for the longer scale boundaries at the start and stop of the glide.

These examples highlight several important aspects of auditory boundaries. First, the boundaries may be of different strengths. Second, they may be caused by different aspects of perception. Most importantly, the key phrase "perceptibly different" is not always transparent since exactly *which* aspect of perception dominates at any given time is a complex issue that depends on the training of the listener, on the focus of the listener's attention, as well as on a myriad of physical factors.

Isolated auditory boundaries are typically perceived as starting or stopping times of auditory events. When auditory boundaries occur periodically or in a regular succession, then a new phenomenon emerges: the beat.

Definition 2: A *beat* is a regular succession of auditory boundaries.

For example, suppose that a series of audio boundaries occur at times

$$\tau,\ \tau+T,\ \tau+2T,\ \tau+3T,\ \tau+4T,\ldots \tag{1}$$

where T is the time between adjacent audio boundaries and τ specifies the starting time of the series. Of course, actual sounds will not be so precise: T might change (slightly) between repetitions, some of the $\tau+nT$ terms might be missing, and there may be some extra auditory boundaries interspersed among the T-width lattice. Thus, while (1) is idealized to a periodic sequence, the term *regular succession* is used to emphasize that a sequence of the auditory boundaries need not be strictly periodic in order to evoke the beat sensation. The beat tracking problem requires finding the best regular lattice of points (for instance, the τ and T above) that fit the sequence of auditory boundaries dictated by the performance of a particular piece of music.

III. Rhythm Tracks

Just as the definition of beat has two components (the auditory boundaries and their regular spacing), our algorithm for beat tracking finds auditory boundaries and then parses the sequence of boundaries for regularities. This section details the creation of *rhythm tracks*, which process and downsample a digitized audio stream so that when auditory boundaries occur, the rhythm track takes large values. Between auditory boundaries, when there is little change in the sound stimulus, the rhythm track is takes small values. Such rhythm tracks can be combined and searched for regularities in many ways. For example, in [19] we demonstrated a particle filter based method that operates directly on a collection of rhythm tracks.

Auditory boundaries are not measurable instantaneously, as indicated by the presence of the ϵ parameter in the definition. Rhythm tracks are created by partitioning a sound waveform $w(t)$ into N ϵ-sized segments and assigning a measure (or number) to each partition. To be concrete, let $w(t)$ be sampled at intervals T_s and let $w(jT_s)$ designate the jth sample. For notational simplicity, let $\epsilon = nT_s$ be an integer multiple of the sampling interval. Then the kth partition contains the n samples from $w(knT_s - nT_s + 1)$ to $w(knT_s)$, that is,

$$s(k) \equiv [w(knT_s - nT_s + 1), \ldots, w(knT_s)],\ k = 1, 2, \ldots, N. \tag{2}$$

Rhythm tracks can be built from the sequence of partitions $s(k)$ in a variety of ways. Perhaps the simplest approach uses a direct function of the time elements in $s(k)$.

Example 1: The energy in the kth partition is

$$e(k) = \sum_{i=1}^{n} w^2(knT_s - nT_s + iT_s). \tag{3}$$

The values of $e(k)$ can be used directly as the terms of a rhythm track. Observe that this rhythm track is downsampled by a factor of n from the raw audio data.

Example 2: Motivated by the idea that changes are significant, the energy differences $e(k+1) - e(k)$ can be used to define a rhythm track.

Example 3: Another alternative considers the percent change in the energy $\frac{e(k+1)-e(k)}{e(k+1)+e(k)}$. This rhythm track is insensitive to the overall volume (power) of the sound, potentially allowing identification of auditory boundaries in both soft and loud passages.

Other functions of the partition $s(k)$ are also reasonable.

Example 4: The function $\sum x^2 \log(x^2)$ gives a measure of the entropy in the sequence x, and this can be used to create rhythm tracks by assigning the number

$$ent(k) = \sum_{i=1}^{n} w^2(knT_s - nT_s + iT_s) \log(w^2(knT_s - nT_s + iT_s))$$

to the kth partition. The values of $ent(k)$ may be used directly as the terms of a rhythm track, the change in entropy $ent(k+1) - ent(k)$ may be used, or the percent change $\frac{ent(k+1)-ent(k)}{|ent(k+1)|+|ent(k)|}$ may be appropriate.

Example 5: Alternatively, the $\log$ of the energy can be used to create a rhythm track

$$logen(k) = \sum_{i=1}^{n} \log(w^2(knT_s - nT_s + iT_s)).$$

As in Example 4, the values of $logen(k)$ may be used directly as the terms of a rhythm track, the change $logen(k+1) - logen(k)$ may be used, or the percent change $\frac{logen(k+1)-logen(k)}{|logen(k+1)|+|logen(k)|}$ may be appropriate.

Example 6: Another kind of measure of the signal $w(kT_s)$ is the *total variation* of the kth partition

$$tv(k) = \sum_{i=1}^{n-1} |w(knT_s - nT_s + (i+1)T_s) - w(knT_s - nT_s + iT_s)|.$$

This measures the "wiggliness" of the waveform, and is larger when the waveform has significant energy in the higher frequencies. Again, the values $tv(k)$, their differences, or their percent change may be useful in generating rhythm tracks.

The central dilemma of this paper should now be clear: there are many ways to measure auditory boundaries. Which of the methods in examples 1-3 is the "best" in terms of clearly delineating energy-related auditory boundaries? Which of the methods in examples 4-5 is the "best" in terms of clearly

delineating entropy-related auditory boundaries? How do these methods compare to the total variation? A pragmatic way to answer this question is the focus of section IV. Indeed, we are not the first to confront these issues. For instance, [6] uses three rhythm tracks: the energy, the spectral flatness, and the energy within a single subband.

There are other ways to use the data in the partition $s(k)$. For instance, the Fourier transform gives the magnitude and phase spectra, and these can be also used to define rhythm tracks. Let $W_k(mf_s)$ for $m = 1, 2, \ldots, n$ represent the FFT of the kth partition $s(k)$ where $f_s = \frac{1}{T_s}$. Then any of the functions in examples 1 to 5 can be applied to either the magnitude spectrum $|W_k(mf_s)|$ or to the phase spectrum $\angle W_k(mf_s)$. Moreover, it is common to do some kind of (nonrectangular) windowing and overlapping of the data when measuring the spectrum [12], and the details of the window and the amount of overlap provide another set of parameters that may be varied. There are also other ways of using the spectra. Perhaps the most obvious is to apply a masking operation to the magnitude spectrum and to build a rhythm track from just (say) the low frequencies, or just the high frequencies. Indeed, this is the strategy followed in [21] in the construction of the *audio matrix* and in [16], though the latter uses a bank of filters rather than an FFT to accomplish the decomposition. There are other, less obvious ways of utilizing the spectra.

Example 7: The *center* of the spectrum is the frequency value f^c at which half of the energy lies below f^c and half lies above. Let $f^c(k)$ be the spectral center of partition k. Then $f^c(k)$, the difference $f^c(k+1) - f^c(k)$, and the percent change can all be used as rhythm tracks.

Example 8: The *spectral dispersion* is the dispersion of the magnitude spectrum about the center. Let $f^d(k)$ be the spectral dispersion of partition k. Then $f^d(k)$, the difference $f^d(k+1) - f^d(k)$, and the percent change can all be used as rhythm tracks.

Example 9: When unwrapped, the phase spectrum often lies (approximately) along a line. Let $\tau^p(k)$ be the slope of the closest line to the phase spectrum in the k partition. Then $\tau^p(k)$, the difference $\tau^p(k+1) - \tau^p(k)$, and the percent change can all be used as rhythm tracks.

The methods of examples 7-9 are discussed at length in [19].

Once the partitions are chosen, the data may be used directly in the time domain or it may be translated into the frequency domain (using either the magnitude or phase spectra) as suggested above. There are other ways of preprocessing or transforming the data. For example, the cepstrum may be used instead of the spectrum. If the data in each partition is presumed to be generated by some kind of stochastic process, then a histogram of the data gives the empirical probability density function (PDF) and this can

be transformed into the empirical cumulative distribution function (CDF). Both the PDF and the CDF may then be used in any of the ways suggested in examples 1-5 to create rhythm tracks.

Recasting the rhythm tracks in the language of probability suggests the use of a variety of statistical tests which may also be of use in distinguishing adjacent partitions. For example, the Kolmogorov-Schmirnov test finds the maximum distance between two CDFs and describes the probability that they arise from the same underlying distribution. Stein's Unbiased Risk Estimator (SURE) can also be applied using (for instance) a log energy measure. Of course, one can readily combine the various methods, taking (say) the CDF of the magnitude of the FFT, or the PDF of the cepstral data. Table I details 17 different ways of preprocessing or transforming the data in the partitions, including six subband (wavelet) transforms that carry out (approximately) the same subband decomposition as in [16].

TABLE I

THE PARTITIONED DATA CAN BE USED IN ANY OF THESE SEVENTEEN DIFFERENT DOMAINS.

label	domain
A	time signal
B	magnitude of FFT
C	phase of FFT
D	cepstrum
E	PDF of time signal
F	CDF of time signal
G	FFT of the PDF of time
H	PDF of FFT magnitude
I	CDF of FFT magnitude
J	PDF of cepstrum
K	CDF of cepstrum
L	subband 1: 40 - 200 Hz
M	subband 2: 200 - 400 Hz
N	subband 3: 400 - 800 Hz
O	subband 4: 800 - 1600 Hz
P	subband 5: 1600 - 3200 Hz
Q	subband 6: 3200 - 22000 Hz

Considering all these factors, the total number of rhythm tracks is approximately the product of

$$\left\{\begin{array}{c}\text{\# ways to}\\ \text{choose partitions}\end{array}\right\} \times \left\{\begin{array}{c}\text{\# of}\\ \text{domains}\end{array}\right\} \times \left\{\begin{array}{c}\text{\# of distance}\\ \text{measures}\end{array}\right\} \times \left\{\begin{array}{c}\text{\# ways of}\\ \text{differencing}\end{array}\right\}$$

Assuming a fixed sampling rate (the CD standard of 44.1 kHz), the "ways of choosing partitions" involves the variable ϵ, the time interval over which the various measures are gathered and hence n, the number of samples in each partition. Also included are the overlap factors and different window functions that might be applied when extracting the data in a single partition from the complete waveform. Given the desire to utilize the faster FFT (rather than the slower DFT), we considered partitions with lengths that are powers of two. Early simulations suggested that partitions smaller than 2^9 did not have enough points to ensure reliable measures. Partitions longer than 2^{10} have an effective sampling rate on the order of 20 Hz or greater, which is only barely faster than the underlying rhythmic phenomenon. Using overlap is common when transforming into the frequency domain, though it is less common when running statistical tests. Accordingly, we considered three cases: no overlap, overlap of two, and overlap of four. When using no overlap, we used the rectangular window; otherwise we used the Hamming window. (Early simulations showed very little difference between the standard windows such as Kaiser, Dolph, Hamming, and Hann.) In total, these choices give six different partitions.

Once the data is partitioned and the domain chosen from table I, then a way must be chosen to measure the distance between the vector in the kth partition and the vector in the $k+1$st partition. This can be done in many ways. In table II, $z_1, z_2, \ldots, z_n$ represents the data in partition k and $y_1, y_2, \ldots, y_n$ represents the data in partition $k-1$. The x_i may represent either z_i or the difference $z_i - y_i$. Thus the first six methods actually specify twelve different ways of measuring the distance, and there are a total of 24 different methods.

Finally, three "ways of differencing" were discussed in examples 1-3:

a. using the value directly,

b. using the difference between successive values, and

c. using the percent change.

Considering all these possibilities, there are $6 \cdot 17 \cdot 24 \cdot 3 = 7344$ ways of creating rhythm tracks! Actually there are somewhat fewer because some of the measures are redundant (for instance, using the sum square of the magnitude of the FFT is directly proportional to the energy in time by Parseval's theorem) and some are degenerate (for instance, measuring the total variation of a CDF always leads to the same answer). Nonetheless, a large number remain, and the next section is devoted to the design of

TABLE II

IN THIS TABLE, $z_1, z_2, \ldots, z_n$ REPRESENTS (POSSIBLY TRANSFORMED) DATA IN PARTITION k AND $y_1, y_2, \ldots, y_n$ REPRESENTS THE (POSSIBLY TRANSFORMED) DATA IN PARTITION $k-1$. THE x_i MAY REPRESENT EITHER z_i (METHODS 1-6) OR THE DIFFERENCE $z_i - y_i$ (METHODS 19-24).

	measure	method
1	$\sum_i x_i^2$	l^2 energy
2	$\sum_i \lvert x_i \rvert$	l^1 norm
3	$\sum_i \log(x_i^2)$	log energy
4	$\sum_i x_i^2 \log(x_i^2)$	entropy
5	$\sum_i \lvert x_i \rvert \log(\lvert x_i \rvert)$	absolute entropy
6	argmax $\lvert x_i \rvert$	location of maximum
7	max $\lvert z_i - y_i \rvert$	KS test (for CDF)
8	$\#\{i : \lvert z_i \rvert > \eta\}, \eta = \text{mean}(z_i)$	number of $\lvert z_i \rvert$ larger than mean
9	$\eta = \sqrt{2 \log_e(n \log_2(n))}$	SURE threshold in measure 8
10	$\max(z_i) - \min(z_i)$	range of data
11	$\min_m \sum_i (z_i - mi - b)^2$	slope
12	$\min_m \lvert \sum_{j=1}^m z_j^2 - \frac{1}{2} \sum_{i=1}^n z_i^2 \rvert$	center
13	$\sum_i (m - z_i^2)^2$	dispersion about center m
14	$\sum_i \lvert z_{i+1} - z_i \rvert$	total absolute variation
15	$\sum_i (z_{i+1} - z_i)^2$	total square variation
16	$\sum_i \log(\frac{\lvert z_i \rvert}{\lvert y_i \rvert})$	cross information
17	$\sum_i \lvert z_i \rvert \log(\frac{\lvert z_i \rvert}{\lvert y_i \rvert}) + \lvert y_i \rvert \log(\frac{\lvert y_i \rvert}{\lvert z_i \rvert})$	symmetrized cross entropy
18	$\sum_i i z_i^2$	weighted energy

an empirical test that can sift through the myriad of possible rhythm tracks, leaving only the best.

IV. MEASURING THE QUALITY OF RHYTHM TRACKS

What makes a good rhythm track? First, it must clearly delineate auditory boundaries. Second, it should emphasize those auditory boundaries that occur at regular intervals, that is, it should emphasize the beat. Schematically, the rhythm track should have a lattice-like structure in which large values regularly occur among a sea of smaller values. Suppose that the beat locations $b_1, b_2, \ldots, b_M$ are known for a

particular piece of music. Then the quality of a given rhythm track for that piece can be measured by how well the rhythm track reflects the known structure, that is, if the rhythm track regularly has large values near the b_i and small values elsewhere.

To make this concrete, rhythm tracks may be modeled as a collection of normal random variables with changing variances: the variance is small when "between" the auditory boundaries and large when boundaries occur, as illustrated in Fig. 1. Musically, the variance is small when "between" the beats and large when "on" the beat, and this model is explored in detail in [19]. In the simplest setting, the rhythm track values are assumed independent so that the probability of a block of values is the product of the probability of each value. This allows a concrete measure of the quality of a rhythm track by measuring the fidelity of the rhythm track to the model.

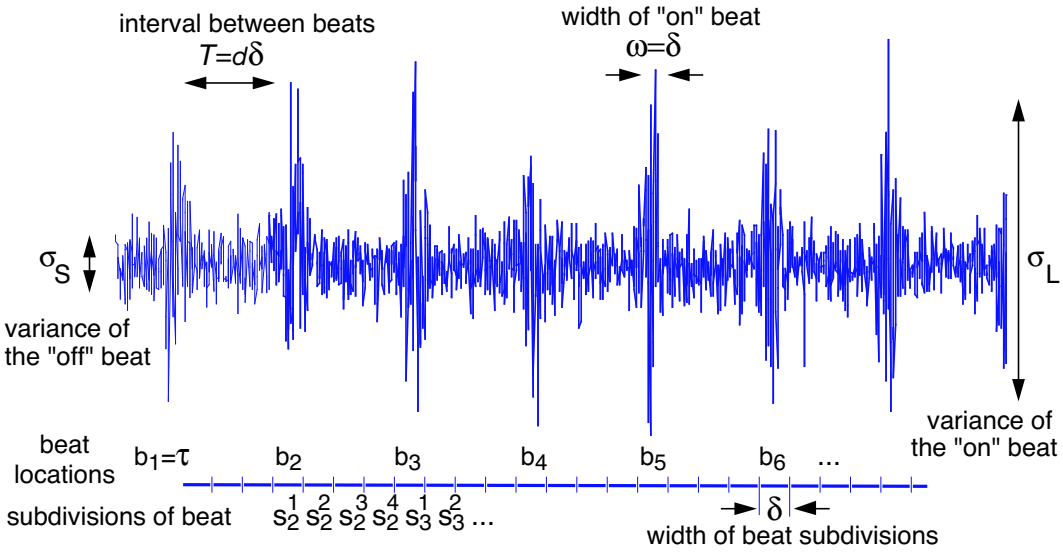


Fig. 1. Parameters of the rhythm track model are $\mathcal{T}$, τ, ω, σ_S, and σ_L. Parameters used in the evaluation of a rhythm track are the identified (known) beat locations b_i and the d δ-width subdivisions of the ith beat s_i^j. The $d = 4$ case is illustrated.

As a practical matter, there is some inaccuracy in the measurement of the b_i: let δ be chosen so that the rhythm track is expected to have large variance σ_L within each interval $\{b_i - \delta/2, b_i + \delta/2\}$ and small variance σ_S elsewhere. Since the average distance between beat boundaries is avg$\{b_{i+1} - b_i\}$, this may be divided into $d = \frac{\text{avg}\{b_{i+1}-b_i\}}{\delta}$ segments. Let $s_i^1, s_i^2, \ldots, s_i^d$ designate the d segments between the beat boundaries b_i and b_{i+1}.

A stochastic model for the rhythm track can be used to assign a probability to the two states "on the

beat" and "off the beat" for each segment. Bayes' Theorem gives

$$p_{on}(s_i) \propto p(s_i|\sigma_L)p(\text{on the beat})$$

where σ_L is the variance on the beat, and $p(\text{on the beat})$ is the prior probability of being on the beat, i.e., $1/d$. Similarly

$$p_{off}(s_i) \propto p(s_i|\sigma_S)p(\text{off the beat})$$

where $p(\text{off the beat}) = (d-1)/d$. The terms $p(s_i|\sigma_{L,S})$ are computed by evaluating the Gaussian PDF for each sample in the segment for the two values of variance, σ_L and σ_S, (σ_S is the variance off the beat), and forming the product over the samples.

For the purposes of assessing the quality of a rhythm track, a *utility function* [15] is defined which depends on the on/off beat state from the *known* beat locations. This is:

- *if a segment s_i is on the beat, an on-beat estimate has value M_{on}; an off-beat estimate has value 0.*
- *if the segment s_i is off the beat, an off-beat estimate has value M_{off}; an on-beat estimate has value 0.*

Typically $M_{on} > M_{off}$ because on-beats occur more rarely and are more important than off-beats. Using the on- and off-beat probabilities $p_{on}(s_i)$ and $p_{off}(s_i)$, the *expected value* for each segment is given by

$$E(s_i) = \begin{cases} M_{on} \times p_{on}(s_i) + 0 \times p_{off}(s_i) & \text{if on the beat} \\ M_{off} \times p_{off}(s_i) + 0 \times p_{on}(s_i) & \text{if off the beat} \end{cases} . \quad (4)$$

In practice, because of the very simple model for the stochastic structure of the rhythm tracks, for any given segment, one of $p_{on}(s_i)$ or $p_{off}(s_i)$ will be much larger than the other, and the prior terms will be negligible compared with $p(s_i|\sigma_{L,S})$. Thus, a good approximation to $E(s_i)$ is

$$E'(s_i) = \begin{cases} M_{on} & \text{if } p(s_i|\sigma_L) > p(s_i|\sigma_S), \text{ and on the beat} \\ M_{off} & \text{if } p(s_i|\sigma_S) > p(s_i|\sigma_L), \text{ and off the beat} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

and the total quality measure Q for the entire rhythm track is the average of $E'(s_i)$ over all segments in the rhythm track.

In words, the quality measure counts up how many times the segments are small when they are meant to be small, and how many times the segments are large when they are meant to be large. Because they are rarer, the latter are weighted (by the constants M_{on} and M_{off}), and the average of the values provides a measure of the fidelity of the rhythm track to the piece of music.

The procedure for determining the best rhythm tracks for a given corpus of music is now straightforward:

(a) Choose a set of test pieces for which the beat boundaries are known.

(b) For each piece and for each candidate rhythm track, calculate the quality measure Q.

(c) Those rhythm tracks which score highest over the complete set of test pieces are the best rhythm tracks.

The next section details extensive tests of the quality of the rhythm tracks of section III when applied to a large corpus of music. In order to calibrate the meaning of the tests, consider the value of the quality measure Q in the two extreme cases where (a) the rhythm track exactly fits the model and (b) where it is an undifferentiated sequence of independent identically distributed random variables. These two cases, which are presented in examples 10 and 11 below, bracket the quality of real rhythm tracks.

Let $x_j, j = 1, 2, \ldots n$ be a collection of i.i.d. Gaussian random variables in the segment s_i. The probability that $p(s_i|\sigma_L)$ is greater than $p(s_i|\sigma_S)$ is given by the probability that

$$\prod_{j=1}^{n} \frac{1}{\sigma_L} e^{-\frac{1}{2}\frac{x_j^2}{\sigma_L^2}} > \prod_{j=1}^{n} \frac{1}{\sigma_S} e^{-\frac{1}{2}\frac{x_j^2}{\sigma_S^2}}.$$

Taking the natural logarithm of both sides, this becomes

$$-\frac{1}{2\sigma_L^2}\sum_{j=1}^{n} x_j^2 - n\log(\sigma_L) > -\frac{1}{2\sigma_S^2}\sum_{j=1}^{n} x_j^2 - n\log(\sigma_S),$$

which can be solved for the sum as

$$\sum_{j=1}^{n} x_j^2 > \frac{2n\sigma_L^2\sigma_S^2(\log(\sigma_L) - \log(\sigma_S))}{\sigma_L^2 - \sigma_S^2}. \tag{6}$$

When the x_j are $N(0,1)$, the sum is distributed as a χ^2 random variable with n degrees of freedom, while the right hand side is constant. Accordingly, the desired probability can be read directly from a table of χ^2 distribution. Using the nominal values of $\sigma_L = 0.2$ and $\sigma_S = 0.1$, the right hand side of (6) is $0.0185n$.

Example 10: Using the nominal values for σ_L and σ_S, when $x_j \sim N(0, \sigma)$, the right hand side of (6) becomes $\frac{0.0185n}{\sigma^2}$. For a rhythm track with beat length of 300 ms, there are about 30 samples per beat. Divided into $d = 5$ segments implies $n = 6$ samples per segment. Thus, for $\sigma = .12$, the probability that $p(s_i|\sigma_L)$ is greater than $p(s_i|\sigma_S)$ is the same as the probability that a χ^2 random variable with $n = 6$ degrees of freedom is greater than $(0.0185 \cdot 6)/0.12^2 \approx 4.9$. This occurs about 50% of the time. Hence the quality measure for this rhythm track is

$$Q = 0.5\frac{M_{on}}{d} + 0.5\frac{(d-1)M_{off}}{d}.$$

For the nominal values $M_{on} = 3$ and $M_{off} = 1$, this gives a quality assessment of $Q = 0.7$.

Example 11: When the rhythm track exactly fits the model, $x_j \sim N(0, \sigma_L)$ for segments which are on the beat and $x_j \sim N(0, \sigma_S)$ for segments that are off the beat. Using the same numerical values as in Example 10, the probability that $p(s_i|\sigma_L)$ is greater than $p(s_i|\sigma_S)$ when on the beat is the same as the probability that a χ^2 random variable with $n = 6$ degrees of freedom is greater than $(0.0185 \cdot 6)/0.2^2 \approx 2.78$. This occurs about 85% of the time. On the other hand, the probability that $p(s_i|\sigma_L)$ is greater than $p(s_i|\sigma_S)$ when off the beat is the same as the probability that a χ^2 random variable with $n = 6$ degrees of freedom is greater than $(0.0185 \cdot 6)/0.1^2 \approx 11.1$, which occurs about 12% of the time. Thus the quality measure is

$$Q = 0.85\frac{M_{on}}{d} + (1 - 0.12)\frac{(d-1)M_{off}}{d}.$$

For the nominal values $M_{on} = 3$ and $M_{off} = 1$, this gives a quality assessment of $Q = 1.2$.
Examples 10 and 11 present the two extremes; it is reasonable to expect that the quality of rhythm tracks derived from real music should fall somewhere between $Q = 0.7$ and $Q = 1.2$.

V. Results

The procedure outlined in (a)-(c) above requires a body of music for which the beat boundaries are already known. In earlier work [19], we had used a set of four rhythm tracks combined with a particle filter (Bayesian estimator) to locate the beat locations in a number of different renditions of Scott Joplin's *Maple Leaf Rag*. These results were verified by superimposing a series of clicks (at the hypothesized beat locations) over the music. It was easy to hear that the algorithm correctly identified the beat. Audio examples are provided at the website [24].

The first step in the assessment was to measure the quality of all 7344 rhythm tracks on these known pieces. We selected the best rhythm tracks from this test, and used these to find the beat boundaries in 20 'pop' tunes, 20 'piano' pieces, and 20 'orchestral' pieces. These 60 pieces formed the body of music on which we then tested the rhythm tracks using (a)-(c) and the quality measure of section IV. Again, the success of the beat finding was verified by careful listening. The result of this procedure is a collection of rhythm tracks which are able to pinpoint the kinds of auditory boundaries that occur in a wide range of music. Details of the tests follow, and the results are analyzed and discussed in section VII.

A. *All the* Rags

The gnutella (peer to peer) file sharing network [4] was used to locate twenty-six versions of the 'Maple Leaf Rag' by Scott Joplin. About half were piano renditions, the instrument it was originally composed for. Other versions were performed on solo guitar, banjo, or marimba. Renditions were performed in

diverse styles: a kletzmer version, a bluegrass version, Sidney Bechet's big band version, one by the Canadian brass ensemble, and an orchestral version from the film 'The Sting.' In 22 of the 26 versions the beat was correctly located using the particle filter algorithm described in [19]. The procedure (a)-(c) was then applied for each of the 7344 rhythm tracks. Since the goal is to find rhythm tracks which perform well over a large collection of musical pieces, summary information about the quality of the rhythm tracks is reported in four complementary ways:

(I) the rhythm tracks that are among the best 10 (of 7344) for more than 5 performances,

(II) the rhythm tracks that are among the best 10 for more than 3 performances, which also have quality values $Q > 1$.

(III) the rhythm tracks with mean value $Q > 1$ when averaged over all performances,

(IV) the rhythm tracks with median over all performances of $Q > 1$.

Each row in table III lists all the rhythm tracks fulfilling conditions (I)-(IV) over the complete set of *Maple Leaf Rags*. The capital letter refers to the domain into which the data was transformed, as specified in table I. The integer indicates one of the distance measures 1-24 of table II. (Numbers 19-24 use the same processing as 1-6, but applied to the difference between adjacent partitions rather than the data within the partition.) The lower case letter specifies which method of differencing is used ('a'=raw data, 'b'=difference, 'c'=percent difference). In all cases, the 'winners' of this competition used a partition of size 2^{10} and an overlap of 2. Accordingly, these parameters were adopted for all succeeding tests, thus reducing the number of rhythm tracks to a more manageable 1224.

The single highest quality rating of any rhythm track was $Q = 1.32$, which was achieved by rhythm track B19a (which squares the energy in the FFT). This rhythm track does not appear in table III because its high Q is limited to few performances; it was only slightly better than the average rating of about $Q = 0.76$ overall. Of more interest are the rhythm tracks of general applicability, those which fulfill several of the conditions (I)-(IV). Inspection of table III reveals that there are eight rhythm tracks which fulfill all four conditions

B2b B9b I7a I7b I16a I18b I19a P3b

and nine more which fulfill three of the four conditions

B8b B13b B20b I22a J6b J12b K12b K13b K24b.

Of particular note are rhythm tracks B2b (energy in the FFT) and B13b (dispersion about the spectral center) which were two of the original four rhythm tracks proposed in [19]. Detailed discussion of the most successful performance measures is postponed to section VII.

TABLE III

RHYTHM TRACKS WITH THE HIGHEST OVERALL QUALITY RATINGS OVER THE SET OF 22 RENDITIONS OF THE *Maple Leaf Rag*.

Condition	Rhythm Tracks
(I)	B2b B9b B13c B18c I7a I7b I16a I18b I19a P3b
(II)	B2b B8b B9b B13b B20b I7a I7b I16a I19a I22a I18b I19b J6b J12b K12b K13b K24b P16a P3b
(III)	B2b B8b B9b B13b B20b I7a I7b I16a I19a I22a I18b I19b J6b J12b J24b K12b K13b K24b P3b Q18b
(IV)	B2b B2c B8b B9b B13b B20b I7a I7b I16a I18b I19a I22a I24b J6b J6c J12b J24b K12b K13b K24b P3b Q18b Q18c

B. The Best Performance Measures

Using the 17 best rhythm tracks from table III in conjunction with the particle filter beat tracking algorithm of [19], beat boundaries were successfully derived in three sets of pieces from three musical genres: "pop," "piano," and "large ensemble." The twenty "pop" pieces included songs by the Monkees ("I'm a Believer"), the Byrds ("Mr. Tambourine Man"), Leo Kottke ("Everybody Lies"), Creedence Clearwater Revival ("Proud Mary"), the Beatles ("Norwegian Wood"), among others. The twenty "piano" pieces included several of the preludes from the Well Tempered Clavier (in both harpsichord and piano versions), Scarlattis's K517 and K450 sonatas, and several pieces of early music from the Baltimore Consort. The twenty large ensemble pieces included rhythmic instrumentals such as Handel's "Water Music" and "Hornpipe," Strauss's "Fire Polka," and Sousa's "Stars and Stripes Forever."

The testing procedure (a)-(c) was applied to these three sets of pieces, and quality values Q were obtained for each of the 1224 rhythm tracks. Because there were so many rhythm tracks with $Q > 1$, conditions (II)-(IV) were amended to report only those rhythm tracks with quality values $Q > 1.1$. The results are summarized in tables IV-VI. As in the analysis of the *Maple Leaf Rag* performance measures, the 'best' rhythm tracks are those which simultaneously fulfill three or more of the conditions (I)-(IV). These are

B2b H1b H5b H10b H10c H18b H18c H22a
I1b I1c I2b I3b I9b I16a I18b I18c I22a K8b K13b

for the pop songs,

B2b B13b H1b H5b H9b H18b I1b I2b I3b I4b I5b I9b I18b P3b P16a

for the piano tunes, and

H1b H4b H5b H9b H10b H15b H18b K18b

for the large ensemble pieces. There are 28 distinct rhythm tracks in these three groups.

TABLE IV

RHYTHM TRACKS WITH THE HIGHEST OVERALL QUALITY RATINGS OVER THE SET OF TWENTY "POP" SONGS.

Condition	Rhythm Tracks
(I)	B2b B20b H1b H5b H10b H18b H18c H20b H22a
	I2b I7b I9b I16a I18b I18c I22a K8b K13b K18b
(II)	B2b B18b B19b B20b H1b H1c H5b H7a H9b H10b
	H14b H14c H15c H17a H18b H20b H22a H23b I1b
	I1c I2b I3b I7a I9b I18b I18c I22a
(III)	B2b B5b B8b B13b B17b H1b H1c H5b H9b H10b
	H14b H18b H22a H23b H10c H14c H15c H18c I1b I1c I2b
	I2c I3b I5b I9b I16a I16b I18b I18c I18c I19b I20b K8B
	I22a J6b J6c J12b J24b J24c K12b K13b K24b K24c Q2b Q13b Q18b
(IV)	B2b H1b H1c H5b H10b H18b H18c I1b I1c I2b I2c I3b
	I9b I16a I18b I18c I19b I22a J12b J24c K2b K8b K13b K18b

VI. INDEPENDENCE OF RHYTHM TRACKS

The tests of the previous sections examine a large number of rhythm tracks but provide no guarantee that individual tracks contain unique information. For example, if a rhythm track has a high quality value over a large number of pieces, a copy of that rhythm track would also have a high quality value, and so both would pass into the tables IV-VI. This section discusses a simple way to identify the most redundant rhythm tracks.

Suppose there are n rhythm tracks $P_1, P_2, \ldots P_n$ derived from the same piece of music containing L beats. Then each P_i can be thought of as a vector in L-dimensional space. If m of these vectors are

TABLE V

RHYTHM TRACKS WITH THE HIGHEST OVERALL QUALITY RATINGS OVER THE SET OF TWENTY "PIANO" TUNES.

Condition	Rhythm Tracks
(I)	B2b B13b B20b H1b H5b H9b H18b I16a I19a I23b P3b P16a
(II)	B13b H5b I19a I23b P3b P16a
(III)	B2b B9b H1b H4b H5b H9b H10b H18b I1b I2b I3b I4b I5b I9b I18b I19b
(IV)	B2b B13b B17a B17b B20b H1b H4b H5b H5c H7a H7b H9b H10b H14b H15b H18b H18c I1b I1c I2b I2c I3b I4b I5b I7a I9b I16a I18b I20b P3b P16a P16b

TABLE VI

RHYTHM TRACKS WITH THE HIGHEST OVERALL QUALITY RATINGS OVER THE SET OF TWENTY "LARGE ENSEMBLE" PIECES.

Condition	Rhythm Tracks
(I)	B9b H1b H4b H5b H9b H10b H15b H18b I18b P3b
(II)	H1b H4b H5b H9b H9c H10b H14b H15b H18b K13b K18b P3b
(III)	H1b H4b H5b H9b H10b H18b I18c I19b K13b K18b
(IV)	H1b H4b H5b H9b H9c H10b H14b H15b H18b I19b K13b K18b

linearly dependent, then (after some possible relabelling) there are constants $a_1, a_2, \ldots, a_m$ with

$$a_1 P_1 + a_2 P_2 + \ldots + a_m P_m = 0.$$

This can be written in matrix form as $Pa = 0$ where the P_i form the columns of P and a is a vector of the a_i. If the vectors P_i are close to dependent, then Pa should be small. This can be written as the problem of minimizing $(Pa)'Pa = a'P'Pa$. 'Closeness' to dependence can thus be quantified by calculating the singular values of P (i.e., the eigenvalues of $P'P$). A singular value near zero corresponds to near linear dependence of the rhythm tracks.

A complete test of all the best rhythm tracks would involve far too much computation (for example, locating the 6 'most independent' vectors among the 28 best rhythm tracks would require $\binom{28}{6}$ eigenvalue calculations). Since we had observed that some pairs of rhythm tracks appeared to be quite similar to

each other, we tested for dependencies among all pairs of rhythm tracks over the set of 60 musical pieces. Considering rhythm tracks dependent if the ratio of the singular values was greater than 100:1, there were six independent groups, as specified in table VII. Accordingly, only one rhythm track from each group is needed. The meaning of these rhythm tracks is discussed in the next section.

TABLE VII

MANY OF THE RHYTHM TRACKS CLUSTER INTO DEPENDENT GROUPS WHICH ESSENTIALLY DUPLICATE THE INFORMATION PROVIDED BY OTHERS IN THE SAME GROUP.

cluster	Rhythm Tracks
(i)	P3b P16a
(ii)	H1b H22a I1b
(iii)	H10c H18c I1c I18c I19c
(iv)	K8b K13b K18b
(v)	B2b I2b I3b I4b I5b I9b I16a
(vi)	B13b I18b H4b H5b H9b H10b H15b H18b

VII. DISCUSSION

A number of the 'best' performance measures in tables III-VII are based on the magnitude of the FFT (designated B in the tables). Since auditory perception is tightly tied with spectral features of the sound, frequency domain processing forms a common workhorse in audio processing. Indeed, this is why three of the four measures from our earlier work utilized the spectrum, though only two of these survive in the present results (B2b and B13b).

One interesting aspect of the results are the domains (recall table I) that are missing from the tables. None of the measures based directly on the time signal appear in tables III-VII. None of the measures based on the phase of the FFT, the cepstrum, or the direct histogram methods (the PDF or CDF of the time signal) appear in the tables either. Of all the subbands, only the highest ones appear, and only the P-subband appears reliably. Of course, many of these missing domains can form good rhythm tracks. For example, the time domain method A combined with the total variation measure 14 achieved a quality rating of $Q > 1.2$ for three different pieces in the 'pop' music tests. It does not appear in the tables, however, because its average was below $Q = 1$ over all the pieces.

What does appear repeatedly are the domains B, H, and I: the magnitude of the FFT, its CDF and its

PDF. The latter two are surprising because of their hybrid nature (the combination one of the histogram methods with the spectrum) and because they are, for the most part, absent from the literature. In addition, domains J and K (the CDF and PDF of the cepstrum) occur in several places in tables III-VII though they rarely pass the '3 of 4' criterion. Again, these are somewhat unexpected.

Despite the wide variety of successful measures, table VII suggests that there are only a handful of truly different features being measured. Some of these are easy to interpret and others are not.

Probably the easiest to understand is cluster (i) which contains two methods that measure the change in entropy and the cross information of the high frequency subband P. It is reasonable to interpret these rhythm tracks as reflecting strong rhythmic overtones that occur in regular on-beat patterns. Interestingly, the low frequency subbands (L, M, N, and O) are completely absent from the tables, though some were successful for particular musical pieces. This suggests that the use of bass frequencies alone (as in [1] and [2]) is unlikely to be applicable to a wide variety of musical styles.

Cluster (ii) contains both I1b and H1b: the change in the energy in the PDF and CDF of the magnitude spectrum. These histogram-based *histospectral* methods may be interpreted as distinguishing dense spectra (such as noises) from thinner line spectra (such as tonal material). To see this, observe that the PDF (the histogram) of the FFT magnitude shows how many frequencies there are at each magnitude level. When the spectrum is uniformly dense (such as during the attack phase of a note), the histogram tends to have a sharp cutoff. When the spectrum has more shape (such as the gentle rolloff of the spectrum of a string or wind instrument) then the histogram also has a more gentle slope.

The unifying factor in cluster (iii) is the appearance of the distance measure c, which gives the percent change (rather than the change itself) between successive partitions. This would be most crucial when dealing with music that has a large dynamic range since the percent change in a soft passage is (roughly) the same as the percent change in the same passage played loudly. The fact that H18c and I18c can both be found in group (iii) is also easily understood since these are weighted versions of the energy. The fact that H10c is also redundant with these suggests that the energy measurements may be dominated by extreme values.

All the K domain measures are clustered in (iv), so it is reasonable to suppose that the CDF of the cepstrum provides only one unique dimension of information. The cepstrum provides information about the rate of change in the different frequency bands. It has been used in speech processing to separate the glottal frequencies from the vocal tract resonances and in speech recognition studies, where the squared error between cepstral coefficients in adjacent frames provides a distance measure [13]. In our studies, none

of the performance measures that were based on the cepstrum alone were competitive. However, taking histograms and comparing the empirical PDF and/or CDF of the cepstrum (taking the *histocepstrum*) in successive segments provided successful performance measures in the beat tracking application.

Cluster (v) is dominated by measures based on domain I. One way that distances 2, 3, 4, and 5 can all result in similar rhythm tracks would be if the CDF's consist primarily of zeroes and ones. The distinguishing feature would then be the location of the jump discontinuity between successive partitions. Translating back to the PDF, this would represent the magnitude at which the frequencies begin to drop off rapidly. In terms of the spectrum, this would be a cutoff frequency; below this cutoff the magnitudes are large, above this cutoff, the magnitudes are small. Assuming this line of reasoning is correct, cluster (v) is measuring changes in the bandwidth of the signal over time.

The final cluster (vi) contains B13b, the dispersion about the center frequency. This can be interpreted as distinguishing sounds with widely scattered spectra from those with more compact spectra. This interpretation is also consistent with the entropy measures (H4 and H5) since entropy is maximized when all values x_i are equal and is small when the x_i are disparate.

VIII. Conclusions

This paper has considered a variety of performance measures for the delineation of auditory boundaries with the goal of tracking beats in musical performances. A quantitative test for the quality of the measures was proposed and each of the measures was applied to a set of sixty musical performances drawn from three distinct musical styles. A large variety of measures were considered, though after appropriate clustering, there appeared to be only six distinct features. We attempted to interpret these features in terms of their possible perceptual meaning.

Perhaps the most interesting result of these experiments is the uncovering of several classes of performance measures: the histospectral methods (the CDF and PDF of the magnitude spectrum) and histocepstral methods (the CDF and PDF of the cepstrum) that have not been previously applied to audio processing. Our results show that these measures are clearly useful in the beat tracking application. It is reasonable to suppose that these same measures may also find use in other areas where it is necessary to delineate auditory (or other kinds of signal) boundaries.

References

[1] M. Alghoniemy and A. H. Tewfik, "Rhythm and periodicity detection in polyphonic music," 1999 International Workshop on Multimedia Signal Processing, Copenhagen, Denmark, 1999.

[2] T. L. Blum, D. F. Keislar, J. A. Wheaton, and E. H. Wold, "Method and article of manufacture for content-based analysis, storage, retrieval, and segmentation of audio information," U.S. Patent No. 5,918,223, 1999.

[3] P. Desain, "A (de)composable theory of rhythm perception," *Music Perception*, Vol. 9, No. 4, 439-454, Summer 1992.

[4] http://www.gnutella.com

[5] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *J. New Music Research*, Vol. 30, No. 2, pp. 159-171, 2001.

[6] F. Gouyon and P. Herrera, "Determination of the meter of musical audio signals: seeking recurrences in beat segment descriptors," *Proc. of AES*, 114th Convention, 2003.

[7] F. Gouyon and B. Meudic, "Towards rhythmic content processing of musical signals: fostering complementary approaches," *J. New Music Research*, Vol. 32, No. 1, pp. 159-171, 2003.

[8] M. R. Jones and M. Boltz, "Dynamic attending and responses to time," *Psychological Review*, 96 (3) 459-491, 1989.

[9] E. W. Large and J. F. Kolen, "Resonance and the perception of musical meter," *Connection Science* 6: 177-208, 1994.

[10] M. Leman, *Music and Schema Theory: Cognitive Foundations of Systematic Musicology*, Berlin, Heidelberg: Springer-Verlag 1995.

[11] R. D. Morris and W. A. Sethares, "Beat Tracking," Seventh Valencia International Meeting on Bayesian Statistics, Tenerife, Spain, June 2002.

[12] B. Porat, *Digital Signal Processing*, Wiley 1997.

[13] L. R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals* Prentice-Hall, New Jersey, 1978.

[14] C. Roads, *Microsound*, MIT Press, 2002.

[15] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Springer-Verlag, 2001.

[16] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoustical Society of America*, 103 (1), 588-601, Jan. 1998.

[17] J. Seppänen, "Computational models of musical meter recognition," MS Thesis, Tempere Univ. Tech. 2001.

[18] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale*, Springer-Verlag 1997.

[19] W. A. Sethares, R. D. Morris and J. C. Sethares, "Beat tracking of audio signals," accepted for publication in *IEEE Trans. On Speech and Audio Processing*. A preliminary version is available on line at http://eceserv0.ece.wisc.edu/~sethares/beatstuff/beatrack4.pdf

[20] W. A. Sethares and T. Staley, "The periodicity transform," *IEEE Trans. Signal Processing*, Vol. 47, No. 11, Nov. 1999.

[21] W. A. Sethares and T. Staley, "Meter and periodicity in musical performance", *J. New Music Research*, Vol. 30, No. 2, June 2001.

[22] N. P. M. Todd, D. J. O'Boyle and C. S. Lee, "A sensory-motor theory of rhythm, time perception and beat induction," *J. of New Music Research*, 28:1:5-28, 1999.

[23] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals", *IEEE Trans. Speech Audio Processing*, 10(5), July 2002.

[24] Website for musical examples can be found at http://eceserv0.ece.wisc.edu/~sethares/beatrack

[25] T. Zhang and J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Processing*, Vol. 9, pp. 441457, May 2001.