A. P. Klapuri, A. J. Eronen, J. T. Astola, “Analysis of the meter of acoustic musical signals,” *IEEE Trans. Speech and Audio Proc.* (in press).

Analysis of the Meter of Acoustic Musical Signals

Anssi P. Klapuri, Antti J. Eronen, and Jaakko T. Astola

***Abstract*— A method is decribed which analyzes the basic pattern of beats in a piece of music, the musical meter. The analysis is performed jointly at three different time scales: at the temporally atomic *tatum* pulse level, at the *tactus* pulse level which corresponds to the tempo of a piece, and at the musical *measure* level. Acoustic signals from arbitrary musical genres are considered. For the initial time-frequency analysis, a new technique is proposed which measures the degree of musical accent as a function of time at four different frequency ranges. This is followed by a bank of comb filter resonators which extracts features for estimating the periods and phases of the three pulses. The features are processed by a probabilistic model which represents primitive musical knowledge and uses the low-level observations to perform joint estimation of the tatum, tactus, and measure pulses. The model takes into account the temporal dependencies between successive estimates and enables both causal and noncausal analysis. The method is validated using a manually annotated database of 474 music signals from various genres. The method works robustly for different types of music and improves over two state-of-the-art reference methods in simulations.**

***Index Terms*—Acoustic signal analysis, music, musical meter analysis, music transcription.**

EDICS: *2-MUSI*

I. Introduction

Meter analysis, here also called *rhythmic parsing*, is an essential part of understanding music signals and an innate cognitive ability of humans even without musical education. Perceiving the meter can be characterized as a process of detecting moments of musical stress (accents) in an acoustic signal and filtering them so that underlying periodicities are discovered [1], [2]. For example, tapping a foot to music indicates that the listener has abstracted metrical information about music and is able to predict when the next beat will occur.

Musical meter is a hierarchical structure, consisting of pulse sensations at different levels (time scales). Here, three metrical levels are considered. The most prominent level is the *tactus*, often referred to as the foot tapping rate or the beat. Following the terminology of [1], we use the word *beat* to refer to the individual elements that make up a pulse. A musical meter can be illustrated as in Fig. 1, where the dots denote beats and each sequence of dots corresponds to a particular pulse level. By the *period* of a pulse we mean the time duration between successive beats and by *phase* the time when a beat occurs with respect to the beginning of the piece. The *tatum* pulse has its name stemming from "temporal atom" [3]. The period of this pulse corresponds to the shortest durational values in music that are still more than incidentally encountered. The other durational values, with few exceptions, are integer multiples of the tatum period and the onsets of musical events occur approximately at a tatum beat. The *musical measure* pulse is typically related to the harmonic change rate or to the length of a rhythmic pattern. Although sometimes ambiguous, these three metrical levels are relatively well-defined and span the metrical hierarchy at the aurally most important levels. The *tempo* of a piece is defined as the rate of the tactus pulse. In order that a meter would make sense musically, the pulse periods must be slowly varying and, moreover, each beat at the larger levels must coincide with a beat at all the smaller levels.

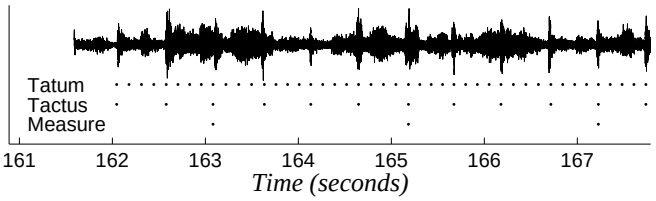

Fig. 1. A music signal with three metrical levels illustrated.

The concept *phenomenal accent* is important for meter analysis. Phenomenal accents are events that give emphasis to a moment in music. Among these are the beginnings of all discrete sound events, especially the onsets of long pitched events, sudden changes in loudness or timbre, and harmonic changes. Lerdahl and Jackendoff define the role of phenomenal accents in meter perception compactly by saying that "the moments of musical stress in the raw signal serve as cues from which the listener attempts to extrapolate a regular pattern" [1, p.17].

Automatic rhythmic parsing has several applications. A metrical structure facilitates cut-and-paste operations and editing of music signals. It enables synchronization with light effects, video, or electronic instruments, such as a drum machine. In a disc jockey application, metrical information can be used to mark the boundaries of a rhythmic loop or to synchronize two audio tracks. Provided that a time-stretching algorithm is available, rhythmic modifications can be made to audio signals [4]. Rhythmic parsing for symbolic (MIDI[1]) data is required for *time quantization*, an indispensable subtask of score typesetting from keyboard input [5]. The particular motivation for the present work is to utilize metrical information in further signal analysis and in music transcription [6], [7], [8].

A. Previous work

The work on automatic meter analysis originated from algorithmic models that attempted to explain how a human

A. P. Klapuri is with Institute of Signal Processing, Tampere University of Technology, FIN-33720 Tampere, Finland (e-mail: Anssi.Klapuri@tut.fi).

[1]Musical Instrument Digital Interface. A standard interface for exchanging performance data and parameters between electronic musical devices.

listener arrives at a particular metrical interpretation of a piece. An extensive analysis of the early models has been given by Lee in [9] and later augmented by Desain and Honing in [10]. In brief, the early models performed meter analysis for symbolic data (impulse patterns) and can be seen as being based on a *set of rules* that were used to define what makes a musical accent and to infer the most natural meter.

More recently, Rosenthal proposed a system to emulate the human rhythm perception for piano performances, presented as MIDI files [11]. Parncutt developed a detailed algorithmic model of meter perception based on systematic listening tests [12]. Brown analyzed the meter of musical scores by processing the onset times and durations of note events using the autocorrelation function [13]. Large and Kolen used adaptive oscillators which adjust their period and phase to an incoming pattern of impulses, located at the onsets of musical events [14].

Temperley and Sleator [15] proposed a meter analysis algorithm for arbitrary MIDI files by implementing the preference rules that were described in verbal terms by Lerdahl and Jackendoff in [1]. Dixon proposed a rule-based system to track the tactus pulse of expressive MIDI performances and introduced a simple onset detector to make the system applicable for audio signals [16]. The source codes of both Temperley's and Dixon's systems are publicly available for testing.

Cemgil and Kappen developed a probabilistic generative model for the timing deviations in expressive musical performances [5]. Then, the authors used Monte Carlo methods to infer a hidden continuous tempo variable and quantized ideal note onset times from observed noisy onset times in a MIDI file. A similar Bayesian model was independently proposed by Raphael [17].

Goto and Muraoka were the first to achieve a reasonable meter analysis accuracy for audio signals [18], [19]. Their system operated in real time and was based on an architecture where multiple agents tracked competing meter hypotheses. Beat positions at the larger levels were inferred by detecting certain drum sounds [18] or chord changes [19].

Scheirer proposed an approach to tactus tracking where no discrete onsets or sound events are detected as a middle-step, but periodicity analysis is performed directly on the half-wave rectified differentials of subband power envelopes [20]. The source code of Scheirer's system is publicly available. Sethares and Staley took a similar approach, but used a periodicity transform for periodicity analysis instead of a bank of comb filters [21]. Laroche proposed a noncausal algorithm where spectral change was measured as a function of time, the resulting signal was correlated with impulse trains of different periods, and dynamic programming was used to find a continuous time-varying tactus pulse [22].

Hainsworth and Macleod [23] developed a method which is loosely related to that of Cemgil et al. [5]. They extracted discrete onsets from an audio signal and then used particle filters to associate the onsets to a time-varying tempo process and to find the locations of the beats. Gouyon et al. proposed a system for detecting the tatum pulse in percussive audio tracks of constant tempo [24].

In summary, most of the earlier work on meter analysis has concentrated on symbolic (MIDI) data and typically analyzed the tactus pulse only. Some of the systems ([14], [16], [5], [17]) can be immediately extended to process audio signals by employing an onset detector which extracts the beginnings of discrete acoustic events from an audio signal. Indeed, the authors of [16] and [17] have introduced an onset detector themselves. Elsewhere, onset detection methods have been proposed that are based on using subband energies [25], an auditory model [26], support vector machines [27], independent component analysis [28], or a complex-domain distance measure [29]. However, if a rhythmic parser has been originally developed for symbolic data, the extended system is usually not robust to diverse acoustic material (e.g. classical vs. rock music) and cannot fully utilize the acoustic cues that indicate phenomenal accents in music signals.

There are a few basic problems that need to be addressed in a successful meter analysis system. First, the degree of musical accentuation as a function of time has to be measured. Some systems do this in a continuous manner ([20], [21]), whereas others extract discrete onsets from an audio signal ([18], [24], [22]). Secondly, the periods and phases of the underlying metrical pulses have to be estimated. The methods which detect discrete events as a middle-step have often used inter-onset-interval histograms for estimating the periods [16], [18], [19], [24]. Thirdly, a system has to choose the metrical level which corresponds to the tactus or some other specially designated pulse level. This may take place implicitly, or by using a prior distribution for pulse periods [12] or by rhythmic pattern matching [18].

B. Proposed method

The aim of this paper is to describe a method which analyzes the meter of acoustic musical signals at the tactus, tatum, and measure pulse levels. The target signals are not limited to any particular music type but all the main Western genres, including classical music, are represented in the validation database.

An overview of the method is shown in Fig. 2. For the time-frequency analysis part, a technique is proposed which aims at measuring the degree of accentuation in a music signal. The technique is robust to diverse acoustic material and can be loosely seen as a synthesis and generalization of two earlier state-of-the-art methods [18] and [20]. Feature extraction for estimating the pulse periods and phases is performed using comb filter resonators very similar to those used by Scheirer in [20]. This is followed by a probabilistic model where the period-lengths of the tactus, tatum, and measure pulses are jointly estimated and temporal continuity of the estimates is modelled. At each time instant, the periods of the pulses are estimated first and act as inputs to the phase model. The probabilistic models encode prior musical knowledge and lead to a more reliable and temporally stable meter tracking. Both causal and non-causal algorithms are presented.

This paper is organized as follows. Section II will describe the different elements of the system shown in Fig. 2. Section III will present experimental results and compare

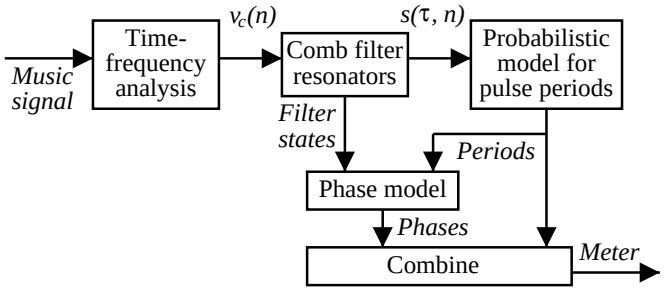

Fig. 2. Overview of the meter estimation method. The two intermediate data representations are bandwise accent signals $v_c(n)$ and metrical pulse saliences (weights) $s(\tau, n)$.

the proposed method with two reference methods. The main conclusions will be summarized in Section IV.

II. Meter Analysis Model

This section will describe the different parts of the meter analysis method illustrated in Fig. 2. Subsection II-A will describe the time-frequency analysis part. In Subsection II-B, the comb filter resonators will be introduced. Subsections II-C and II-D will describe the probabilistic models which are used to estimate the periods and phases of the three pulse levels.

A. Calculation of bandwise accent signals

All the phenomenal accent types mentioned in the introduction can be observed in the time-frequency representation of a signal. Although an analysis using a model of the human auditory system might seem theoretically advantageous (since meter is basically a cognitive phenomenon), we did not manage to obtain a performance advantage using a model similar to [26] and [30]. Also, the computational complexity of such models makes them rather impractical.

In a time-frequency plane representation, some data reduction must take place to discard information which is irrelevant for meter analysis. A big step forward in this respect was taken by Scheirer who demonstrated that the perceived rhythmic content of many music types remains the same if only the power envelopes of a few subbands are preserved and then used to modulate a white noise signal [20]. Approximately five subbands were reported to suffice. Scheirer proposed a method where periodicity analysis was carried out within the subbands and the results were then combined across bands.

Although Scheirer's method was indeed very successful, a problem with it is that it applies primarily to music with a "strong beat". Harmonic changes for example in classical or vocal music go easily unnoticed using only a few subbands. In order to detect harmonic changes and note beginnings in *legato*[2] passages, approximately 40 logarithmically-distributed subbands would be needed.[3] However, this leads to a dilemma: the resolution is sufficient to distinguish harmonic changes but measuring periodicity at each narrow band separately is no longer appropriate. The power envelopes of individual narrow bands are not guaranteed to reveal the correct metrical periods—or even to show periodicity at all, because individual events may occupy different frequency bands.

To overcome the above problem, consider another state-of-the-art system, that of Goto and Muraoka [18]. They detect narrowband frequency components and sum their power differentials across predefined frequency ranges *before* onset detection and periodicity analysis takes place. This has the advantage that harmonic changes are detected, yet periodicity analysis takes place at wider bands.

There is a continuum between the above two approaches. The tradeoff is: how many adjacent subbands are combined before the periodicity analysis and how many at the later stage when the bandwise periodicity analysis results are combined. In the following, we propose a method which can be seen as a synthesis of the approaches of Scheirer and Goto et al.

Acoustic input signals are sampled at 44.1 kHz rate and 16-bit resolution and then normalized to have zero mean and unity variance. Discrete Fourier transforms are calculated in successive 23 ms time frames which are Hanning-windowed and overlap 50 %. In each frame, 36 triangular-response bandpass filters are simulated that are uniformly distributed on a critical-band scale between 50 Hz and 20 kHz [31, p.176]. The power at each band is calculated and stored to $x_b(k)$, where k is the frame index and $b = 1, 2, \ldots, b_0$ is the band index, with $b_0 = 36$. The exact number of subbands is not critical.

There are many potential ways of measuring the degree of change in the power envelopes at critical bands. For humans, the smallest detectable change in intensity, ΔI, is approximately proportional to the intensity I of the signal, the same amount of increase being more prominent in a quiet signal. That is, $\frac{\Delta I}{I}$, the Weber fraction, is approximately constant perceptually [31, p.134]. This relationship holds for intensities from about 20 dB to about 100 dB above the absolute hearing threshold. Thus it is reasonable to normalize the differential of power with power, leading to $\frac{d}{dt}x_b(k)/x_b(k)$ which is equal to $\frac{d}{dt}\ln(x_b(k))$. This measures spectral change and can be seen to approximate the differential of *loudness*, since the perception of loudness for steady sounds is rougly proportional to the sum of log-powers at critical bands.

The logarithm and differentiation operations are both represented in a more flexible form. A numerically robust way of calculating the logarithm is the μ-law compression,

$$y_b(k) = \frac{\ln(1 + \mu x_b(k))}{\ln(1 + \mu)}, \tag{1}$$

which performs a logarithmic-like transformation for $x_b(k)$ as motivated above but behaves linearly near zero. The constant μ determines the degree of compression and can be used to adjust between a close-to-linear ($\mu < 0.1$) and a close-to-logarithmic ($\mu > 10^4$) transformation. The value $\mu = 100$ is employed, but all values in the range $[10, 10^6]$ were found to perform almost equally well.

To achieve a better time resolution, the compressed power envelopes $y_b(k)$ are interpolated by factor two by adding zeros between the samples. This leads to the sampling rate

[2] A smooth and connected style of playing in which no perceptible gaps are left between notes.

[3] In this case, the center frequencies are approximately one *whole tone* apart, which is the distance between e.g. the notes c and d.

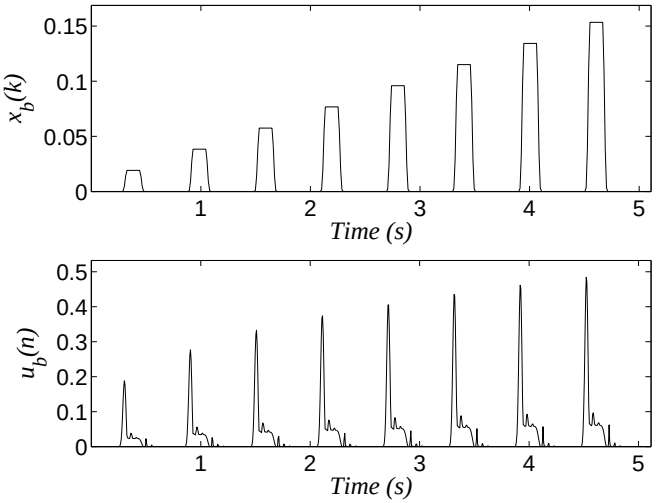

Fig. 3. Illustration of the dynamic compression and weighted differentiation steps for an artificial signal. Upper panel shows $x_b(k)$ and the lower panel shows $u_b(n)$.

$f_r = 172$ Hz. A sixth-order Butterworth lowpass filter with $f_{\mathrm{LP}} = 10$ Hz cutoff frequency is then applied to smooth the compressed and interpolated power envelopes. The resulting smoothed signal is denoted by $z_b(n)$.

Differentiation of $z_b(n)$ is performed as follows. First, a half-wave rectified (HWR) differential of $z_b(n)$ is calculated as

$$z_b{}'(n) = \mathrm{HWR}(z_b(n) - z_b(n-1)), \tag{2}$$

where the function $\mathrm{HWR}(x) = \max(x, 0)$ sets negative values to zero and is essential to make the differentiation useful. Then a weighted average of $z_b(n)$ and its differential $z_b{}'(n)$ is formed as

$$u_b(n) = (1-\lambda)z_b(n) + \lambda\frac{f_r}{f_{\mathrm{LP}}}z_b{}'(n), \tag{3}$$

where $0 \leq \lambda \leq 1$ determines the balance between $z_b(n)$ and $z_b{}'(n)$, and the factor f_r/f_{LP} compensates for the fact that the differential of a lowpass-filtered signal is small in amplitude. A prototypical meter analysis system and a subset of our acoustic database (see Sec. III) were used to thoroughly investigate the effect of λ. Values between 0.6 and 1.0 performed well and $\lambda = 0.8$ was taken into use. Using this value instead of 1.0 makes a slight but consistent improvement in the analysis accuracy.

Figure 3 illustrates the described dynamic compression and weighted differentiation steps for an artificial subband-power signal $x_b(k)$. Although the present work is motivated purely from a practical application point of view, it is interesting to note that the graphs in Fig. 3 bear considerable resemblance to the response of Meddis's auditory-nerve model to acoustic stimulation [32].

Finally, each m_0 adjacent bands are linearly summed to get $c_0 = \lceil b_0/m_0 \rceil$ accent signals at different frequency ranges c:

$$v_c(n) = \sum_{b=(c-1)m_0+1}^{cm_0} u_b(n), \quad c = 1, \ldots, c_0. \tag{4}$$

The accent signals $v_c(n)$ serve as an intermediate data representation for musical meter analysis. They represent the degree of musical accent as a function of time at the wider frequency bands (channels) c. We use $b_0 = 36$ and $m_0 = 9$, leading to $c_0 = 4$.

It should be noted that combining each m_0 adjacent bands at this stage is not primarily an issue of computational complexity, but improves the analysis accuracy. Again, a prototypical meter analysis system was used to investigate the effect of different values of m_0. It turned out that neither of the extreme values $m_0 = b_0$ or $m_0 = 1$ is optimal, but using a large number of initial bands, $b_0 > 20$, and three or four "accent bands" (channels) c_0 leads to the most reliable meter analysis. Other parameters were re-estimated in each case to ensure that this was not merely a symptom of parameter couplings. Elsewhere, at least Scheirer [20] and Laroche [22] have noted that a single accent signal (the case $m_0 = b_0$) appears not to be sufficient as an intermediate representation for rhythmic parsing.

The presented form of calculating the bandwise accent signals is very flexible when varying μ, λ, b_0, and m_0. A representation similar to that used by Scheirer in [20] is obtained by setting $\mu = 0.1$, $\lambda = 1$, $b_0 = 6$, $m_0 = 1$. A representation roughly similar to that used by Goto in [18] is obtained by setting $\mu = 0.1$, $\lambda = 1$, $b_0 = 36$, $m_0 = 6$. In the following, the fixed values $\mu = 100$, $\lambda = 0.8$, $b_0 = 36$, $m_0 = 9$ are used.

B. Bank of comb filter resonators

Periodicity of the bandwise accent signals $v_c(n)$ is analyzed to estimate the *salience* (weight) of different pulse period candidates. Four different period estimation algorithms were evaluated: a method based on autocorrelation, another based on the *YIN* method of de Cheveigné and Kawahara [33], different types of comb-filter resonators [20], and banks of phase-locking resonators [14].

As an important observation, three of the four period estimation methods performed equally well after a thorough optimization. This suggests that the key problems in meter analysis are in measuring the degree of musical accentuation and in modeling higher-level musical knowledge, not in finding exactly the correct period estimator. The period estimation method presented in the following was selected because it is by far the least complex among the three best-performing algorithms, requiring only few parameters and no additional postprocessing steps.

Using a bank of comb-filter resonators with a constant half-time was originally proposed for tactus tracking by Scheirer [20]. The comb filters that we use have an exponentially-decaying impulse response where the *half-time* refers to the delay during which the response decays to a half of its initial value. The output of a comb filter with delay τ for input $v_c(n)$ is given by

$$r_c(\tau, n) = \alpha_\tau r_c(\tau, n-\tau) + (1-\alpha_\tau)v_c(n), \tag{5}$$

where the feedback gain $\alpha_\tau = 0.5^{\tau/T_0}$ is calculated based on a selected half-time T_0 in samples. We used a half-time

equivalent to three seconds, i.e., $T_0 = 3.0\text{s} \cdot f_r$, which is short enough to react to tempo changes but long enough to reliably estimate pulse-periods of up to four seconds in length.

The comb filters implement a frequency response where the frequencies kf_r/τ, $k = 0, \ldots, \lfloor \tau/2 \rfloor$ have a unity response and the maximum attenuation between the peaks is $((1 - \alpha_\tau)/(1 + \alpha_\tau))^2$. The overall power $\gamma(\alpha_\tau)$ of a comb filter with feedback gain α_τ can be calculated by integrating over the squared impulse response, which yields

$$\gamma(\alpha_\tau) = \frac{(1 - \alpha_\tau)^2}{1 - \alpha_\tau^2}. \tag{6}$$

A bank of such resonators was applied, with τ getting values from 1 to $\tau_{\max}$, where $\tau_{\max} = 688$ corresponds to four seconds. The computational complexity of one resonator is $O(1)$ per input sample, and the overall resonator filterbank requires of the order $c_0 f_r \tau_{\max}$ operations per second, which is not too demanding for real-time applications.

Instantaneous energies $\hat{r}_c(\tau, n)$ of each comb filter in channel c at time n are calculated as

$$\hat{r}_c(\tau, n) = \frac{1}{\tau} \sum_{i=n-\tau+1}^{n} r_c(\tau, i)^2. \tag{7}$$

These are then normalized to obtain

$$s_c(\tau, n) = \frac{1}{1 - \gamma(\alpha_\tau)} \left(\frac{\hat{r}_c(\tau, n)}{\hat{v}_c(n)} - \gamma(\alpha_\tau) \right), \tag{8}$$

where $\hat{v}_c(n)$ is the energy of the accent signal $v_c(n)$, calculated by squaring $v_c(n)$ and by applying a leaky integrator, i.e., a resonator which has $\tau = 1$ and the same three-second half-time as the other resonators. Normalization with $\gamma(\alpha_\tau)$ compensates for the differences in the overall power responses for different α_τ. The proposed normalization is advantageous because it preserves a unity response at the peak frequencies and at the same time removes a τ-dependent trend for a white-noise input.

Figure 4 shows the resonator energies $\hat{r}_c(\tau, n)/\hat{v}_c(n)$ and the normalized energies $s_c(\tau, n)$ for two types of artificial input $v_c(n)$: an impulse train and a white-noise signal. It is important to notice that all resonators that are in rational-number relations to the period of the impulse train (24 samples) show response to it. In the case of the autocorrelation function, for example, only integer multiples of 24 come up and an explicit postprocessing step was necessary to generate responses to the subharmonic lags and to achieve the same meter analysis performance. This step is not needed for comb filter resonators where the conceptual complexity and the number of free parameters thus remains smaller.

Finally, a function $s(\tau, n)$ which represents the overall saliences of different metrical pulses at time n is obtained as

$$s(\tau, n) = \sum_{c=1}^{c_0} s_c(\tau, n). \tag{9}$$

This function acts as the *observation* for the probabilistic model that estimates the pulse periods.

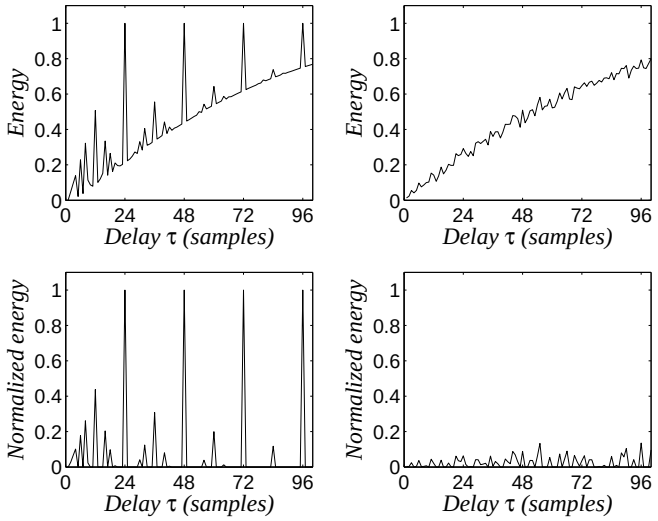

Fig. 4. Resonator energies for an impulse train with a period-length of 24 samples (left) and for white noise (right). Upper panels show the energies $\hat{r}_c(\tau, n)$ and the lower panels normalized energies $s_c(\tau, n)$.

For tatum period estimation, the discrete power spectrum $S(f, n)$ of $s(\tau, n)$ is calculated as

$$S(f, n) = f \left| \frac{1}{\tau_{\max}} \sum_{\tau=1}^{\tau_{\max}} \left(s(\tau, n) \zeta(\tau) e^{-i2\pi f(\tau-1)/\tau_{\max}} \right) \right|^2, \tag{10}$$

where the emphasis with f compensates for a spectral trend and the window function $\zeta(\tau)$ is half-Hanning:

$$\zeta(\tau) = 0.5(1 - \cos(\pi(\tau - 1 + \tau_{\max})/\tau_{\max})). \tag{11}$$

The rationale behind calculating the discrete Fourier transform (DFT) in (10) is that, by definition, other pulse periods are integer multiples of the tatum period. Thus the overall function $s(\tau, n)$ contains information about the tatum and this is conveniently gathered for each tatum-frequency candidate f using the DFT as in (10). For comparison, Gouyon et al. [24] used an inter-onset-interval histogram and Maher's two-way mismatch procedure [34] served the same purpose. Their idea was to find a tatum period which best explained the multiple harmonically related peaks in the histogram. Frequencies above 20 Hz can be discarded from $S(f, n)$, since tatum frequencies faster than this are very rare.

It should be noted that the observation $s(\tau, n)$ and its spectrum $S(f, n)$ are zero-phase, meaning that the phases of the pulses at different metrical levels have to be estimated using some other source of information. As will be discussed in Subsection II-D, the phases are estimated based on the states of the comb filters, after the periods have been decided first.

C. Probabilistic model for pulse periods

Period-lengths of the metrical pulses can be estimated independently of their phases and it is reasonable to compute the phase only for the few winning periods.[4] Thus the proposed

[4]For comparison, Laroche [22] estimates periods and phases simultaneously, at the expense of a larger search space. Here three pulse levels are being estimated jointly and estimating periods and phases separately serves the purpose of retaining a moderately-sized search space.

method finds periods first and then the phases (see Fig. 2). Although estimating the phases is not trivial, the search problem is largely completed when the period-lengths have been found.

Musical meter cannot be assumed to remain static over the whole duration of a piece. It has to be estimated causally at successive time instants and there must be some tying between the successive estimates. Also, the dependencies between different metrical pulse levels have to be taken into account. These require prior musical knowledge which is encoded in the probabilistic model to be presented.

For period estimation, a hidden Markov model that describes the simultaneous evolution of four processes is constructed. The observable variable is the vector of instantaneous energies of the resonators, $s(\tau, n)$, denoted $\boldsymbol{s}_n$ in the following. The unobservable processes and the corresponding hidden variables are the tatum period τ_n^{A}, tactus period τ_n^{B}, and measure period τ_n^{C}. As a mnemonic for this notation, recall that the tatum is the temporally atomic (A) pulse level, the tactus pulse is often called "beat" (B), and the musical measure pulse is related to the harmonic (i.e., chord) change rate (C). For convenience, we use $\boldsymbol{q}_n = [j, k, l]$ to denote a "meter state", equivalent to $\tau_n^{\mathrm{A}} = j$, $\tau_n^{\mathrm{B}} = k$, and $\tau_n^{\mathrm{C}} = l$. The hidden state process is a time-homogenous first-order Markov model which has an initial state distribution $P(\boldsymbol{q}_1)$ and transition probabilities $P(\boldsymbol{q}_n|\boldsymbol{q}_{n-1})$. The observable variable is conditioned only on the current state, i.e., we have the state-conditional observation densities $p(\boldsymbol{s}_n|\boldsymbol{q}_n)$.

The joint probability density of a state sequence $Q = (\boldsymbol{q}_1\boldsymbol{q}_2 \ldots \boldsymbol{q}_N)$ and observation sequence $O = (\boldsymbol{s}_1\boldsymbol{s}_1 \ldots \boldsymbol{s}_N)$ can be written as

$$p(Q, O) = P(\boldsymbol{q}_1)p(\boldsymbol{s}_1|\boldsymbol{q}_1) \prod_{n=2}^{N} P(\boldsymbol{q}_n|\boldsymbol{q}_{n-1})p(\boldsymbol{s}_n|\boldsymbol{q}_n), \quad (12)$$

where the term $P(\boldsymbol{q}_n|\boldsymbol{q}_{n-1})$ can be decomposed as

$$P(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}) = P(\tau_n^{\mathrm{B}}|\boldsymbol{q}_{n-1})P(\tau_n^{\mathrm{A}}|\tau_n^{\mathrm{B}}, \boldsymbol{q}_{n-1})P(\tau_n^{\mathrm{C}}|\tau_n^{\mathrm{B}}, \tau_n^{\mathrm{A}}, \boldsymbol{q}_{n-1}). \quad (13)$$

It is musically meaningful to assume that

$$P(\tau_n^{\mathrm{C}}|\tau_n^{\mathrm{B}}, \tau_n^{\mathrm{A}}, \boldsymbol{q}_{n-1}) = P(\tau_n^{\mathrm{C}}|\tau_n^{\mathrm{B}}, \boldsymbol{q}_{n-1}), \quad (14)$$

i.e., given the tactus period, the tatum period does not give additional information regarding the measure period. We further assume that given τ_{n-1}^{B}, the other two hidden variables at time $n-1$ give no additional information regarding τ_n^{B}. For the tatum and measure periods τ_n^i, $i \in \{\mathrm{A}, \mathrm{C}\}$, we assume that given τ_{n-1}^i and τ_n^{B}, the other two hidden variables at time $n-1$ give no additional information regarding τ_n^i. It follows that (13) can be written as

$$P(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}) = P(\tau_n^{\mathrm{B}}|\tau_{n-1}^{\mathrm{B}})P(\tau_n^{\mathrm{A}}|\tau_n^{\mathrm{B}}, \tau_{n-1}^{\mathrm{A}})P(\tau_n^{\mathrm{C}}|\tau_n^{\mathrm{B}}, \tau_{n-1}^{\mathrm{C}}). \quad (15)$$

Using the same assumptions, $P(\boldsymbol{q}_1)$ is decomposed and simplified as

$$P(\boldsymbol{q}_1) = P(\tau_1^{\mathrm{B}})P(\tau_1^{\mathrm{A}}|\tau_1^{\mathrm{B}})P(\tau_1^{\mathrm{C}}|\tau_1^{\mathrm{B}}). \quad (16)$$

The described modeling assumptions lead to a structure which is represented as a directed acyclic graph in Figure 5.

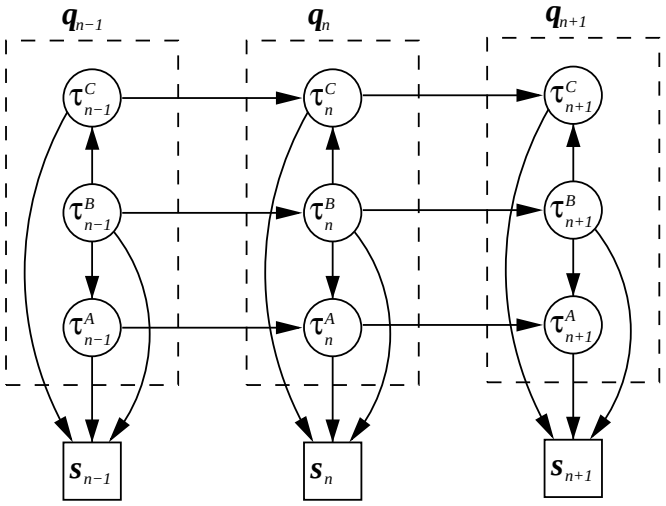

Fig. 5. Hidden markov model for the temporal evolution of the tatum, beat, and measure pulse periods.

The arrows in the graph represent conditional dependencies between the variables. The circles denote hidden variables and the observed variable is marked with boxes. The tactus pulse has a central role in meter perception and it is not by chance that the other two variables are drawn to depend on it [1, pp.73–74]. The assumption in (14) is not valid if the variables are permuted.

1) Estimation of the state-conditional observation likelihoods: The remaining problem is to find reasonable estimates for the model parameters, i.e., for the probabilities that appear in (12)–(16). In the following, we ignore the time indices for a while for simplicity. The state-conditional observation likelihoods $p(\boldsymbol{s}|\boldsymbol{q})$ are estimated from a database of musical recordings where the musical meter has been hand-labeled (see Sec. III). However, the data is very limited in size compared to the number of parameters to be estimated. Estimation of the state densities for each different $\boldsymbol{q} = [j, k, l]$ is impossible since each of the three discrete hidden variables can take on several hundreds of different values. By making a series of assumptions we arrive at the following approximation for $p(\boldsymbol{s}|\boldsymbol{q})$:

$$p(\boldsymbol{s}|\boldsymbol{q} = [j, k, l]) \propto s(k)s(l)S(1/j). \quad (17)$$

where $s(\tau)$ and $S(f)$ are as defined in (9)–(10), omitting the time indices. Appendix I presents the derivation of (17) and the underlying assumptions in detail. An intuitive rationale of (17) is that a truly existing tactus or measure pulse appears as a peak in $s(\tau)$ at the lag that corresponds to the pulse period. Analogously, the tatum period appears as a peak in $S(f)$ at the frequency that corresponds to the inverse of the period. The product of these three values correlates approximately linearly with the likelihood of the observation given the meter.

2) Estimation of the transition and initial probabilities: In (15), the term $P(\tau_n^{\mathrm{A}}|\tau_n^{\mathrm{B}}, \tau_{n-1}^{\mathrm{A}})$ can be decomposed as

$$P(\tau_n^{\mathrm{A}}|\tau_n^{\mathrm{B}}, \tau_{n-1}^{\mathrm{A}}) = P(\tau_n^{\mathrm{A}}|\tau_{n-1}^{\mathrm{A}}) \frac{P(\tau_n^{\mathrm{A}}, \tau_n^{\mathrm{B}}|\tau_{n-1}^{\mathrm{A}})}{P(\tau_n^{\mathrm{A}}|\tau_{n-1}^{\mathrm{A}})P(\tau_n^{\mathrm{B}}|\tau_{n-1}^{\mathrm{A}})}, \quad (18)$$

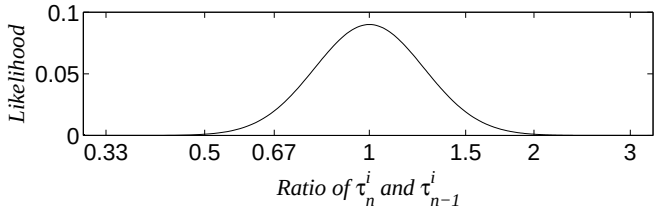

Fig. 6. The likelihood function $f(\tau_n^i/\tau_{n-1}^i)$ which describes the tendency that the periods are slowly-varying.

where the first factor represents transition probabilities between successive period estimates and the second term represents the relation dependencies of simultaneous periods, τ_n^{A} and τ_n^{B}, independent of their actual frequencies of occurrence (in practice τ_n^{B} tends to be integer multiple of τ_n^{A}). Similarly,

$$P(\tau_n^{\mathrm{C}}|\tau_n^{\mathrm{B}},\tau_{n-1}^{\mathrm{C}}) = P(\tau_n^{\mathrm{C}}|\tau_{n-1}^{\mathrm{C}})\frac{P(\tau_n^{\mathrm{C}},\tau_n^{\mathrm{B}}|\tau_{n-1}^{\mathrm{C}})}{P(\tau_n^{\mathrm{C}}|\tau_{n-1}^{\mathrm{C}})P(\tau_n^{\mathrm{B}}|\tau_{n-1}^{\mathrm{C}})}, \tag{19}$$

The transition probabilities $P(\tau_n^i|\tau_{n-1}^i)$, $i \in \{\mathrm{A},\mathrm{B},\mathrm{C}\}$ between successive period estimates are obtained as follows. Again, the number of possible transitions is too large for any reasonable estimates to be obtained by counting occurrences. The transition probability is modeled as a product of the prior probability for a certain period, $P(\tau_1^i)$, and a term $f(\tau_n^i/\tau_{n-1}^i)$ which describes the tendency that the periods are slowly-varying:

$$P(\tau_n^i|\tau_{n-1}^i) = P(\tau_1^i)\frac{P(\tau_n^i,\tau_{n-1}^i)}{P(\tau_n^i)P(\tau_{n-1}^i)} \approx P(\tau_1^i) f\left(\frac{\tau_n^i}{\tau_{n-1}^i}\right), \tag{20}$$

where $i \in \{\mathrm{A},\mathrm{B},\mathrm{C}\}$. The function f,

$$f\left(\frac{\tau_n^i}{\tau_{n-1}^i}\right) = \frac{1}{\sigma_1\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma_1^2}\left(\ln\left(\frac{\tau_n^i}{\tau_{n-1}^i}\right)\right)^2\right], \tag{21}$$

implements a normal distribution as a function of the logarithm of the ratio of successive period values. It follows that the likelihood of large changes in period is higher for long periods, and that period doubling and halving are equally probable. The parameter $\sigma_1 = 0.2$ was found by monitoring the performance of the system in simulations. The distribution (21) is illustrated in Fig. 6.[5]

Prior probabilities for tactus period lengths, $P(\tau^{\mathrm{B}})$, have been measured from actual data by several authors [12], [35], [36]. As suggested by Parncutt [12], we apply the two-parameter lognormal distribution

$$p(\tau^i) = \frac{1}{\tau^i\sigma^i\sqrt{2\pi}}\exp\left[-\frac{1}{2(\sigma^i)^2}\left(\ln\left(\frac{\tau^i}{m^i}\right)\right)^2\right], \tag{22}$$

where m^i and σ^i are the scale and shape parameters, respectively. For the tactus period, the values $m^{\mathrm{B}} = 0.55$ and $\sigma^{\mathrm{B}} = 0.28$ were estimated by counting the occurrences of different period lengths in our hand-labeled database (see Sec. III) and by fitting the lognormal distribution to the histogram data. The parameters depend somewhat on genre (see [35], [36]) but since the genre is generally not known, common parameter values are used here. Figure 7 shows the period-length histograms and the corresponding lognormal distributions for the tactus, measure, and tatum periods. The scale and shape parameters for the tatum and measure periods are $m^{\mathrm{A}} = 0.18$, $\sigma^{\mathrm{A}} = 0.39$, $m^{\mathrm{C}} = 2.1$, and $\sigma^{\mathrm{C}} = 0.26$, respectively. These were estimated from the hand-labeled data in the same way.

[5]For comparison, Laroche uses a cost function where tempo changes exceeding a certain threshold are assigned a fixed cost and smaller tempo changes cause no cost at all [22].

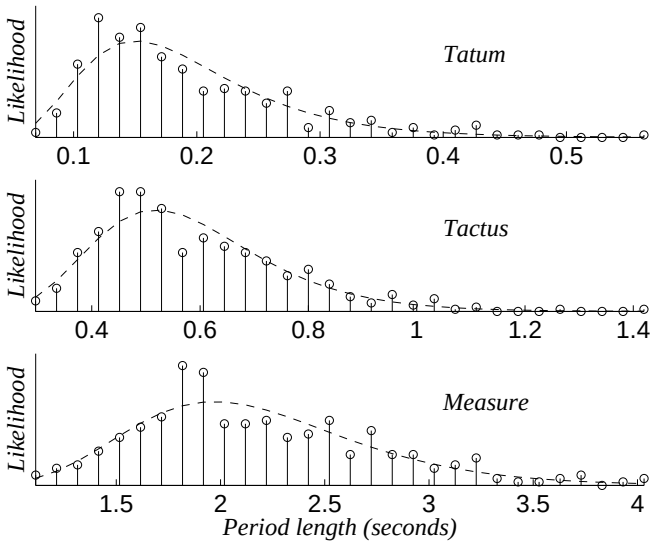

Fig. 7. Period-length histograms and the corresponding lognormal distributions for tatum, tactus, and measure pulses.

The relation dependencies of simultaneous periods are modeled as follows. We model the latter terms in (18)–(19) as

$$\frac{P(\tau_n^{\mathrm{A}},\tau_n^{\mathrm{B}}|\tau_{n-1}^{\mathrm{A}})}{P(\tau_n^{\mathrm{A}}|\tau_{n-1}^{\mathrm{A}})P(\tau_n^{\mathrm{B}}|\tau_{n-1}^{\mathrm{A}})} \approx g\left(\frac{\tau^{\mathrm{B}}}{\tau^{\mathrm{A}}}\right), \tag{23}$$

$$\frac{P(\tau_n^{\mathrm{C}},\tau_n^{\mathrm{B}}|\tau_{n-1}^{\mathrm{C}})}{P(\tau_n^{\mathrm{C}}|\tau_{n-1}^{\mathrm{C}})P(\tau_n^{\mathrm{B}}|\tau_{n-1}^{\mathrm{C}})} \approx g\left(\frac{\tau^{\mathrm{C}}}{\tau^{\mathrm{B}}}\right), \tag{24}$$

where $g(x)$ is a Gaussian mixture density of the form

$$g(x) = \sum_{l=1}^{9} w_l N(x; l, \sigma_2), \tag{25}$$

where w_l are the component weights and sum to unity, l are the component means, and $\sigma_2 = 0.3$ is the common variance. The function models the relation dependencies of simultaneous periods, independent of their actual frequencies of occurrence. The exact weight values are not critical, but are designed to realize a tendency towards binary or ternary integer relationships between concurrent pulses. For example, it happens quite often that one tactus period consists of two, four, or six tatum periods, but multiples five and seven are much less likely in music and thus have lower weights. The distribution is shown in Fig. 8. The Gaussian mixture model was employed to allow some deviation from strictly integral ratios. In theory, the period-lengths should be precisely in integral ratios but, in practice, there are inaccuracies since the period candidates are chosen from discrete vectors $\boldsymbol{s}_n$ and $\boldsymbol{S}_n$.

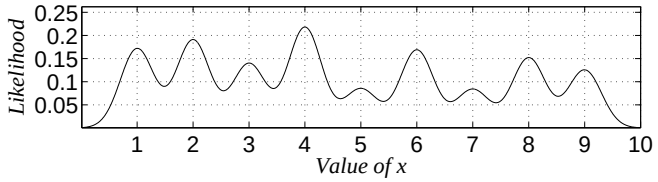

Fig. 8. Distribution $g(x)$ which models the relation dependencies of simultaneous periods (see (25)).

These inaccuracies are conveniently handled by choosing an appropriate value for σ_2 in the above model. The weights w_l were obtained by first assigning them values according to a musical intuition. Then the dynamic range of the weights was found by raising them to a common power which was varied between 0.1 and 10. The value which performed best in small-scale simulations was selected. Finally, small adjustments to the values were made.

It should be noted that here the model parameters were specified in part by hand, considering one probability distribution at a time. It seems possible to devise an algorithm that would learn the model parameters jointly by Bayesian optimization, that is, by maximizing the posterior probability of training data given the prior distributions. However, even after all the described modeling assumptions and simplifications, deriving an expectation-maximization algorithm [37] for the described model, for example, is not easy and such an algorithm does not exist at the present time.

3) Finding the optimal sequence of period estimates: Now we must obtain an estimate for the unobserved state variables given the observed resonator energies and the model parameters. We do this by finding the most likely sequence of state variables $Q = (\boldsymbol{q}_1 \boldsymbol{q}_2 \dots \boldsymbol{q}_N)$ given the observed data $O = (\boldsymbol{s}_1 \boldsymbol{s}_1 \dots \boldsymbol{s}_N)$. This can be straighforwardly computed using the Viterbi algorithm widely applied in speech recognition [38]. Thus, we seek the sequence of period estimates,

$$\hat{Q} = \arg\max_{Q} \left(p(Q, O) \right), \tag{26}$$

where $p(Q, O)$ denotes the joint probability density of the hidden and observed variables (see (12)).

In a causal model, the meter estimate $\boldsymbol{q}_n$ at time n is determined according to the end-state of the best partial path at that point in time. A noncausal estimate after seeing a complete sequence of observations can be computed using backward decoding.

Evaluating all the possible path candidates would be computationally very demanding. Therefore, we apply a suboptimal beam-search strategy and evaluate only a predefined number of the most promising path candidates at each time instant. The selection of the most promising candidates is made using a greedy selection strategy. Once in a second, we select K best candidates independently for the tatum, tactus, and measure periods. The number of candidates $K = 5$ was found to be safe and was used in simulations. The selection is made by maximizing $p(\tau_n^i) p(\boldsymbol{s}_n | \tau_n^i)$ for $i \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$. The probabilities in (23)–(24) could be included to ensure that the selected candidates are consistent with each other, but in practice this is unnecessary. After selecting the best candidates for each, we need only to compute the observation likelihoods for $K^3 = 125$ meter candidates, i.e., for the different combinations of the tatum, tactus, and measure periods. This is done according to (17) and the results are stored into a data vector. The transition probabilities are computed using (15) and stored into a 125-by-125 matrix. These data structures are then used in the Viterbi algorithm.

D. Phase estimation

The phases of the three pulses are estimated at successive time instants, after the periods have been decided at these points. We use $\hat{\tau}_n^i$, $i \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ to refer to the estimated periods of the tatum, tactus, and measure pulses at time n, respectively. The corresponding phases of the three pulses, φ_n^i, are expressed as "temporal anchors", i.e., time values when the nearest beat unit occurs with respect to the beginning of a piece. The periods and phases, τ_n^i and φ_n^i, completely define the meter at time n.

In principle, the phase of the measure pulse, φ_n^{C}, determines the phases of all the three levels. This is because in a well-formed meter each measure-level beat must coincide with a beat at all the lower metrical levels. However, determining the phase of the measure pulse is difficult and turned out to require rhythmic pattern matching techniques, whereas tactus phase estimation is more straightforward and robust. We therefore propose a model where the tactus and measure phases are estimated separately using two parallel models. For the tatum pulse, phase estimation is not needed but the tactus phase can be used.

Scheirer proposed using the state vectors of comb filters to determine the phase of the tactus pulse [20]. This is equivalent to using the latest τ outputs of a resonator with delay τ. We have resonators at several channels c and, consequently, an output matrix $r_c(\tau, j)$ where $c = 1, 2, \dots, c_0$ is the channel index and the phase index j takes on values between $n - \tau + 1$ and n when estimation is taking place at time n. For convenience, we use R_n^i to denote the output matrix $r_c(\hat{\tau}_n^i, j)$ of a found pulse period $\hat{\tau}_n^i$ and the notation $(R_n^i)_{c,j}$ to refer to the individual elements of R_n^i. The matrix R_n^i acts as the observation for phase estimation at time n.

Figure 9 shows an example of the observation matrix R_n^{B} when tactus phase estimation is taking place 20 seconds after the beginning of a piece. The four signals at different channels are the outputs of the comb filter which corresponds to the estimated tactus period $\hat{\tau}_n^{\mathrm{B}} = 0.51$ seconds. The output matrix R_n^{B} contains the latest 0.51 seconds of the output signals, as indicated with the rectangle. The correct phase φ_n^{B} is marked with a dashed line.

Two separate hidden Markov models are evaluated in parallel, one for the tactus phase and another for the measure phase. No joint estimation is attempted. The two models are very similar and differ only in how the state-conditional observation densities are defined. In both models, the observable variable is the output matrix R_n^i of the resonator $\hat{\tau}_n^i$ which corresponds to the found pulse period. The hidden variable is the phase of the pulse, φ_n^i, taking on values between $n - \hat{\tau}_n^i + 1$ and n. The

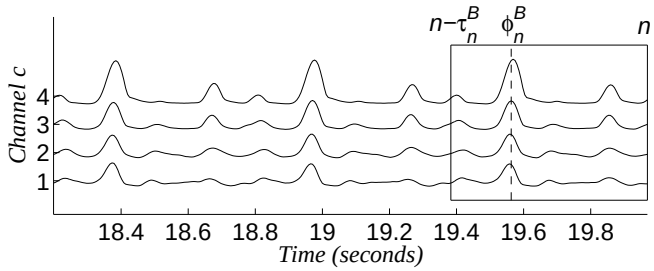

Fig. 9. The rectangle indicates the observation matrix R_n^{B} for tactus phase estimation at time n (here period τ_n^{B} is 0.51 s.). Dashed line shows the correct phase in this case.

hidden state process is a time-homogenous first-order Markov model which has an initial state distribution $P(\varphi_1)$ and transition probabilities $P(\varphi_n|\varphi_{n-1})$. The observable variable is conditional only on the current state, thus we have the state-conditional observation densities $p(R_n^i|\varphi_n^i)$.

Again, the remaining problem is to find reasonable estimates for the model parameters. State-conditional observation likelihoods $p(R_n^{\mathrm{B}}|\varphi_n^{\mathrm{B}})$ for the tactus pulse are approximated as

$$p(R_n^{\mathrm{B}}|\varphi_n^{\mathrm{B}}=j) \propto \sum_{c=1}^{c_0}(c_0-c+2)(R_n^{\mathrm{B}})_{c,j}, \qquad (27)$$

where $c=1$ corresponds to the lowest-frequency channel. That is, the likelihood is proportional to a weighted sum of the resonator outputs across the channels. Across-band summing is intuitively meaningful and earlier used in [20], [30]. Emphasizing the low frequencies is motivated by the "stable bass" rule as stated in [1], and improved the robustness of phase estimation in simulations. The exact weight values are not critical.

For the purpose of estimating the phase of the measure pulse, a formula for the state-conditional observation likelihoods analogous to that in (27) is derived, but so that different channels are weighted and delayed in a more complex manner. It turned out that rhythmic pattern matching of some kind is necessary to analyze music at this time scale and to estimate the measure phase φ_n^{C} based on the output matrix R_n^{C}. That is, no simple formula such as (27) exists. The drawback of this is that rhythmic pattern matching is more genre-specific than for example the stable bass rule which appears to be quite universal. In the case that the system would have access to the pitch content of an incoming piece, the points of harmonic change might serve as cues for estimating the measure phase in a more straightforward manner. However, this remains to be proved. Estimation of the higher-level metrical pulses in audio data has been earlier attempted by Goto and Muraoka who resorted to pattern matching [18] or to straightforward chord change detection [19]. The method presented in the following is the most reliable that we found.

First, a vector $h_n(l)$ is constructed as

$$h_n(l) = \sum_{c=1}^{c_0}\sum_{k=0}^{3}\eta_{c,k}(R_n^{\mathrm{C}})_{c,j(k,l,n)}, \qquad (28)$$

where

$$l = 0, 1, \ldots, \hat{\tau}_n^{\mathrm{C}} - 1, \qquad (29)$$

$$j(k,l,n) = n - \hat{\tau}_n^{\mathrm{C}} + 1 + \left(\left(l + \frac{k\hat{\tau}_n^{\mathrm{C}}}{4}\right) \bmod \hat{\tau}_n^{\mathrm{C}}\right), \qquad (30)$$

and $(x \bmod y)$ denotes modulus after division. The scalars $\eta_{c,k}$ are weights for the resonator outputs at channels c and with delays k. The weights $\eta_{c,k}$ are used to encode a typical pattern of energy fluctuations within one measure period, so that the maximum of $h_n(l)$ indicates the measure phase. The delay k is expressed in quarter-measure units so that k corresponds to the delay $k\hat{\tau}_n^{\mathrm{C}}/4$. For example, a simple pattern consisting of two events, a low-frequency event (at channel $c=1$) in the beginning of a measure ($k=0$) and a loud event in the middle of the measure ($k=2$), could be represented by defining the weights $\eta_{1,0}=3$ (low), $\eta_{c,2}=1$ for all c (loud), and $\eta_{c,k}=0$ otherwise.

Two rhythmic patterns were found that generalized quite well over our database. The weight matrices $\eta_{c,k}^{(1)}$ and $\eta_{c,k}^{(2)}$ of these patterns are given in Appendix II and lead to the corresponding $h_n^{(1)}(l)$ and $h_n^{(2)}(l)$. The patterns were found by trial and error, trying out various arrangements of simple atomic events and monitoring the behaviour of $h_n(l)$ against manually annotated phase values. Both of the two patterns can be characterized as a pendulous motion between a low-frequency event and a high-intensity event. The first pattern can be summarized as "low, loud, –, loud", and the second as "low, –, loud, –". The two patterns are combined into a single vector to perform phase estimation according to whichever pattern matches better to the data

$$h_n^{(1,2)}(l) = \max\left(h_n^{(1)}(l), h_n^{(2)}(l)\right). \qquad (31)$$

The state-conditional observation likelihoods are then defined as

$$p(R_n^{\mathrm{C}}|\varphi_n^{\mathrm{C}}=j) \propto h_n^{(1,2)}(j-(n-\hat{\tau}_n^{\mathrm{C}}+1)). \qquad (32)$$

Obviously, the two patterns imply a *binary time signature*: they assume that one measure period consists of two or four tactus periods. Analysis results for ternary meters will be separately discussed in Sec. III-C.

Other pattern-matching approaches were evaluated, too. In particular, we attempted to sample R_n^{C} at the times of the tactus beats and to train a statistical classifier to choose the beat which corresponds to the measure beat (see [36] for further elaboration on this idea). However, the methods were basically equivalent to that described above, yet less straightforward to implement and performed slightly worse.

Transition probabilities $P(\varphi_n^i|\varphi_{n-1}^i)$ between successive phase estimates are modeled as follows. Given two phase estimates (i.e., beat occurrence times), the conditional probability which ties the successive estimates is assumed to be normally distributed as a function of a *prediction error* e which measures the deviation of φ_n^i from the predicted next beat occurence time given the previous beat time φ_{n-1}^i and the period $\hat{\tau}_n^i$:

$$P(\varphi_n^i|\varphi_{n-1}^i) = \frac{1}{\sigma_3\sqrt{2\pi}}\exp\left(-\frac{e^2}{2\sigma_3^2}\right), \qquad (33)$$

where

$$e = \frac{1}{\hat{\tau}_n^i}\left\{\left[\left(\left|\varphi_n^i - \varphi_{n-1}^i\right| + \frac{\hat{\tau}_n^i}{2}\right) \bmod \hat{\tau}_n^i\right] - \frac{\hat{\tau}_n^i}{2}\right\}, \qquad (34)$$

and $\sigma_3 = 0.1$ is common for $i \in \{\mathrm{B}, \mathrm{C}\}$. In (34), it should be noted that any integer number of periods $\hat{\tau}_n^i$ may elapse between φ_{n-1}^i and φ_n^i. Since estimates are produced quite frequently compared to the pulse rates, in many cases $\varphi_n^i = \varphi_{n-1}^i$. The initial state distributions $P(\varphi_1^i)$ are assumed to be uniform.

Using (27), (32), and (33), causal and noncausal computation of phase is performed using the Viterbi algorithm as described in Sec. II-C. Fifteen phase candidates for both the winning tactus and the winning measure period are generated once in a second. The candidates are selected in a greedy manner by picking local maxima in $p(R_n^i|\varphi_n^i = j)$. The corresponding probability values are stored into a vector and transition probabilities between successive estimates are computed using (33).

E. Sound onset detection and extrametrical events

Detecting the beginnings of discrete acoustic events one-by-one has many uses. It is often of interest whether an event occurs at a metrical beat or not, and what is the exact timing of an event with respect to its ideal metrical position. Also, in some musical pieces there are extrametrical events, such as *triplets*, where an entity of e.g. four tatum periods is exceptionally divided into three parts, or *grace notes* which are pitched events that occur shortly before a metrically stable event.

In this paper, we used an onset detector as a front-end to one of the reference systems (designed for MIDI input) to enable it to process acoustic signals. Rather robust onset detection is achieved by using an *overall accent signal* $v(n)$ which is computed by setting $m_0 = b_0$ in (4). Local maxima in $v(n)$ represent onset candidates and the value of $v(n)$ at these points reflects the likelihood that a discrete event occurred. A simple peak-picking algorithm with a fixed threshold level can then be used to distinguish genuine onsets from the changes and modulations that take place during the ringing of a sound. Automatic adaptation of the threshold would presumably further improve the detection accuracy.

III. Results

This section looks at the performance of the proposed method in simulations and compares the results with two reference systems. Also, the importance of different processing elements will be validated.

A. Experimental setup

Table I shows the statistics of the database[6] that was used to evaluate the accuracy of the proposed meter analysis method and the two reference methods. Musical pieces were collected from CD recordings, downsampled to a single channel, and stored to a hard disc using 44.1 kHz sampling rate and 16 bit resolution. The database was created for the purpose of musical signal classification in general and the balance between genres is according to an informal estimate of what people listen to.

[6]Details of the database can be found on-line at URL http://www.cs.tut.fi/~klap/iiro/meter.

TABLE I
Statistics of the evaluation database.

	# Pieces with annotated pulses		
Genre	Tatum	Tactus	Measure
Classical	69	84	0
Electronic / dance	47	66	62
Hip hop / rap	22	37	36
Jazz / blues	70	94	71
Rock / pop	114	124	101
Soul / RnB / funk	42	54	46
Unclassified	12	15	4
Total	376	474	320

The metrical pulses were manually annotated for approximately one-minute long excerpts which were selected to represent each piece. Tactus and measure-pulse annotations were made by a musician who tapped along with the pieces. The tapping signal was recorded and the tapped beat times were then detected semiautomatically using signal level thresholding. The tactus pulse could be annotated for 474 of a total of 505 pieces. The measure pulse could be reliably marked by listening for 320 pieces. In particular, annotation of the measure pulse was not attempted for classical music without the musical scores. Tatum pulse was annotated by the first author by listening to the pieces together with the annotated tactus pulse and by determining the integer ratio between the tactus and the tatum period lengths. The integer ratio was then used to interpolate the tatum beats between the tapped tactus beats.

Evaluating a meter analysis system is not trivial. The issue has been addressed in depth by Goto and Muraoka in [39]. As suggested by them, we use the longest *continuous* correctly analyzed segment as a basis for measuring the performance. This means that one inaccuracy in the middle of a piece leads to 50 % performance. The longest continuous sequence of correct pulse estimates in each piece is sought and compared to the length of the segment which was given to be analyzed. The ratio of these two lengths determines the performance rate for one piece and these are then averaged over all pieces. However, prior to the meter analysis, all the algorithms under consideration were given a four-second "build-up period" in order to make it theoretically possible to estimate the correct period already from the beginning of the evaluation segment. Also, it was taken care that none of the input material involved tempo discontinuities. More specifically, the interval between two tapped reference beat times (pulse period) does not change more than 40 % at a time, between two successive beats. Other tempo fluctuations were naturally allowed.

A correct period estimate is defined to deviate less than 17.5 % from the annotated reference and a correct phase to deviate from an annotated beat time less than 0.175 times the annotated period length. This precision requirement has been suggested in [39] and was found perfectly appropriate here since inaccuracies in the manually tapped beat times allow meaningful comparison of only up to that precision. However, for the measure pulse, the period and phase requirements were tightened to 10 % and 0.1, respectively, because the measure-

period lengths are large and allow the creation of a more accurate reference signal. For the tatum pulse, tactus phase is used and thus the phase is correct always when the tactus phase is correct, and only the period has to be considered separately.

Performance rates are given for three different criteria [39]:

- "Correct": A pulse estimate at time n is accepted if both its period and phase are correct.
- "Accept d/h": Consistent period doubling or halving is accepted. More exactly, a pulse estimate is accepted if its phase is correct, the period matches either 0.5, 1.0, or 2.0 times the annotated reference, and the factor does not change within the continuous sequence. Correct meter analysis is taking place but a wrong metrical level is chosen to be e.g. the tactus pulse.
- "Period correct": A pulse estimate is accepted if its period is correct. Phase is ignored. For the tactus pulse, this can be interpreted as the *tempo estimation* accuracy.

Which is the single best number to characterize the performance of a pulse estimator? This was investigated by auralizing meter analysis results.[7] It was observed that temporal continuity of correct meter estimates is indeed very important aurally (see also [1, pp.74,104]). Secondly, phase errors are very disturbing. Third, period doubling or halving is not very disturbing; tapping *consistently* twice too fast or slow does not matter much and selecting the correct metrical level is in some cases ambiguous even for a human listener [12]. In summary, it appears that the "accept d/h" criterion gives a single best number to characterize the performance of a system.

B. Reference systems

To put the results in perspective, two reference methods are used as a baseline in simulations. This is essential because the principle of using a continuous sequence of correct estimates for evaluation gives a somewhat pessimistic picture of the absolute performance.

The methods of Scheirer [20] and Dixon [16] are very different, but both systems represent the state-of-the-art in tactus pulse estimation and their source codes are publicly available. Here, the used implementations and parameter values were those of the original authors. However, for Scheirer's method, some parameter tuning was made which slightly improved the results. Dixon developed his system primarily for MIDI-input, and provided only a simple front-end for analyzing acoustic signals. Therefore, a third system denoted "O+Dixon" was developed where an independent onset detector (described in Sec. II-E) was used prior to Dixon's tactus analysis. Systematic phase errors were compensated for.

C. Experimental results

In Table II, the tactus tracking performance of the proposed causal and noncausal algorithms is compared with those of the two reference methods. As the first observation, it was noticed that the reference methods did not maintain the temporal continuity of acceptable estimates. For this reason, the performance rates are also given as percentages of individual acceptable estimates (right half of Table II). Dixon's method has difficulties in choosing the correct metrical level for tactus, but performs well according to the "accept d/h" criterion when equipped with the new onset detector. The proposed method outperforms the previous systems in both accuracy and temporal stability.

[7] Samples are available at URL http://www.cs.tut.fi/∼klap/iiro/meter.

TABLE II
TACTUS ANALYSIS PERFORMANCE (%) OF DIFFERENT METHODS.

	Continuity required			Individual estimates		
Method	Correct	Accept d/h	Period c.	Correct	Accept d/h	Period c.
Causal	57	**68**	74	63	78	76
Noncausal	59	**73**	74	64	80	75
Scheirer [20]	27	**31**	30	48	69	57
Dixon [16]	7	**26**	10	15	53	25
O+Dixon	12	**39**	15	22	63	30

TABLE III
METER ANALYSIS PERFORMANCE OF THE PROPOSED METHOD.

		Continuity required			Individual estimates		
Method	Pulse	Correct	Accept d/h	Period	Correct	Accept d/h	Period
Causal	Tatum	44	**57**	62	51	**72**	65
	Tactus	57	**68**	74	63	78	76
	Measure	42	**48**	**78**	43	51	81
Non-causal	Tatum	45	**63**	62	52	**74**	65
	Tactus	59	**73**	74	64	80	75
	Measure	46	**54**	**79**	47	55	81

Table III shows the meter analysis performance of the proposed causal and noncausal algorithms. As for human listeners, meter analysis seems to be easiest at the tactus pulse level. For the measure pulse, period estimation can be done robustly but estimating the phase is difficult. A reason for this is that in a large part of the material, a drum pattern recurs twice within one measure period and the system has difficulties in choosing which one is the first. In the case that π-phase errors (each beat is displaced by a half-period) would be accepted, the performance rate would be essentially the same as for the tactus pulse. However, π-phase errors *are* disturbing and should not be accepted.

For the tatum pulse, in turn, deciding the period is difficult. This is because the temporally atomic pulse rate typically comes up only occasionally, making temporally stable analysis hard to attain. The method often has to halve its period hypothesis when the first rapid event sequence occurs. This appears in the performance rates so that the method is not able to produce a consistent tatum period over time but alternates between e.g. the reference and double the reference. This degrades the temporally continuous rate, although the "accept d/h" rate is very good for individual estimates. The produced errors are not very disturbing when listening to the results.

As mentioned in Sec. II-D, the phase analysis of the measure pulse using rhythmic patterns assumes a binary time signature. Nine percent of the pieces in our database have a ternary (3/4) meter but, unfortunately, most of these represent the classical genre where the measure pulse was not annotated. Among the

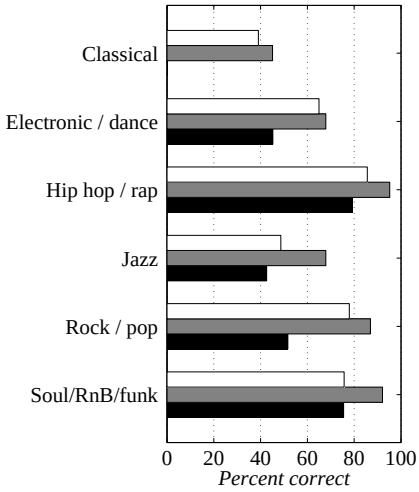

Fig. 10. Performance of the proposed causal system within different musical genres. The "accept d/h" (continuity required) percentages are shown for the tatum (white), tactus (gray), and measure pulses (black).

TABLE IV
METER ANALYSIS PERFORMANCE (%) FOR DIFFERENT SYSTEM CONFIGURATIONS.

	Continuity required, accept d/h			Individual estimates, accept d/h		
Method	Tatum	Tactus	Measure	Tatum	Tactus	Measure
0. Baseline	**63**	**73**	**54**	74	80	55
1. No joint estim.	**58**	**68**	**49**	71	75	50
2. No temporal proc.	**45**	**54**	**31**	72	77	50
3. Neither of the two	**41**	**50**	**25**	70	72	44

other genres, there were *only five* pieces with ternary meter. For these, the measure-level analysis was approximately twice less accurate than for the rest of the database. For the tactus and tatum, there were 41 and 30 annotated ternary pieces, respectively, and no significant degradation in performance was observed. On the contrary, the ternary pieces were rhythmically easier than the others within the same genre.

Figure 10 shows the "accept d/h" (continuity required) performance rates for the proposed causal system within different musical genres. For classical music, the proposed method is only moderately successful, although e.g. the tactus estimation error rate still outperforms the performance of the reference methods for the whole material (31 % and 26 % for Scheirer's and Dixon's methods, respectively). However, this may suggest that pitch analysis would be needed to analyze the meter of classical music. In jazz music, the complexity of musical rhythms is higher on the average and the task thus harder.

D. Importance of the different parts of the probability model

Table IV shows the performance rates for different system configurations. Different elements of the proposed model were disabled in order to evaluate their importance. In each case, the system was kept otherwise fixed. The baseline method is the noncausal system.

In the first test, the dependencies between the different pulse levels were broken by using a non-informative (flat) distribution for $g(x)$ in (25). This slightly degrades the performance in all cases. In the second test, the dependencies between temporally successive estimates were broken by using a non-informative distribution for the transition probabilities between successive period and phase estimates, $P(\tau_n^i|\tau_{n-1}^i)$ and $P(\varphi_n^i|\varphi_{n-1}^i)$, respectively. This degrades the temporal stability of the estimates considerably and hence collapses the performance rates which use the longest continuous correct segment for evaluation. In the third case, the both types of dependencies were broken. The system still performs moderately, indicating that the initial time-frequency analysis method and the comb-filter resonators provide a high level of robustness.

IV. CONCLUSIONS

A method has been described which can successfully analyze the meter of acoustic musical signals. Musical genres of very diverse types can be processed with a common system configuration and parameter values. For most musical material, relatively low-level acoustic information can be used, without the need to model the higher-level auditory functions such as sound source separation or multipitch analysis.

Similarly to human listeners, computational meter analysis is easiest at the tactus pulse level. For the measure pulse, period estimation can be done equally robustly but estimating the phase is less straightforward. Either rhythmic pattern matching or pitch analysis seems to be needed to analyze music at this time scale. For the tatum pulse, in turn, phase estimation is not difficult at all, but deciding the period is very difficult for both humans and a computational algorithm. This is because the temporally atomic pulse rate typically comes up only occasionally. Thus causal processing is difficult and it is often necessary to halve the tatum hypothesis when the first rapid event sequence occurs.

The critical elements of a meter analysis system appear to be the initial time-frequency analysis part which measures musical accentuation as a function of time and the (often implicit) internal model which represents primitive musical knowledge. The former is needed to provide robustness for diverse instrumentations in classical, rock, or electronic music, for example. The latter is needed to achieve temporally stable meter tracking and to fill in parts where the meter is only faintly implied by the musical surface. A challenge in this part is to develop a model which is generic for jazz and classical music, for example. The proposed model describes sufficiently low-level musical knowledge to generalize over different genres.

APPENDIX I

This appendix presents the derivation and underlying assumptions in the estimation of the state-conditional observation likelihoods $p(\boldsymbol{s}|\boldsymbol{q})$. We first assume that the realizations of τ^{A} are independent of the realizations of τ^{B} and τ^{C}, that is, $P(\tau^{\mathrm{A}} = j|\tau^{\mathrm{B}} = k, \tau^{\mathrm{C}} = l) = P(\tau^{\mathrm{A}} = j)$. This violates the dependencies of our model but significantly simplifies

the computations and makes it possible to obtain reasonable estimates. Using the assumption, we can write

$$P(\boldsymbol{s}|\tau^{\mathrm{A}}=j,\tau^{\mathrm{B}}=k,\tau^{\mathrm{C}}=l) = P(\boldsymbol{s}|\tau^{\mathrm{B}}=k,\tau^{\mathrm{C}}=l)P(\boldsymbol{s}|\tau^{\mathrm{A}}=j)/P(\boldsymbol{s}). \quad (35)$$

Furthermore, tatum information is most clearly visible in the spectrum of the resonator outputs. Thus we use

$$P(\boldsymbol{s}|\tau^{\mathrm{A}}=j) = P(\boldsymbol{S}|\tau^{\mathrm{A}}=j), \quad (36)$$

where $\boldsymbol{S}$ is the spectrum of $\boldsymbol{s}$, according to (10). We further assume the components of $\boldsymbol{s}$ and $\boldsymbol{S}$ to be conditionally independent of each other given the state, and write the nominator of (35) as

$$P(\boldsymbol{s}|\tau^{\mathrm{B}}=k,\tau^{\mathrm{C}}=l)P(\boldsymbol{S}|\tau^{\mathrm{A}}=j) = \prod_{k'=1}^{\tau_{\max}} P(s(k')|\tau^{\mathrm{B}}=k,\tau^{\mathrm{C}}=l) \prod_{j'=1}^{\tau_{\max}} P(S(1/j')|\tau^{\mathrm{A}}=j). \quad (37)$$

We make two more simplifying assumptions. First, we assume that the value of $\boldsymbol{s}$ and $\boldsymbol{S}$ at the lags corresponding to a period actually present in the signal depends only on the particular period, not on other periods. Second, the value at lags where there is no period present in the signal is independent of the true periods τ^{A}, τ^{B}, and τ^{C}, and is dominated by the fact that no period corresponds to that particular lag. Hence, (35) can be written as

$$P(\boldsymbol{s}|\boldsymbol{q}=[j,k,l]) = \frac{1}{P(\boldsymbol{s})} P(s(k)|\tau^{\mathrm{B}}=k) \cdot P(s(l)|\tau^{\mathrm{C}}=l) \prod_{k'\neq k,l} P(s(k')|\tau^{\mathrm{B}},\tau^{\mathrm{C}}\neq k') \cdot P(S(1/j)|\tau^{\mathrm{A}}=j) \prod_{j'\neq j} P(S(1/j')|\tau^{\mathrm{A}}\neq j'), \quad (38)$$

where $P(s(\tau)|\tau^{\mathrm{B}}=\tau)$ denotes the probability of value $s(\tau)$ given that τ is a tactus pulse period and $P(s(\tau)|\tau^{\mathrm{B}}\neq\tau)$ denotes the probability of value $s(\tau)$ given that τ is not a tactus pulse period. These conditional probability distributions (tactus, measure, and tatum each have two distributions) were approximated by discretizing the value range of $s(\tau)$, $s(\tau)\in[0,1]$, and by calculating a histogram of $s(\tau)$ values in the cases that τ is or is not an annotated metrical pulse period.

Then, by defining

$$\beta(\boldsymbol{s}) = \frac{1}{P(\boldsymbol{s})} \prod_{k'=1}^{\tau_{\max}} P(s(k')|\tau^{\mathrm{B}},\tau^{\mathrm{C}}\neq k') \prod_{j'=1}^{\tau_{\max}} P\left(S\left(\frac{1}{j'}\right)|\tau^{\mathrm{A}}\neq j'\right), \quad (39)$$

Equation (38) can be written as

$$P(\boldsymbol{s}|\boldsymbol{q}=[j,k,l]) = \beta(\boldsymbol{s}) \cdot \frac{P(s(k)|\tau^{\mathrm{B}}=k)}{P(s(k)|\tau^{\mathrm{B}},\tau^{\mathrm{C}}\neq k)} \cdot \frac{P(s(l)|\tau^{\mathrm{C}}=l)}{P(s(l)|\tau^{\mathrm{B}},\tau^{\mathrm{C}}\neq l)} \frac{P(S(1/j)|\tau^{\mathrm{A}}=j)}{P(S(1/j)|\tau^{\mathrm{A}}\neq j)}, \quad (40)$$

where the scalar $\beta(\boldsymbol{s})$ is a function of $\boldsymbol{s}$ but does not depend on $\boldsymbol{q}$.

By using the two approximated histograms for the tactus, measure, and tatum pulses, each of the three terms of the form $P(s(\tau)|\tau^i=\tau)/P(s(\tau)|\tau^i\neq\tau)$ in (40) can be represented by a single discrete histogram. These were modeled with first-order polynomials. The first two terms depend linearly on the value $s(\tau)$ and the last term depends linearly on the value $S(1/\tau)$. Thus we can write

$$p(\boldsymbol{s}|\boldsymbol{q}=[j,k,l]) \propto s(k)s(l)S(1/j). \quad (41)$$

The histograms could be more accurately modeled with third-order polynomials, but this did not bring performance advantage over the simple linear model in (41).

APPENDIX II

Numerical values of the matrices used in Sec. II-D:

$$\eta^{(1)}_{c,k} = \begin{bmatrix} 12 & 1.0 & 0 & 5.7 \\ 0 & 2.0 & 0 & 2.0 \\ 0 & 3.0 & 0 & 3.0 \\ 0 & 4.0 & 0 & 4.0 \end{bmatrix}, \quad \eta^{(2)}_{c,k} = \begin{bmatrix} 10 & 0 & 1.4 & 1.3 \\ 0 & 0 & 2.8 & 0.8 \\ 0 & 0 & 4.3 & 1.2 \\ 0 & 0 & 5.8 & 1.5 \end{bmatrix}.$$

Here channel c determines the row and delay k the column. The first row corresponds to the lowest-frequency channel.

REFERENCES

[1] F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, Cambridge, Massachusetts, 1983.

[2] E. F. Clarke. Rhythm and timing in music. In D. Deutsch, editor, *The Psychology of Music*. Academic Press, 1999.

[3] J. Bilmes. Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm. Master's thesis, Massachusetts Institute of Technology, 1993.

[4] F. Gouyon, L. Fabig, and J. Bonada. Rhythmic expressiveness transformations of audio recordings: Swing modifications. In *Proceedings of 6th International Conf. on Digital Audio Effects*, London, UK, 2003.

[5] A. T. Cemgil and B. Kappen. Monte carlo methods for tempo tracking and rhythm quantization. *Journal of Artificial Intelligence Research*, 18:45–81, 2003.

[6] A. P. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Trans. Speech and Audio Processing*, 11(6):804–815, 2003.

[7] J. K. Paulus and A. P. Klapuri. Conventional and periodic N-grams in the transcription of drum sequences. In *Proceedings of IEEE International Conference on Multimedia and Expo*, Baltimore, Maryland, July 2003.

[8] M. Ryynänen and A. P. Klapuri. Modelling of note events for singing transcription. In *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Proc.*, Jeju, Korea, October 2004.

[9] C. S. Lee. The perception of metrical structure: Experimental evidence and a model. In P. Howell, R. West, and Cross I., editors, *Representing musical structure*. Academic Press, London, 1991.

[10] P. Desain and H. Honing. Computational models of beat induction: The rule-based approach. *Journal of New Music Research*, 28(1):29–42, 1999.

[11] D. F. Rosenthal. *Machine rhythm: Computer emulation of human rhythm perception*. PhD thesis, Massachusetts Institute of Technology, 1992.

[12] R. Parncutt. A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, 11(4):409–464, 1994.

[13] J. C. Brown. Determination of the meter of musical scores by autocorrelation. *J. Acoust. Soc. Am.*, 94(4):1953–1957, 1993.

[14] E. W. Large and J. F. Kolen. Resonance and the perception of musical meter. *Connection science*, 6(1):177–208, 1994.

[15] D. Temperley and D. Sleator. Modeling meter and harmony: A preference-rule approach. *Computer Music Journal*, 23(1):10–27, 1999. URL: http://www.link.cs.cmu.edu/music-analysis/.

[16] S. Dixon. Automatic extraction of tempo and beat from expressive performances. *J. New Music Research*, 30(1):39–58, 2001. URL: http://www.ai.univie.ac.at/∼simon/beatroot/.

[17] C. Raphael. Automated rhythm transcription. In *Proceedings of International Symposium on Music Information Retrieval*, pages 99–107, Indiana, October 2001.

[18] M. Goto and Y. Muraoka. Music understanding at the beat level — Real-time beat tracking for audio signals. In *Proceedings of IJCAI-95 Workshop on Computational Auditory Scene Analysis*, pages 68–75, 1995.

[19] M. Goto and Y. Muraoka. Real-time rhythm tracking for drumless audio signals — Chord change detection for musical decisions. In *Proceedings of IJCAI-97 Workshop on Computational Auditory Scene Analysis*, pages 135–144, 1997.

[20] E. D. Scheirer. Tempo and beat analysis of acoustic musical signals. *J. Acoust. Soc. Am.*, 103(1):588–601, 1998. URL: http://sound.media.mit.edu/∼eds/beat/tapping.tar.gz.

[21] W. A. Sethares and T. W. Staley. Meter and periodicity in musical performance. *Journal of New Music Research*, 22(5), 2001.

[22] J. Laroche. Efficient tempo and beat tracking in audio recordings. *J. Audio Eng. Soc.*, 51(4):226–233, April 2003.

[23] S. Hainsworth and M. Macleod. Beat tracking with particle filtering algorithms. In *Proceedings of IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics*, New Paltz, New York, 2003.

[24] F. Gouyon, P. Herrera, and P. Cano. Pulse-dependent analyses of percussive music. In *Proceedings of AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, pages 396–401, Espoo, Finland, June 2002.

[25] A. P. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3089–3092, Phoenix, Arizona, 1999.

[26] D. Moelants and C. Rampazzo. A computer system for the automatic detection of perceptual onsets in a musical signal. In A. Camurri, editor, *KANSEI — The Technology of Emotion*, pages 141–146. AIMI/DIST, Genova, 1997.

[27] M. Davy and S. Godsill. Detection of abrupt spectral changes using support vector machines. An application to audio signal segmentation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1313–1316, Orlando, Florida, May 2002.

[28] S. A. Abdallah and M. D. Plumbley. Probability as metadata: Event detection in music using ICA as a conditional density model. In *Proceedings of 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, Nara, Japan, April 2003.

[29] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler. Complex domain onset detection for musical signals. In *Proceedings of 6th Int. Conf. on Digital Audio Effects*, London, UK, September 2003.

[30] R. Meddis and M. J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *J. Acoust. Soc. Am.*, 89(6):2866–2882, 1991.

[31] B. C. J. Moore, editor. *Hearing. Handbook of Perception and Cognition*. Academic Press, second edition, 1995.

[32] R. Meddis. Simulation of mechanical to neural transduction in the auditory receptor. *J. Acoust. Soc. Am.*, 79(3):702–711, 1986.

[33] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4):1917–1930, 2002.

[34] R. C. Maher and J. W. Beauchamp. Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *J. Acoust. Soc. Am.*, 95(4):2254–2263, 1994.

[35] L. van Noorden and D. Moelants. Resonance in the perception of musical pulse. *Journal of New Music Research*, 28(1):43–66, 1999.

[36] J. Seppänen. Computational models of musical meter recognition. Master's thesis, Tampere University of Technology, Tampere, Finland, 2001.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, 39:1–38, 1977.

[38] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, New Jersey, 1993.

[39] M. Goto and Y. Muraoka. Issues in evaluating beat tracking systems. In *Proceedings of IJCAI-97 Workshop on Issues in AI and Music*, pages 9–16, 1997.