Extracting Predominant Local Pulse Information From Music Recordings

Peter Grosche, *Student Member, IEEE*, and Meinard Müller, *Member, IEEE*

***Abstract*—The extraction of tempo and beat information from music recordings constitutes a challenging task in particular for non-percussive music with soft note onsets and time-varying tempo. In this paper, we introduce a novel mid-level representation that captures musically meaningful local pulse information even for the case of complex music. Our main idea is to derive for each time position a sinusoidal kernel that best explains the local periodic nature of a previously extracted note onset representation. Then we employ an overlap-add technique accumulating all these kernels over time to obtain a single function that reveals the predominant local pulse (PLP). Our concept introduces a high degree of robustness to noise and distortions resulting from weak and blurry onsets. Furthermore, the resulting PLP curve reveals the local pulse information even in the presence of continuous tempo changes and indicates a kind of confidence in the periodicity estimation. As further contribution, we show how our PLP concept can be used as a flexible tool for enhancing tempo estimation and beat tracking. The practical relevance of our approach is demonstrated by extensive experiments based on music recordings of various genres.**

***Index Terms*—Audio feature, beat tracking, mid-level representation, music signal processing, musical pulse, onset detection, tempo estimation.**

I. Introduction

MUSIC signal processing constitutes a difficult field of research because of the complexity and diversity of music. When analyzing audio recordings, one has to account for various musical dimensions such as pitch, harmony, timbre, and rhythm. In this paper, we address the aspects of beat and tempo, which are of fundamental importance for understanding and interacting with music [1]. It is the *beat*, the steady pulse that drives music forward and provides the temporal framework of a piece of music [2]. Intuitively, the beat can be described as a sequence of perceived pulses that are equally spaced in time and corresponds to the pulse a human taps along when listening to the music [3]. The term *tempo* then refers to the rate of the pulse. Because tempo and beat are of fundamental musical importance, the automated extraction of this information from music recordings is a central topic in the field of *music information retrieval*. Most approaches to tempo estimation and beat tracking proceed in two steps. In the first step, positions of note onsets within the music signal are estimated. Here, most approaches capture changes of the signal's energy or spectrum and derive a so-called *novelty curve*. The peaks of such a curve yield good indicators for note onset candidates [4]–[6]. In the second step, the novelty curve is analyzed to detect reoccurring patterns and quasi-periodic pulse trains [7]–[10]. For non-percussive music with soft note onsets, however, the extraction of beat and tempo information becomes a difficult problem. Even more challenging becomes the detection of local periodic patterns in the presence of tempo changes.

As the main contribution of this paper, we introduce a novel approach that allows for a robust extraction of musically meaningful local pulse information even for the case of complex music. Intuitively speaking, our idea is to construct a mid-level representation that explains the local periodic nature of a given (possibly very noisy) onset representation. More precisely, starting with a novelty curve, we determine for each time position a sinusoidal kernel that best captures the local peak structure of the novelty curve. Since these kernels localize well in time, even continuous tempo variations and local changes of the pulse level can be handled. Now, instead of looking at the local kernels individually, our crucial idea is to employ an overlap-add technique by accumulating all local kernels over time. As a result, one obtains a single curve that can be regarded as a local periodicity enhancement of the original novelty curve. Revealing *predominant local pulse* (PLP) information, this curve is referred to as *PLP curve*.

Our PLP concept yields a powerful mid-level representation that can be applied as a flexible tool for various music analysis tasks. In particular, we discuss in detail how the PLP concept can be applied for improving on tempo estimation as well as for validating the local tempo estimates. Furthermore, we show that state-of-the-art beat trackers can be improved when applied to a PLP-enhanced novelty representation. Here, one important feature of our work is that we particularly consider music recordings that reveal changes in tempo, whereas most of the previous tempo estimation and beat tracking approaches assume a (more or less) constant tempo throughout the recording. As it turns out, our PLP concept is capable of capturing continuous tempo changes as implied by ritardando or accelerando. However, as our approach relies on the assumption of a locally quasi-periodic behavior of the signal it reaches its limits in the presence of strong local tempo distortions as found in highly expressive music (e.g. romantic piano music). To demonstrate the practical relevance of our PLP concept, we have conducted extensive experiments based on several music datasets consisting of 688

Manuscript received July 08, 2010; revised October 11, 2010; accepted November 08, 2010. Date of publication December 03, 2010; date of current version June 01, 2011. This work was supported by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sylvain Marchand.

The authors are with the Saarland University and the Max-Planck Institut für Informatik, 66123 Saarbrücken, Germany (e-mail: pgrosche@mpi-inf.mpg.de; meinard@mpi-inf.mpg.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2096216

recordings amounting to more than 36 hours of annotated audio material. The datasets cover various genres including popular music, Jazz music, and classical music.

The remainder of this paper is organized as follows. In Section II, we give a detailed background on onset detection, beat tracking, and tempo estimation, while discussing relevant work. Section III contains the main conceptual contribution of this paper, where we introduce the PLP concept and discuss the main properties of our novel mid-level representation. In Section IV, we review the concept of novelty curves while introducing a variant used in our experiments. The applications to tempo estimation and beat tracking as well as the corresponding experiments are discussed in Section V. Finally, we conclude with Section VI, where we also discuss future work. Parts of this work have been published earlier in [11] and [12].

II. Background

As mentioned in [1], the beat is a sequence of repeating short-duration stimuli perceived as points in time. The beat is a perceptual phenomenon and perceptual beat times do not necessarily coincide with physical beat times [13]. Furthermore, the perception of beats varies between listeners. However, beat positions typically go along with note onsets or percussive events. Therefore, in most tempo and beat tracking approaches the first step consists in locating such events in the given signal—a task often referred to as *onset detection* or *novelty detection*. To determine the physical starting times of the notes occurring in the music recording, the general idea is to capture changes of certain properties of the signal to derive a *novelty curve*. The peaks of this curve indicate candidates for note onsets. Many different methods for computing novelty curves have been proposed; see [4] and [5] for an overview. When playing a note, the onset typically goes along with a sudden increase of the signal's energy. Having a pronounced attack phase, note onset candidates may be determined by locating time positions, where the signal's amplitude envelope starts to increase [4]. Much more challenging, however, is the detection of onsets in the case of non-percussive music, where one has to deal with soft onsets or blurred note transitions. This is often the case for classical music dominated by string instruments. As a consequence, more refined methods have to be used for computing a novelty curve, e.g., by analyzing the signal's spectral content [4], [6], pitch [6], harmony [14], [15], or phase [4], [16]. To handle the variety of different signal types, a combination of novelty curves can improve the detection accuracy [6], [17]. Furthermore, in complex polyphonic mixtures of music, simultaneously occurring events of high intensities lead to masking effects that prevent any observation of an energy increase of a low intensity onset. To circumvent these masking effects, detection functions were proposed that analyze the signal in a bandwise fashion [18] to extract transients occurring in certain frequency regions of the signal. As a side-effect of a sudden energy increase, there appears an accompanying broadband noise burst in the signal's spectrum. This effect is mostly masked by the signal's energy in lower frequency regions but well detectable in the higher frequency regions [19] of the spectrum. Here, logarithmic compression [18] and spectral whitening [20] are techniques for enhancing the high-frequency information. Some of these approaches are employed for computing our novelty curves; see Section IV.

To derive the beat period and the tempo from a novelty curve, one strategy is to explicitly determine note onset positions and then to reveal the structure of these events. For the selection of onset candidates, one typically employs peak picking strategies based on adaptive thresholding [4]. Each pair of note onset positions then defines an inter-onset-interval (IOI). Considering suitable histograms over the occurring IOIs, one may derive hypotheses on the beat period and tempo [21]–[23]. The idea is that IOIs frequently appear at integer multiples and fractions of the beat period. Similarly, one may compute the autocorrelation of the extracted onset times [15] to derive the beat period. The drawback of these approaches is that they rely on an explicit localization of a discrete set of note onsets—a fragile and error-prone step. In particular, in the case of weak and blurry onsets the selection of the relevant peaks of the novelty curve that correspond to true note onsets becomes a difficult or even infeasible problem.

Avoiding the explicit extraction of note onset, the novelty curves can directly be analyzed with respect to reoccurring or quasi-periodic patterns. Here, generally speaking, one can distinguish between three different methods for measuring periodicities. The autocorrelation method allows for detecting periodic self-similarities by comparing a novelty curve with time-shifted copies [8]–[10], [24]. Another widely used method is based on a bank of comb filter resonators, where a novelty curve is compared with templates consisting of equally spaced spikes representing various frequencies [7], [25]. Similarly, one can use a short-time Fourier transform to derive a frequency representation of the novelty curve [8]. Here, the novelty curve is compared with sinusoidal templates representing specific frequencies. Each of the methods reveals periodicities of the underlying novelty curve, from which one can estimate the tempo or beat. The characteristics of the periodicities typically change over time and can be visualized by means of spectrogram-like representations referred to as *tempogram* [26], *rhythmogram* [27], or *beat spectrogram* [28].

More challenging becomes the detection of local periodic patterns in the case that the music recordings reveal significant tempo changes. This often occurs in performances of classical music as a result of ritardandi, accelerandi, fermatas, and so on [22]. Furthermore, the extraction problem is complicated by the fact that the notions of tempo and beat are ill-defined and highly subjective due to the complex hierarchical structure of rhythm [1]. For example, there are various levels that are presumed to contribute to the human perception of tempo and beat. Typically, previous work focuses on determining musical pulses on the *tactus* (the foot tapping rate or beat [3]) level [8]–[10], but only few approaches exist for analyzing the signal on the measure level [15], [25] or finer tatum level [23], [29], [30]. Here, a *tatum* or *temporal atom* refers to the fastest repetition rate of musically meaningful accents occurring in the signal [31]. Various approaches have been suggested that simultaneously analyze different pulse levels [24], [25], [32].

In contrast to previous approaches, our goal is to extract the predominant local periodicity of accents in the music signal, which may be a pulse on the tatum, the tactus, or measure level.

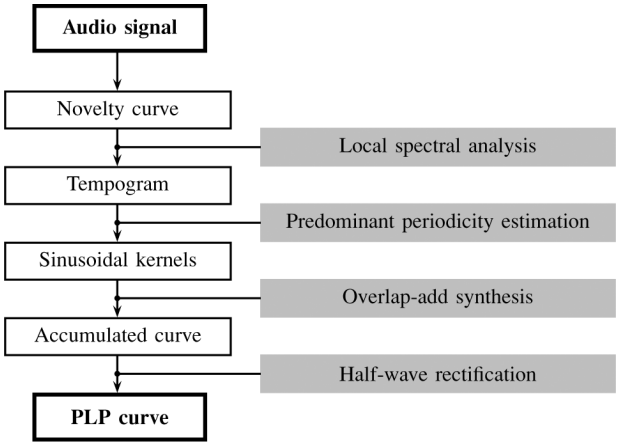

Fig. 1. Flowchart of the steps involved in the PLP computation.

Furthermore, our approach does not assume constant tempo throughout the recording. Actually, our PLP curve exhibits the predominant pulse for each time position, thus making local tempo information explicit.

III. Predominant Local Pulse Estimation

Our approach for computing a PLP curve is based on a general procedure that can be applied to any given novelty curve. In some sense, the resulting PLP mid-level representation can be regarded as a kind of *periodicity enhancement*, which immediately reveals for each time position the predominant pulse hidden in the novelty curve. In this section, we describe in detail our PLP concept. We start with an overview, then elaborate on the mathematical details, and finally discuss general properties of PLP curves. Since the usage of specific novelty curves is not in the focus of this paper, we postpone the description of the computation of such curves to Section IV, where we also describe the novelty curve used in our experiments.

A. Overview

We now give an overview of the steps involved in the PLP computation; see Fig. 1 for a schematic overview and Fig. 2 for an example. The input of our procedure consists of a spike-like novelty curve; see Fig. 2(a). In the first step, we derive a time-pulse representation, referred to as *tempogram*, by performing a local spectral analysis of the novelty curve; see Fig. 2(b). Here, we avoid the explicit determination of note onsets, which generally is an error-prone and fragile step. Then, from the tempogram, we determine for each time position the sinusoidal periodicity kernel that best explains the local periodic nature of the novelty curve in terms of period (frequency) and timing (phase); see Fig. 2(c). Since there may be a number of outliers among these kernels, one usually obtains unstable information when looking at these kernels in a one-by-one fashion. Therefore, as one main idea of this paper, we use an overlap-add technique by accumulating all these kernels over time to obtain a single curve; see Fig. 2(d). In a final step, we apply a half-wave rectification (only considering the positive part of the curve) to obtain the mid-level representation we refer to as predominant local pulse (PLP) curve; see Fig. 2(e). As it turns out, such PLP curves are robust to outliers and reveal musically meaningful periodicity information even when starting with relatively poor onset information.

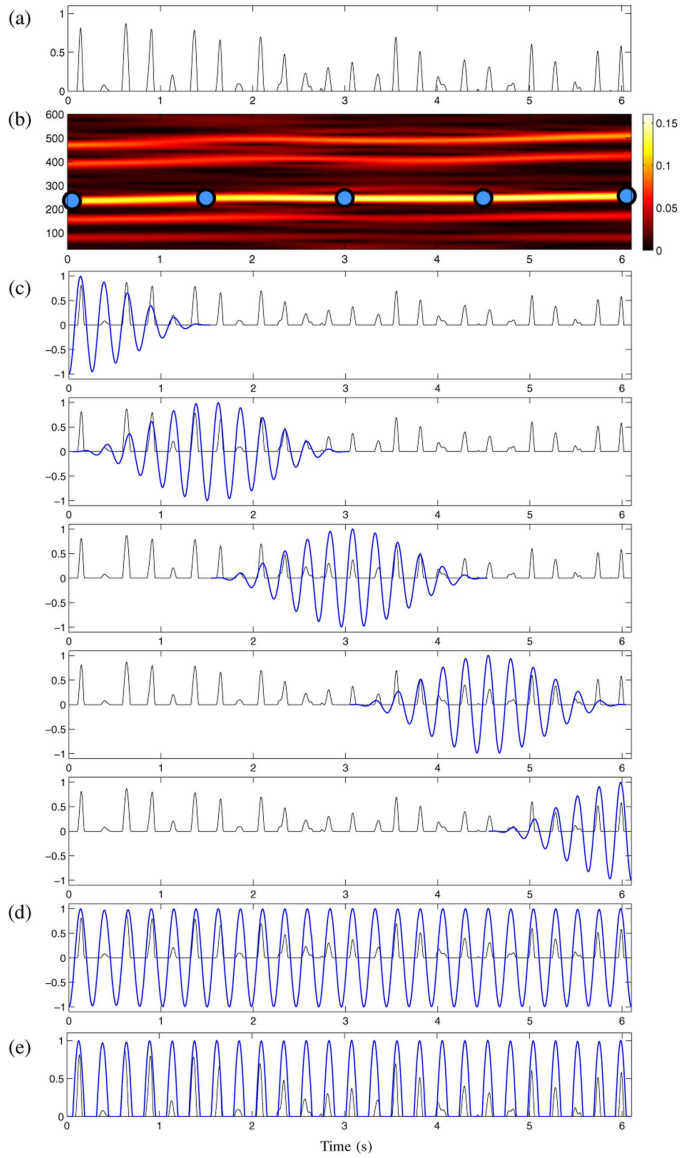

Fig. 2. Illustration of the PLP computation. (a) Novelty curve Δ. (b) Magnitude tempogram $|\mathcal{T}|$ with maxima (indicated by circles) shown at five time positions t. (c) Optimal sinusoidal kernels κ_t (using a kernel size of 3 s) corresponding to the maxima. (d) Accumulation of all kernels (overlap-add). (e) PLP curve Γ obtained after half-wave rectification.

B. Tempogram

A novelty curve typically reveals the note onset candidates in the form of impulse-like spikes. Because of extraction errors and local tempo variations, the spikes may be noisy and irregularly spaced over time. Dealing with spiky novelty curves, autocorrelation methods [10] as well as comb filter techniques [7] may have difficulties in capturing the quasi-periodic information. This is due to the fact that spiky structures are hard to identify by means of spiky analysis functions in the presence of irregularities. In such cases, smoothly spread analysis functions such as sinusoids are better suited to detect locally distorted quasi-periodic patterns. Therefore, similar to [8], we use a

short-time Fourier transform to analyze the local periodic structure of the novelty curves.

For the moment, we assume that the novelty curve is simply a function $\Delta : [1 : T] \to \mathbb{R}$, where $[1 : T] := \{1, 2, \ldots, T\}$, for some $T \in \mathbb{N}$, represents the sampled time axis with respect to a fixed sampling rate. To avoid boundary problems, we assume that Δ is defined on $\mathbb{Z}$ by setting $\Delta(t) := 0$ for $t \in \mathbb{Z} \setminus [1 : T]$. Furthermore, we fix a window function $W : \mathbb{Z} \to \mathbb{R}$ centered at $t = 0$ with support $[-N : N]$ for some $N \in \mathbb{N}$. In the following, we use a Hann window of size $2N + 1$, which is normalized to yield $\sum_{t \in \mathbb{Z}} W(s - t) = 1$ for all $s \in [1 : T]$. Then, for a frequency parameter $\omega \in \mathbb{R}_{\geq 0}$, the complex Fourier coefficient $\mathcal{F}(t, \omega)$ is defined by

$$\mathcal{F}(t, \omega) = \sum_{n \in \mathbb{Z}} \Delta(n) \cdot W(n - t) \cdot \mathrm{e}^{-2\pi i \omega n}. \tag{1}$$

Note that the frequency ω corresponds to the period $1/\omega$. In the context of music, we rather think of tempo measured in beats per minutes (BPM) than of frequency measured in Hertz. Therefore, we use a tempo parameter τ satisfying the equation $\tau = 60 \cdot \omega$.

Similar to a spectrogram, which yields a time-frequency representation, a *tempogram* is a two-dimensional *time-pulse representation* indicating the strength of a local pulse over time; see also [8], [26]. Here, intuitively, a *pulse* can be thought of a periodic sequence of accents, spikes, or impulses. We specify the periodicity of a pulse in terms of a tempo value (in BPM). Now, let $\Theta \subset \mathbb{R}_{>0}$ be a finite set of tempo parameters. Then, we model a tempogram as a function $\mathcal{T} : [1 : T] \times \Theta \to \mathbb{C}$ defined by

$$\mathcal{T}(t, \tau) = \mathcal{F}(t, \tau/60). \tag{2}$$

For an example, we refer to Fig. 2(b), which shows the magnitude tempogram $|\mathcal{T}|$ for the novelty curve shown in Fig. 2(a). Intuitively, the magnitude tempogram indicates for each time position how well the novelty curve can be locally represented by a pulse track of a given tempo. Note that the complex-valued tempogram contains not only magnitude information, but phase information as well. In our experiments, we mostly compute $\mathcal{T}$ using the set $\Theta = [30 : 600]$ covering the (integer) musical tempi between 30 and 600 BPM. Here, the bounds are motivated by the assumption that only events showing a temporal separation between roughly 100 ms (600 BPM) and 2 s (30 BPM) contribute to the perception of tempo [1]. This tempo range requires a spectral analysis of high resolution in the lower frequency range. Therefore, a straightforward fast Fourier transform (FFT) is not suitable. However, since only relatively few frequency bands (tempo values) are needed for the tempogram, computing the required Fourier coefficients individually according to (1) has still a reasonable computational complexity. Typically, we set W to be a Hann window with the size $2N+1$ corresponding to 4–12 s of the audio. The overlap of adjacent windows is adjusted to yield a frame rate of 5 Hz (five frames per second).

C. Predominant Local Periodicity

We now make use of both, the magnitudes and the phases given by $\mathcal{T}$, to derive a mid-level representation that captures the *predominant local pulse* (PLP) of the underlying music signal. Here, the term *predominant pulse* refers to the pulse that is most noticeable in the novelty curve in terms of intensity. Furthermore, our representation is *local* in the sense that it yields the predominant pulse for each time position, thus making local tempo information explicit.

For each $t \in [1 : T]$ we compute the tempo parameter $\tau_t \in \Theta$ that maximizes the magnitude of $\mathcal{T}(t, \tau)$:

$$\tau_t := \operatorname*{argmax}_{\tau \in \Theta} |\mathcal{T}(t, \tau)|. \tag{3}$$

Fig. 2(b) exemplarily shows the predominant local periodicity τ_t for five $t \in [1 : T]$ of the magnitude tempogram. The corresponding phase φ_t is defined by [33]:

$$\varphi_t := \frac{1}{2\pi} \arccos\left(\frac{\mathrm{Re}(\mathcal{T}(t, \tau_t))}{|\mathcal{T}(t, \tau_t)|}\right). \tag{4}$$

Using τ_t and φ_t, the optimal sinusoidal kernel $\kappa_t : \mathbb{Z} \to \mathbb{R}$ for $t \in [1 : T]$ is defined as the windowed sinusoid

$$\kappa_t(n) := W(n - t) \cos(2\pi(n \cdot \tau_t/60 - \varphi_t)) \tag{5}$$

for $n \in \mathbb{Z}$ and the same window function W as used for the tempogram computation in (1). Fig. 2(c) shows the five optimal sinusoidal kernels for the five time parameters indicated in Fig. 2(b) using a Hann window of three seconds. Intuitively, the sinusoid κ_t best explains the local periodic nature of the novelty curve at time position t with respect to the set Θ. The period $60/\tau_t$ corresponds to the predominant periodicity of the novelty curve and the phase information φ_t takes care of accurately aligning the maxima of κ_t and the peaks of the novelty curve. The properties of the kernels κ_t depend not only on the quality of the novelty curve, but also on the window size $2N + 1$ of W and the set of frequencies Θ. Increasing the parameter N yields more robust estimates for τ_t at the cost of temporal flexibility. In the following, this duration is referred to as *kernel size* (KS) and is specified in seconds.

D. PLP Curve

The estimation of optimal periodicity kernels in regions with a strongly corrupted peak structure is problematic. This particularly holds in the case of small kernel sizes. To make the periodicity estimation more robust, our idea is to apply an overlap-add technique, where we accumulate these kernels over all time positions to form a single function instead of looking at the kernels in a one-by-one fashion. Furthermore, we only consider the positive part of the resulting curve (half-wave rectification). More precisely, we define a function $\Gamma : [1 : T] \to \mathbb{R}_{\geq 0}$ as follows:

$$\Gamma(n) = \left| \sum_{t \in [1 : T]} \kappa_t(n) \right|_{\geq 0} \tag{6}$$

for $n \in [1 : T]$, where $|x|_{\geq 0} := x$ for a non-negative real number x and $|x|_{\geq 0} := 0$ for a negative real number x. The resulting function is our mid-level representation referred to as the *PLP curve*. Fig. 2(d) shows the accumulated curve for the five optimal periodicity kernels shown in Fig. 2(c). Note, how the maxima of the periodicity kernels not only align well with the

peaks of the novelty curve, but also with the maxima of neighboring kernels in the overlapping areas, which leads to constructive interferences. Furthermore, note that, because of the normalization of the window W (see Section III-B), the values of the curve lie in the interval $[-1, 1]$ and a local maximum is close to the value one if and only if the overlapping kernels align well. From this, the final PLP curve Γ is obtained through half-wave rectification; see Fig. 2(e).

Note that taking the framewise maximum as in (3) has its assets and drawbacks. On the one hand, it allows the PLP curve to quickly adjust to even sudden changes in tempo and in the dominating pulse level; see Fig. 3 for an example. On the other hand, taking the framewise maximum may lead to unwanted jumps (e.g., random switches between tempo octaves) in the tempo trajectory defined by the maximizing tempo parameter. Here, instead of simply using the context-independent framewise maximum, one may use optimization techniques based on dynamic programming to obtain a context-sensitive smooth tempo trajectory [8], [34]. Similarly, one may constrain the set Θ of tempo parameters in the maximization covering only tempo parameters in a suitable neighborhood of an expected (average) tempo value. Because of the subsequent accumulation step, a small number of outliers does not effect the overall properties of the PLP curve. A larger number of outliers or unwanted switches between tempo octaves, however, may deteriorate the result. Our PLP framework allows for incorporating additional constraints and smoothness strategies in the kernel selection to adjust the properties of the resulting PLP curve according to the requirements of a specific application. The issue of kernel selection will be further discussed in Sections III-E and V-F.

E. Discussion of Properties

We now discuss various properties of PLP curves based on representative examples to demonstrate the benefits of our concept. For an extensive quantitative analysis, we refer to Section V.

As first example, we consider the Waltz No. 2 from Dimitri Shostakovich's Suite for Variety Orchestra No. 1. Fig. 3(a) shows an excerpt (measures 25 to 36) of a piano reduced version of the score of this piece, but the audio recording in this example is an orchestral version conducted by Yablonsky. (The audio excerpt corresponding to measures 25 to 36 has a duration of 10 s.) The manually annotated reference onset positions in this audio excerpt are indicated by the vertical lines in Fig. 3(b). A typical novelty curve (see Section IV for details) for this except is shown in Fig. 3(c). Note that the peaks of this curve strongly correlate with the onset positions. However, the first beats (downbeats) in this 3/4 Waltz are played softly by non-percussive instruments leading to relatively weak and blurred onsets, whereas the second and third beats are played staccato supported by percussive instruments. As a result, the peaks of the novelty curve corresponding to downbeats are hardly visible or even missing, whereas peaks corresponding to the percussive beats are much more pronounced.

Fig. 3(d) shows the magnitude tempogram computed from the novelty curve using a kernel size KS $= 3$ s. Obviously, this tempogram indicates a significant tempo at 210 BPM throughout the audio excerpt, which actually corresponds to the quarter

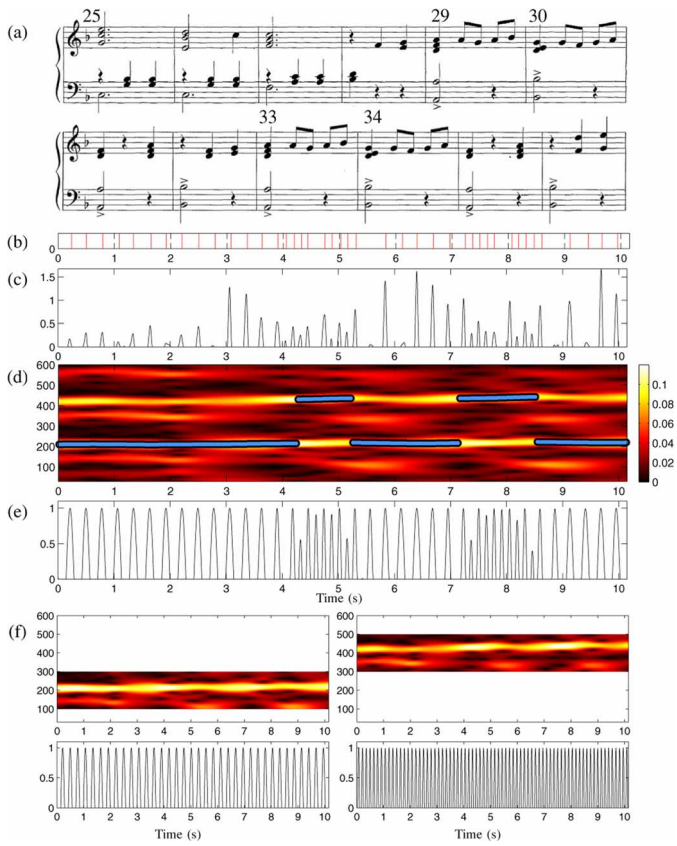

Fig. 3. Excerpt of Shostakovich's Waltz No. 2 from the *Suite for Variety Orchestra No. 1*. (a) Score representation of measures 25 to 36 (in a piano reduced version). (b) Annotated reference onsets (for an orchestral audio recording conducted by Yablonsky). (c) Novelty curve Δ. (d) Magnitude tempogram $|\mathcal{T}|$ using $\Theta = [30 : 600]$ with indication of the predominant tempo. (e) PLP curve Γ. (f) Magnitude tempogram and resulting PLP curve using a constrained tempo set. **Left:** $\Theta = [100 : 300]$ (quarter note tempo range). **Right:** $\Theta = [300 : 500]$ (eighth note tempo range).

note pulse (tactus level) of the piece. Note that this tempo is clearly indicated despite of poor and missing peaks in the novelty curve. Furthermore, the magnitude tempogram additionally reveals high intensities at 420 BPM, which corresponds to the double tempo or eighth note pulse (tatum) level of the piece. Looking at the score, one would expect a predominant tempo which corresponds to the tactus level (score reveals quarter note pulse) for measures 25–28, 31/32, and 35/36 and to the tatum level (score reveals eighth note pulse) for measures 29/30 and 33/34. Indeed, this is exactly reflected by the lines in Fig. 3(d), which indicate the predominant tempo (maximum intensity) for each time position. Note that one has a pulse level switch to the tatum level exactly for the seconds 4–5 (measures 29/30) and seconds 7–8 (measures 33/34).

The PLP curve Γ shown in Fig. 3(e) is obtained from the local tempo estimates. Note that the predominant pulse positions are clearly indicated by the peaks of the PLP curve even though some of the expected peaks were missing in the original novelty curve. Also, the switches between the tactus and tatum level are captured by the PLP curve. In other words, the PLP curve can be regarded as a local periodicity enhancement of the original novelty curve, where the predominant pulse level is taken into account. Although our concept is designed to reveal such locally predominant information, for some applications the

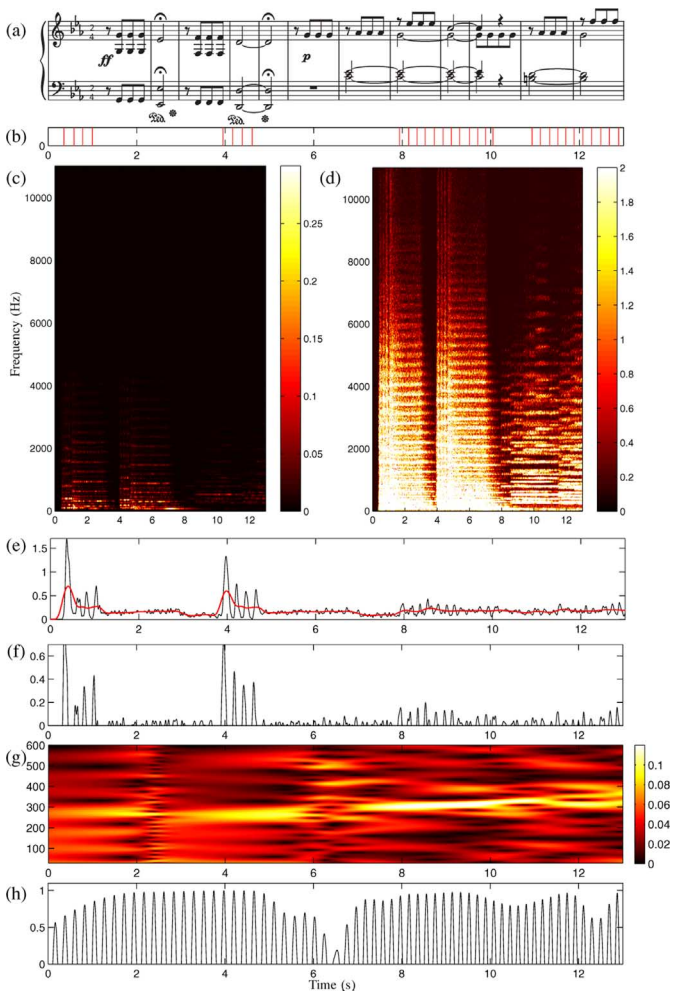

Fig. 4. First 12 measures of Beethoven's Symphony No. 5 (Op. 67). (a) Score representation (in a piano reduced version). (b) Annotated reference onsets (for an orchestral audio recording conducted by Bernstein). (c) Magnitude spectrogram $|X|$. (d) Logarithmically compressed magnitude spectrogram Y. (e) Novelty curve $\bar{\Delta}$ and local mean (red curve). (f) Novelty curve Δ. (g) Magnitude tempogram $|\mathcal{T}|$. (h) PLP curve Γ.

local nature of these estimates might not be desirable. Actually, our PLP framework allows for incorporating prior knowledge on the expected tempo range to exhibit information on different pulse levels. Here, the idea is to constrain the set Θ of tempo parameters in the maximization; see (3). For example, using a constrained set $\Theta = [100 : 300]$ instead of the original set $\Theta = [30 : 600]$, one obtains the tempogram and PLP curve shown in Fig. 3(f) on the left. In this case, the PLP curve correctly reveals the quarter note (tactus) pulse positions with a tempo of 210 BPM. Similarly, using the set $\Theta = [300 : 500]$ reveals the eighth (tatum) note pulse positions and the corresponding tempo of 420 BPM; see Fig. 3(f) on the right. In other words, in the case there is a dominant pulse of (possibly varying) tempo within the specified tempo range Θ, the PLP curve yields a good pulse tracking on the corresponding pulse level.

As second example, we consider a recording of Beethoven's Fifth Symphony. Fig. 4(a) shows the piano reduced version of the first 12 measures of the score. Again, the audio recording is an orchestral version conducted by Bernstein. This piece constitutes a great challenge. First, there are significant local tempo changes (e.g., indicated by hold/pause signs or fermatas). Second, besides very dominant note onsets in the fortissimo section at the beginning of the piece, there are soft and blurred note onsets in the piano section which is mainly played by strings. This is also reflected by the novelty curve shown in Fig. 4(f). The strong onsets in the fortissimo section result in very pronounced peaks, whereas the soft onsets in the piano section (seconds 8–13) can hardly be distinguished from the spurious peaks not related to any note onsets. In particular, the height of a peak is not necessarily a good indicator for the relevance of the peak. However, even though corrupted, the peak structure still possesses some local periodic regularities. These regularities are captured by the periodicity kernels and revealed in the magnitude tempogram shown in Fig. 4(g). Here, at the beginning (second 0 to 6), a tempo of roughly 280 BPM dominates the tempogram. During the second fermata (second 6–7) the tempogram does not show any pronounced tempo. However, in the piano section, the tempogram again indicates a dominating tempo of roughly 300 BPM, which actually corresponds to the eighth note pulse level. Finally, Fig. 4(h) shows the PLP curve Γ. Note that the peaks of Γ align well with the musically relevant onset positions. While note onset positions in the fortissimo section can be directly determined from the original novelty curve, this becomes problematic for the onsets in the piano section. However, exploiting that the note onsets lie on a local rhythmic grid, the PLP curve is capable of capturing meaningful onset information even in the piano passage.

As another important property of our concept, a PLP curve not only reveals positions of predominant pulses but also indicates a kind of confidence in the estimation. Note that the amplitudes of the periodicity kernels do not depend on the amplitude of the novelty curve. This makes a PLP curve invariant under changes in dynamics of the underlying music signal. Recall that we estimate the periodicity kernels using a sliding window technique and add up the kernels over all considered time positions. Since neighboring kernels overlap, constructive and destructive interference phenomena in the overlapping regions influence the amplitude of the resulting PLP curve Γ. Consistent local tempo estimates result in consistent kernels, which in turn produce constructive interferences in the overlap-add synthesis. In such regions, the peaks of the PLP curve assume a value close to one. In contrast, random local tempo estimates result in inconsistent kernels, which in turn cause destructive interferences and lower values of Γ. In Fig. 4(h), this effect is visible in the fermata section (seconds 5 to 8). In Section V-D, we show how this property of PLP curves can be used to detect problematic passages in audio recordings.

Finally, we give a first indication in which way our PLP concept is capable of capturing local tempo changes. To this end, we distinguish between two types of tempo changes. The first type concerns moderate and continuous tempo changes as typically implied by an accelerando or ritardando. To simulate such tempo changes, we generated a pulse train of increasing tempo in the first half and of decreasing tempo in the second half. Fig. 5 shows the resulting novelty curve, the magnitude tempogram, and the PLP curve. As this example indicates, the PLP curve captures well such types of continuous tempo changes—even the amplitude of the PLP curve indicates a high confidence of

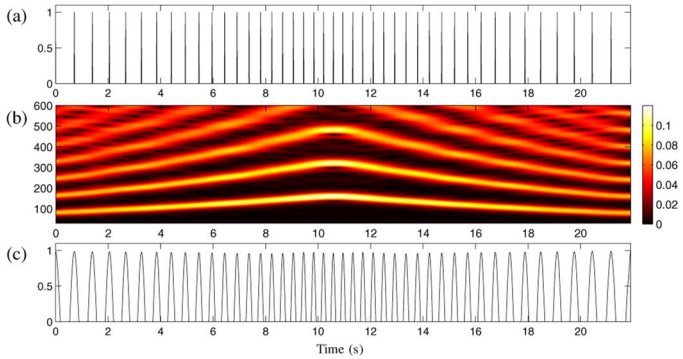

Fig. 5. Behavior of PLP curves under continuous tempo changes (accelerando, ritardando). (a) Impulse train of increasing tempo (80 to 160 BPM, first part) and decreasing tempo (160 to 80 BPM, second part). (b) Magnitude tempogram $|\mathcal{T}|$ for KS $= 4$ s. (c) PLP curve.

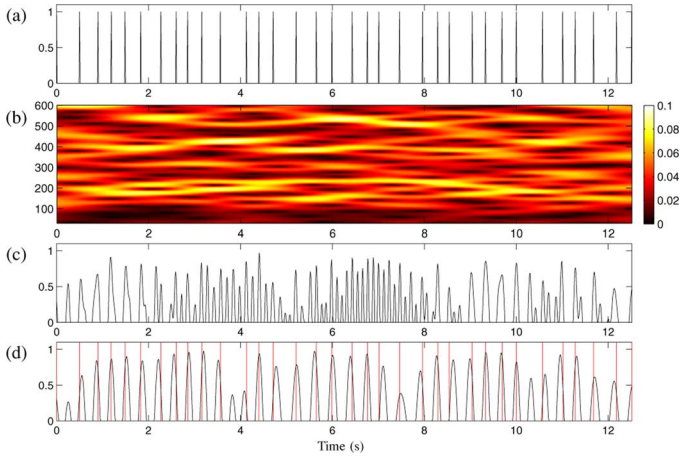

Fig. 6. Behavior of PLP curves under strong local tempo distortions. (a) Impulse train of constant tempo (160 BPM) with random local distortions. (b) Magnitude tempogram $|\mathcal{T}|$. (c) PLP curve. (d) PLP curve for the constrained tempo set $\Theta = [110 : 220]$. Ground-truth pulse positions are indicated by vertical lines.

the estimation; see also Section V-A for a quantitative evaluation. The second type concerns strong local tempo distortions as found in highly expressive music (e.g., romantic piano music). To simulate such tempo changes, we first generated a pulse train of constant tempo (160 BPM) and then locally displaced the impulses in a random fashion. Fig. 6 shows the resulting novelty curve, the magnitude tempogram, and the PLP curve. As this example indicates, the PLP curve fails to capture such extreme distortions—also note the low confidence values. This is not surprising since our PLP concept relies on the assumption of a locally quasi-periodic behavior of the signal. Using a constrained tempo set $\Theta = [110 : 220]$ (similar effects are obtained by using a context-sensitive smooth tempo trajectory; see Section V-F), one obtains an improved PLP curve as shown in Fig. 6(d). However, note that the quasi-periodically spaced peaks of the PLP curve often deviate from the real pulse positions. In Section V, we will further discuss these issues using romantic piano music as extreme example.

IV. Novelty Curve

We now exemplarily describe the approach for computing novelty curves used in our experiments. Here, we combine ideas from [4], [6], [18], [25]. Note, however, that the particular design of the novelty curve is not the focus of this paper. Our mid-level representation as introduced in Section III is designed to work even for noisy novelty curves with a poor peak structure. Naturally, the overall result may be improved by employing more refined novelty curves as suggested in [6].

Given a music recording, a short-time Fourier transform is used to obtain a spectrogram $X = (X(k,t))_{k,t}$ with $k \in [1 : K]$ and $t \in [1 : T]$. Here, K denotes the number of Fourier coefficients, T denotes the number of frames, and $X(k,t)$ denotes the kth Fourier coefficient for time frame t. In our implementation, the discrete Fourier transforms are calculated over Hann-windowed frames of length 46 ms with 50% overlap. Consequently, each time parameter t corresponds to 23 ms of the audio recording. Using suitable binning strategies, various approaches switch over to a logarithmically spaced frequency axis, e.g., by using mel-frequency bands or pitch bands; see [18]. Here, we keep the linear frequency axis, since it puts greater emphasis on the high-frequency regions of the signal, thus accentuating noise bursts that are typically visible in the high-frequency spectrum.

Next, we apply a logarithm to the magnitude spectrogram $|X|$ of the signal yielding $Y := \log(1 + C \cdot |X|)$ for a suitable constant $C > 1$; see [25]. Such a compression step not only accounts for the logarithmic sensation of sound intensity but also allows for adjusting the dynamic range of the signal to enhance the clarity of weaker transients, especially in the high-frequency regions. In our experiments, we use the value $C = 1000$, but our results as well as the findings reported by Klapuri *et al.* [25] show that the specific choice of C does not effect the final result in a substantial way. The effect of this compression step is illustrated by Fig. 4 for our Beethoven example. Fig. 4(c) shows the magnitude spectrogram $|X|$ and Fig. 4(d) the compressed spectrogram Y using $C = 1000$. Here, events with low intensities are considerably enhanced, especially in the high frequency range.

To obtain a novelty curve, we basically compute the discrete temporal derivative of the compressed spectrum Y. In the following, we only consider note onsets (positive derivative) and not note offsets (negative derivative). Therefore, we sum up only over positive intensity changes to obtain the novelty function $\bar{\Delta} : [1 : T-1] \to \mathbb{R}$:

$$\bar{\Delta}(t) := \sum_{k=1}^{K} |Y(k,t+1) - Y(k,t)|_{\geq 0}. \tag{7}$$

Fig. 4(e) shows the resulting curve for the Beethoven example. To obtain our final novelty function Δ, we subtract the local mean [red curve in Fig. 4(e)] from $\bar{\Delta}$ and only keep the positive part (half-wave rectification); see Fig. 4(f). In our implementation, we actually use a higher-order smoothed differentiator [35]. Furthermore, we process the spectrum in a bandwise fashion using five bands. Similar as in [7] these bands are logarithmically spaced and nonoverlapping. Each band is roughly one octave wide. The lowest band covers the frequencies from 0 to 500 Hz, the highest band from 4000 to 11025 Hz. The resulting five novelty curves are summed up to yield the final novelty function. For details, we refer to the quoted literature.

V. Applications and Experiments

In this section, we report on various experiments to demonstrate how our PLP concept can be applied for improving and stabilizing tempo estimation and beat tracking. In Section V-A, we report on a first baseline experiment using synthesized audio material, where we show that our PLP concept can locally estimate the tempo even in the presence of continuous tempo changes. This extends previous approaches to tempo estimation [14], [36] where often one global tempo for the entire recording is determined and used for the evaluation. We then continue with describing our datasets which consist of real audio material and are used in the subsequent experiments (Section V-B), report on our extensive tempo estimation experiments (Section V-C), and show how our PLP concept can be used to measure the confidence of the estimated tempo values (Section V-D). Subsequently, we address the task of beat tracking, which extends tempo estimation in the sense that it additionally considers the phase of the pulses. In Section V-E, we start by reviewing a state-of-the-art beat tracker used in our experiments. Then we report on various experiments, showing that the combined usage of PLP curves with original novelty information significantly improves beat tracking results (Section V-F). Finally, we introduce a novel beat tracking evaluation measure that considers beats in their temporal context (Section V-G).

A. Baseline Experiment to Tempo Estimation

In Section III-E, we indicated that our PLP concept can handle continuous tempo changes; see also Fig. 5. We now give a quantitative evaluation to confirm this property. To this end, we use a representative set of ten pieces from the RWC music database [37] consisting of five classical pieces, three jazz, and two popular pieces; see Table I (first column). The pieces have different instrumentations containing percussive as well as nonpercussive passages of high rhythmic complexity. Using the MIDI files supplied by [37], we manually determined the pulse level that dominates the piece (making the simplistic assumption that the predominant pulse does not change throughout the piece) and set the tempo to a constant value with regard to this pulse; see Table I (second and third columns). The resulting MIDI files are referred to as *original MIDIs*. To simulate continuous tempo changes as implied by accelerandi and ritardandi, we divided the original MIDIs into 20-s segments and alternately applied to each segment a continuous speed up or slow down (referred to as *warping procedure*) so that the resulting tempo of the dominant pulse fluctuates between $+30\%$ and -30% of the original tempo. The resulting MIDI files are referred to as *distorted MIDIs*. Finally, audio files were generated from the original and distorted MIDIs using a high-quality synthesizer.

To evaluate the tempo extraction capability of our PLP concept, we proceed as follows. Given an original MIDI, let τ denote the tempo and let Θ be the set of integer tempo parameters covering the tempo range of $\pm 40\%$ of the original tempo τ. This coarse tempo range reflects the prior knowledge of the respective pulse level (in this experiment, we do not want to deal with tempo octave confusions) and comprises the tempo values of the distorted MIDI. Based on Θ, we compute for each time position t the maximizing tempo parameter $\tau_t \in \Theta$ as defined in (3) for the original MIDI using various kernel sizes. We consider the local tempo estimate τ_t *correct*, if it falls within a 2% deviation of the original tempo τ. The left part of Table I shows the percentage of correctly estimated local tempi for each piece. Note that, even having a constant tempo, there are time positions with incorrect tempo estimates. Here, one reason is that for certain passages the pulse level or the onset information is not suited or simply not sufficient for yielding good local tempo estimations, e.g., caused by musical rests or local rhythmic offsets. For example, for the piece C003 (Beethoven's Fifth), the tempo estimation is correct for 74.5% of the time parameters when using a kernel size (KS) of 4 s. Assuming a constant tempo, it is not surprising that the tempo estimation stabilizes when using a longer kernel. In case of C003, the percentage increases to 85.4% for KS $= 12$ s.

TABLE I
Percentage of Correctly Estimated Local Tempi Using Original MIDI Files (Constant Tempo) and Distorted MIDI Files For Different Kernel Sizes KS $= 4, 6, 8, 12$ s

			original MIDI				distorted MIDI			
Piece	Tempo	Level	4	6	8	12	4	6	8	12
C003	360	1/16	74.5	81.6	83.7	85.4	73.9	81.1	83.3	86.2
C015	320	1/16	71.4	78.5	82.5	89.2	61.8	67.3	71.2	76.0
C022	240	1/8	95.9	100.0	100.0	100.0	95.0	98.1	99.4	89.2
C025	240	1/16	99.6	100.0	100.0	100.0	99.6	100.0	100.0	96.2
C044	180	1/8	95.7	100.0	100.0	100.0	82.6	85.4	77.4	59.8
J001	300	1/16	43.1	54.0	60.6	67.4	37.8	48.4	52.7	52.7
J038	360	1/12	98.6	99.7	100.0	100.0	99.2	99.8	100.0	96.7
J041	315	1/12	97.4	98.4	99.2	99.7	95.8	96.6	97.1	95.5
P031	260	1/8	92.2	93.0	93.6	94.7	92.7	93.7	93.9	93.5
P093	180	1/8	97.4	100.0	100.0	100.0	96.4	100.0	100.0	100.0
average:			86.6	90.5	92.0	93.6	83.5	87.1	87.5	84.6

In any case, the tempo estimates for the original MIDIs with constant tempo only serve as reference values for the second part of our experiment. Using the distorted MIDIs, we again compute the maximizing tempo parameter $\tau_t \in \Theta$ for each time position. Now, these values are compared to the time-dependent distorted tempo values that can be determined from the warping procedure. Analogous to the left part, the right part of Table I shows the percentage of correctly estimated local tempi for the distorted case. The crucial point is that even when using the distorted MIDIs, the quality of the tempo estimations only slightly decreases. For example, in the case of C003, the tempo estimation is correct for 73.9% of the time parameters when using a kernel size of 4 s (compared to 74.5% in the original case). Averaging over all pieces, the percentage decreases from 86.6% (original MIDIs) to 83.5% (distorted MIDIs), for KS $= 4$ s. This clearly demonstrates that our concept allows for capturing even significant tempo changes. As mentioned above, using longer kernels naturally stabilizes the tempo estimation in the case of constant tempo. This, however, does not hold when having music with constantly changing tempo. For example, looking at the results for the distorted MIDI of C044 (Rimski–Korsakov, The Flight of the Bumble Bee), we can note a drop from 82.6% (4-s kernel) to 59.8% (12-s kernel).

B. Datasets

For our subsequent experiments and evaluations, we use five different datasets that consists of real audio recordings (op-

TABLE II
FIVE BEAT-ANNOTATED DATASETS USED IN OUR EXPERIMENTS

Dataset	Audio [#]	Length [s]	Beats [#]	Unannotated [s]	Mean Tempo [BPM]	Std. Tempo [%]
BEATLES	179	28831	52729	1422	116.7	3.3
RWC-POP	100	24406	43659	0	111.7	1.1
RWC-JAZZ	50	13434	19021	0	89.7	4.5
RWC-CLASSIC	61	19741	32733	725	104.8	15.2
MAZURKA	298	45177	85163	1462	126.0	24.6

posed to the synthesized audio material used in Section V-A) and comprise music of various genres and complexities. For all audio recordings, manually generated beat annotations are available. The first collection BEATLES consists of the 12 studio albums by "The Beatles" containing a total number of 179 recordings[1] of Rock/Pop music [38]. Furthermore, we use audio recordings from the RWC Music Database [37], which consists of subcollections of different genres. From this database, we use the three subcollections RWC-POP, RWC-CLASSIC, and RWC-JAZZ containing a total number of 211 recordings. Our fifth dataset MAZURKA consists of piano recordings taken from a collection of 2700 recorded performances for the 49 Mazurkas by Frédéric Chopin. These recordings were collected in the Mazurka Project.[2] For 298 of the 2700 recordings, manually generated beat annotations exist, which have been previously used for the purpose of performance analysis [39]. The dataset MAZURKA consists of exactly these 298 recordings (corresponding to five of the 49 Mazurkas).

Table II gives an overview of the five different datasets. The first four columns of the table indicate the name of the dataset, the number of contained audio recordings, the total length of all audio recordings, and the total number of annotated beat positions. Some recordings contain passages where no meaningful notion of a beat is perceivably. For example, the datasets MAZURKA and RWC-CLASSIC contain some audio files with long passages of silence. Furthermore, in BEATLES, some songs contain noise-like improvisational passages, where the musicians refrain from following any rhythmic pattern. All these passages have not been annotated and are left unconsidered in our evaluation (if not stated otherwise). The fifth column (Unannotated) of Table II indicates the total length of the unannotated passages.

From the beat positions, one can directly derive the local tempo given in BPM. The last two columns of Table II indicate the piecewise mean tempo (in BPM) and standard deviation (in percent) averaged over all recordings of the respective dataset. Note that popular music is often played with constant tempo. This is also indicated by the small values for the standard deviation (e.g., 1.1% for RWC-POP). In contrast, classical music often reveals significant tempo changes, which is indicated by higher values for the standard deviation (e.g., 15.2% for RWC-CLASSIC). These changes can be abrupt as a result of a changing tempo marking (e.g., from *Andante* to *Allegro*) or continuous as indicated by tempo marks such as ritardando or accelerando. Another source for tempo changes is the artistic freedom a musician often takes when interpreting a piece of music. In particular, for romantic piano music such as the Chopin Mazurkas, the tempo consistently and significantly changes from one beat to the next, resulting in pulse sequences similar to the one shown in Fig. 6(a).

[1]Actually, there are 180 songs, but for the song "Revolution 9" no annotations were available. This song is a collage of vocal and music sound clips without any meaningful notion of a beat.

[2]http://mazurka.org.uk/

C. Tempo Estimation Experiments

Continuing the evaluation of Section V-A, we now analyze the tempo estimation capability of our approach on the basis of real audio recordings. To this end, we generate a reference tempo curve for each audio recording of our datasets from the available beat annotations. Here, we first compute the local tempo on the quarter-note level, which is determined by the given inter-beat intervals. The regions before the first beat and after the last beat are left unconsidered in the evaluation. As the tempo values on such a fine temporal level tend to be too noisy, we further smooth the resulting tempo values by considering for each time position the averaged tempo over a range of three consecutive beat intervals. Using the same sampled time axis $[1:T]$ as in Section III-B, we obtain a tempo curve $\tau^{\mathrm{R}}:[1:T]\rightarrow\mathbb{R}_{\geq 0}$ that encodes the local reference tempo for each time position. Now, for each time position t, we compute the maximizing tempo parameter $\tau_t \in \Theta$ as defined in (3). Leaving the problem of tempo octave confusion unconsidered, we say that an estimated local tempo τ_t is *correct*, if it falls within $\pm 4\%$ of an integer multiple[3] $k \in [1,2,\ldots,5,6]$ of the reference tempo $\tau^{\mathrm{R}}(t)$. Here, we choose a tolerance of $\pm 4\%$ as used in [14]. For each recording, we then compute the percentage of correctly estimated tempi and average these values over all recordings of a given dataset.

Table III shows the evaluation results of the local tempo estimation for the five datasets and for different kernel sizes. For popular music, one generally obtains high estimation rates, e.g., an average rate of 94.1% for BEATLES and 95.3% for RWC-POP when using a kernel size (KS) of 4 s. Having constant tempo for most parts, the rates even increase when using longer kernel sizes. For the RWC-JAZZ dataset, the rate is 81.8% (KS $= 4$ s). This lower rate is partly due to passages with soft onsets and complex rhythmic patterns. Using longer kernels, the tempo can be correctly identified even for some of these passages leading to a significantly higher rate of 87.2% (KS $= 12$ s). The situation becomes more complex for classical music, where one has much lower rates, e.g., 70.4% (KS $= 4$ s) for RWC-CLASSIC. Here, a manual inspection reveals two major reasons leading to degradations in the estimation rates. The first reason is again the existence of passages with soft onsets—here, longer kernel sizes help in stabilizing the tempo estimation. The second reason is that for many recordings of classical music one has significant local tempo fluctuation caused by the artistic freedom a musician takes. In such passages, the model assumption of local quasi-periodicity is strongly violated—even within a window of 4 s the tempo may

[3]In general, confusion with integer fractions $k \in [1/2, 1/3, 1/4, \ldots]$ of the tempo may occur, too. However, it can be shown that Fourier-based tempograms (as opposed to, e.g., autocorrelation-based tempograms) respond to tempo harmonics (integer multiples) but suppress tempo subharmonics (integer fractions); see [8], [40]. Since we use Fourier-based tempograms, we only consider confusion with tempo harmonics.

TABLE III
PERCENTAGE OF CORRECTLY ESTIMATED LOCAL TEMPI (±4% TOLERANCE) FOR THE FIVE DATASETS USING THE KERNEL SIZES KS = 4, 6, 8, 12 S

Dataset	4 s	6 s	8 s	12 s
BEATLES	94.1	95.4	95.9	96.3
RWC-POP	95.3	96.7	97.3	98.0
RWC-JAZZ	81.8	85.4	86.6	87.2
RWC-CLASSIC	70.4	70.9	70.3	68.7
MAZURKA	44.5	40.1	37.3	34.3

significantly change by more than 50% percent; see also Fig. 6. Here, it is difficult for the local periodicity kernels to capture meaningful periodic behavior. For such passages, increasing the kernel size has a negative effect on the tempo estimation. In other words, the increase of the kernel size is beneficial for the first type of degradation and detrimental for the second type of degradation. For RWC-CLASSIC, these two effects neutralize each other yielding similar estimation rates for all kernel sizes. However, for the MAZURKA dataset, one mainly has to deal with degradations of the second type. Containing highly expressive romantic piano music, the estimation rate is 44.5% when using KS $= 4$ s. The rate becomes even worse when increasing the kernel size, e.g., 34.3% for KS $= 12$ s. This type of music reveals the limitations of a purely onset-based tempo estimation approach—actually, for such music the notion of local tempo becomes problematic even from a musical point of view; see Section V-D for a continuation of this discussion.

D. Confidence and Limitations

The results for the local tempo estimation significantly degrade in the case that the assumption of local quasi-periodicity is violated. We now show how the PLP concept allows for detecting such problematic passages automatically. As mentioned in Section III-E, constructive and destructive interference phenomena in the overlap-add synthesis influence the amplitude of the resulting PLP curve $\Gamma : [1 : T] \to [0, 1]$. Locally consistent tempo estimations result in amplitude values for the peaks close to one, whereas inconsistent kernel estimations result in lower values. We now exploit this property of Γ to derive a confidence measure for the tempo estimation. To this end, we fix a confidence threshold $\theta \in [0, 1]$ and a length parameter λ. Then, a time interval $I \subseteq [1 : T]$ of length λ is called *reliable* if *all* peaks (local maxima) of Γ positioned in I have a value above θ; otherwise, I is called *unreliable*. The idea is that when I contains at least one peak of lower amplitude, there are inconsistent kernel estimates that make a tempo estimation in I unreliable. Finally, we define the subset $I(\theta, \lambda) \subseteq [1 : T]$ to be the union of all reliable intervals of length λ.

We show that $I(\theta, \lambda)$ indeed corresponds to passages yielding reliable tempo estimates by conducting experiments based on the five datasets of Table II. This time, we include all time positions in the evaluation, even the previously excluded regions without any beat annotations and the regions before the first and after the last beats. Since no meaningful tempo can be assigned to these regions, all estimates within these regions are considered wrong in the evaluation. Here, our motivation is that these regions should automatically be classified as unreliable. The last column of Table IV shows the estimation rates using a kernel size of 4 s. Naturally, including unannotated regions, the rates

TABLE IV
PERCENTAGE OF CORRECTLY ESTIMATED LOCAL TEMPI FOR THE FIVE DATASETS USING RESTRICTED REGIONS $I(\theta, \lambda)$. THE USED PARAMETERS ARE λ = KS = 4 S AND $\theta = 0.95, 0.90, 0.80, 0.70$. THE UNRESTRICTED CASE (LAST COLUMN) CORRESPONDS TO $\theta = 0$. THE RELATIVE SIZE OF $I(\theta, \lambda)$ (IN PERCENT) IS SPECIFIED IN PARENTHESES

Database	0.95	0.90	0.80	0.70	0 (All)
BEATLES	98.5 (59.4)	98.1 (62.9)	97.5 (64.3)	97.2 (66.2)	89.0 (100)
RWC-POP	99.5 (66.5)	99.2 (67.2)	99.1 (69.3)	98.8 (72.5)	92.8 (100)
RWC-JAZZ	94.2 (35.0)	91.4 (40.0)	89.8 (43.8)	89.6 (47.4)	79.0 (100)
RWC-CLASSIC	89.4 (31.4)	84.7 (38.5)	82.4 (43.7)	81.8 (47.1)	67.6 (100)
MAZURKA	74.1 (6.4)	69.2 (11.8)	65.6 (17.8)	62.4 (22.0)	42.0 (100)

are lower compared to the ones reported in the first column of Table III. For example, for the dataset BEATLES, one now has a rate of 89.0% instead of 94.1%.

In our experiments, we use an interval length of $\lambda = 4$ s corresponding to the kernel size KS $= 4$ s. We then compute $I(\theta, \lambda)$ for a fixed threshold θ and evaluate the tempo estimates only on the restricted region $I(\theta, \lambda) \subseteq [1 : T]$. Table IV shows the percentages of correctly estimated local tempi within $I(\theta, \lambda)$ for various thresholds $\theta \in \{0.95, 0.9, 0.8, 0.7\}$. Furthermore, the size of $I(\theta, \lambda)$ relative to $[1 : T]$ is indicated in parentheses (given in percent). For example, for the dataset BEATLES, the restricted region $I(\theta, \lambda)$ with $\theta = 0.95$ covers in average 59.4% of all time positions, while the estimation rate amounts to 98.5%. Lowering the threshold θ, the region $I(\theta, \lambda)$ increases, while the estimation rate decreases. The values of the last column can be seen as the special case $\theta = 0$ resulting in the unrestricted case $I(\theta, \lambda) = [1 : T]$.

Also, for the other datasets, the estimation rates significantly improve when using the restricted region $I(\theta, \lambda)$. In particular, for RWC-POP, the estimation error drops to less than one percent when using $\theta \geq 0.8$, while still covering more than two thirds of all time positions. Actually, for popular music, most of the unreliable regions result from pulse level changes [see Fig. 3(e)] rather than poor tempo estimates. Also, for the classical music dataset RWC-CLASSIC, the estimation rates increase significantly reaching 89.4% for $\theta = 0.95$. However, in this case the restricted regions only cover one third (31.4%) of the time positions. This is even worse for the MAZURKA dataset, where only 6.4% are left when using $\theta = 0.95$. Fig. 7 illustrates one main problem that arises when dealing with highly expressive music where the assumption of local quasi-periodicity is often violated. The passage shows significant tempo fluctuations of the interpretation of the Mazurka Op. 30–2 as indicated by the reference tempo curve τ^{R} in Fig. 7(a). Indeed, the PLP curve allows for detecting regions of locally consistent tempo estimates [indicated by the thick blue lines in Fig. 7(b)]. For these regions the local tempo estimates largely overlap with the reference tempo; see Fig. 7(a).

E. DP Beat Tracking

In the following, we summarize the state-of-the-art beat tracking procedure as introduced in [10]. The input of the algorithm consists of a novelty-like function $\Lambda : [1 : T] \to \mathbb{R}$ (indicating note onset positions) as well as a number $\rho \in \mathbb{Z}$ that yields an estimate of a global (average) beat period $\rho \in \mathbb{Z}$. Assuming a roughly constant tempo, the difference δ of two neighboring beats should be close to ρ. To measure the distance

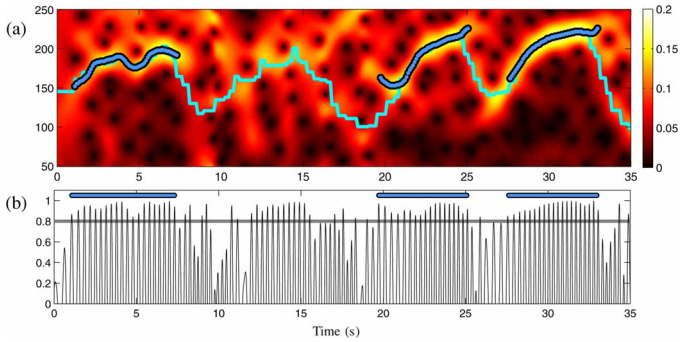

Fig. 7. Tempo estimation for a highly expressive recording (pid9065-14) of Chopin's Mazurka Op30-2. (a) Magnitude tempogram $|\mathcal{T}|$ of the first 35 seconds using $\Theta = [50 : 250]$ with the reference tempo τ^{R} (cyan) and the tempo estimates (thick blue) on $I(\theta, \lambda)$. (b) PLP curve and restricted region $I(\theta, \lambda)$ (blue) for $\theta = 0.8$ and $\lambda = 4$ s.

between δ and ρ, a neighborhood function $N_\rho : \mathbb{N} \to \mathbb{R}$, $N_\rho(\delta) := -(\log_2(\delta/\rho))^2$, is introduced. This function takes the maximum value of 0 for $\delta = \rho$ and is symmetric on a log-time axis. Now, the task is to estimate a sequence $B = (b_1, b_2, \ldots, b_K)$, for some suitable $K \in \mathbb{N}$, of monotonously increasing beat positions $b_k \in [1 : T]$ satisfying two conditions. On the one hand, the value $\Lambda(b_k)$ should be large for all $k \in [1 : K]$, and, on the other hand, the beat intervals $\delta = b_k - b_{k-1}$ should be close to ρ. To this end, one defines the score $S(B)$ of a beat sequence $B = (b_1, b_2, \ldots, b_K)$ by

$$S(B) = \sum_{k=1}^{K} \Lambda(b_k) + \alpha \sum_{k=2}^{K} N_\rho(b_k - b_{k-1}) \tag{8}$$

where the weight $\alpha \in \mathbb{R}$ balances out the two conditions. In our experiments, $\alpha = 5$ turned out to yield a suitable tradeoff. Finally, the beat sequence maximizing S yields the solution of the beat tracking problem. The score-maximizing beat sequence can be obtained by a straightforward dynamic programming (DP) approach; see [10] for details. Therefore, in the following, we refer to this procedure as *DP beat tracking*.

F. Beat Tracking Experiments

We now report on various beat tracking experiments conducted on the five audio datasets described in Section V-B. To investigate the influence of our PLP concept, we consider two different beat tracking approaches. In the first approach, which serves as baseline, we simply perform peak picking based on an adaptive thresholding strategy [4] and define the beat positions to be the detected peak positions. In the second approach, we use the DP beat tracking procedure summarized in Section V-E.

For each of these two approaches, we compare the beat tracking results for five different curves using the original novelty curve Δ, the PLP curve Γ, a constrained PLP curve $\Gamma^{\pm 40}$, a PLP curve Γ^{DP} based on a smooth tempo curve, as well as a combined novelty/PLP curve denoted by Ψ. Here, the PLP curve Γ is computed using $\Theta = [30 : 600]$ and KS $= 4$ s). For the constrained PLP curve $\Gamma^{\pm 40}$, we use a tempo set covering $\pm 40\%$ of the mean tempo of the audio recording, where we assume that a rough estimate of this tempo is given. The curve Γ^{DP} is obtained by first computing a smoothed tempo trajectory based on dynamic programming as described in [34] and then by using these tempo values in the PLP computation instead of the maximizing values, cf. (3). Finally, the combined curve Ψ is defined as $\Psi = (\Delta^{\mathrm{norm}} + \Gamma)/2$, where Δ^{norm} denotes a locally normalized version of Δ that assumes values in the interval [0, 1] (as the PLP curve). The normalization is obtained using a sliding maximum filter of length 4 s (as for the kernel size).

TABLE V
AVERAGE F-MEASURES FOR VARIOUS BEAT TRACKING APPROACHES USING AN ERROR TOLERANCE OF 70 ms

	Peak Picking				DP Beat Tracking		
Dataset	Δ	Γ	$\Gamma^{\pm 40}$	Γ^{DP}	Δ	Γ	Ψ
BEATLES	0.619	0.593	0.671	0.663	0.826	0.741	**0.861**
RWC-POP	0.554	0.507	0.610	0.579	0.786	0.752	**0.819**
RWC-JAZZ	0.453	0.411	0.407	0.407	0.514	**0.573**	0.533
RWC-CLASSIC	0.532	0.514	0.521	0.528	0.618	0.609	**0.644**
MAZURKA	**0.757**	0.618	0.731	0.685	0.641	0.651	0.684

Following the MIREX 2010 Audio Beat Tracking evaluation procedure[4] and the suggestions in the literature [41], a reference beat is considered a *correct detection* (CD) if there is a detected beat within an error tolerance of 70 ms, otherwise a *false negative* (FN). Each detected onset outside all tolerance regions is called a *false positive* (FP). The corresponding numbers are denoted NCD, NFN, and NFP, respectively. From this one obtains precision, recall, and F-measure defined by

$$\mathrm{P} = \frac{\mathrm{NCD}}{\mathrm{NCD} + \mathrm{NFP}}, \quad \mathrm{R} = \frac{\mathrm{NCD}}{\mathrm{NCD} + \mathrm{NFN}}$$
$$\mathrm{F} = \frac{2 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{9}$$

for each piece. The final values are obtained by averaging over all pieces of the respective dataset.

Table V shows the F-measure values for both beat tracking approaches in combination with different curves. Using peak picking based on Δ one obtains an F-measure of $\mathrm{F} = 0.619$ for the dataset BEATLES. Actually, for most music, beat positions go along with onset positions. Consequently, onset positions typically lie on beat positions or on positions corresponding to higher pulse levels. Therefore, even the simple onset detection procedure already yields reasonable F-measures (resulting from a very high recall and a moderate precision). At first sight, it may be surprising that when using peak picking on Γ, one obtains slightly lower F-measure values (e.g., $\mathrm{F} = 0.593$ for BEATLES). Here, note that the peak positions of Γ define a locally periodic pulse grid, where beat positions are likely to occur. As our experiments show, the number of false negatives is reduced in comparison with Δ (leading to a higher recall). However, not all PLP peaks necessarily correspond to beats. Typically, the predominant pulse corresponds to the tatum pulse leading to many false positives (low precision). The situation already improves when using $\Gamma^{\pm 40}$. Constraining the pulse to the correct pulse level reduces the number of false positives (e.g., $\mathrm{F} = 0.671$ for BEATLES). Employing a smooth tempo trajectory for computing Γ^{DP} has a very similar effect (e.g., $\mathrm{F} = 0.663$ for Beatles).

Using the DP beat tracker, the F-measures significantly improve. In general, the best results are achieved when using the DP beat tracker with the combined curve Ψ. In particular, Ψ

[4]http://www.music-ir.org/mirex/wiki/2010: Audio_Beat_Tracking

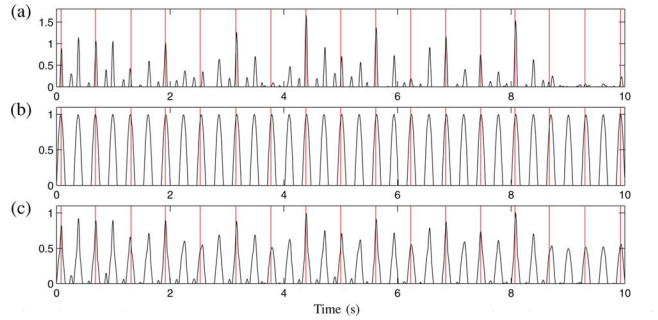

Fig. 8. Illustration of the different curves used in the beat tracking experiments with ground truth beat positions shown as vertical lines. (a) Novelty curve Δ. (b) PLP curve Γ. (c) Combined curve Ψ.

leads to better results than the usual approach exclusively based on Δ. For example, in the case of BEATLES, the F-measure increases from $\mathrm{F} = 0.826$ for Δ to $\mathrm{F} = 0.861$ for Ψ. Using Γ alone, the results seem to degrade ($\mathrm{F} = 0.741$). A manual investigation shows that the PLP curve robustly provides information about likely pulse positions typically at the tatum level, whereas the beat positions correspond to the tactus level. This often results in half-beat shifts (typically one period on the tatum level) in the beat tracking result. In other words, since the peak values are invariant to dynamics, Γ is generally not capable of discriminating between on-beats and off-beats; see Fig. 8(b). The problem of half-beat shifts is not as prominent when using the original novelty curve, since the onset peaks of Δ on the on-beat positions are often more pronounced than the ones on the off-beat positions; see Fig. 8(a). However, the novelty curve Δ often reveals passages with noisy and missing peaks. Here, the combination Ψ inherits the robustness from Γ and the discriminative power from Δ, yielding the best overall beat tracking results; see last column of Table V. Fig. 8(c) illustrates the gain achieved through the combined usage of Γ and Δ.

The evaluation based on the simple F-measure has several weaknesses. First, even completely random false positives only slightly degrade the F-measure. Here, one reason is that the F-measure only moderately punishes peak positions, even if they are not musically meaningful. As a consequence, simple peak picking on Δ, even though ignoring any notion of a beat concept, seems to yield good results. In particular for the MAZURKA dataset, a peak picking based on Δ seems to outperform all other strategies ($\mathrm{F} = 0.757$). Furthermore, this evaluation measure does not account for the issue of half-beat shifts. Finally, evaluating the beats individually, the temporal context of beat tracking is ignored. In Section V-G, we tackle these problems by introducing a novel context-sensitive evaluation measure.

G. Context-Sensitive Evaluation

In the evaluation measure considered so far, the beat positions were evaluated one by one. However, when tapping to the beat of music a listener obviously requires the temporal context of several consecutive beats. Therefore, in evaluating beat tracking procedures, it seems natural to consider beats in the temporal context instead of looking at the beat positions individually [41]. To account for these temporal dependencies, we now introduce a context-sensitive evaluation measure. Let $R = (r_1, r_2, \ldots, r_M)$ be the sequence of monotonously increasing reference beat positions $r_m \in [1 : T], m \in [1 : M]$. Similarly, let $B = (b_1, b_2, \ldots, b_K)$ be the sequence of monotonously increasing detected beat positions $b_k \in [1 : T]$, $k \in [1 : K]$. Furthermore, let $L \in \mathbb{N}$ be a parameter that specifies the temporal context measured in beats, and let ε be the error tolerance corresponding to 70 ms. Then a reference beat r_m is considered an *L-correct detection*, if there exists a subsequence $r_j, \ldots, r_{j+L-1}$ of R containing r_m (i.e., $m \in [j : j + L - 1]$) as well as a subsequence $b_i, \ldots, b_{i+L-1}$ of B such that

$$|r_{j+\ell} - b_{i+\ell}| \leq \varepsilon$$

for all $\ell \in [0 : L - 1]$. Intuitively, for a beat being considered L-correct, one requires an entire track consisting of L consecutive detected beats that match (up to the error tolerance ε) to a track of L consecutive reference beats. Here, a single outlier in the detected beats already destroys this property. Let M^L be the number of L-correct references beats. Then, we define the context-sensitive recall $\mathrm{R}^L := M^L/M$, precision $\mathrm{P}^L := M^L/K$ and F-measure $\mathrm{F}^L := (2 \cdot \mathrm{P}^L \cdot \mathrm{R}^L)/(\mathrm{P}^L + \mathrm{R}^L)$.

In our evaluation, we use the parameter $L = 4$ corresponding to four consecutive beats (roughly a measure). Table VI(a) shows the resulting context-sensitive F^L-measures for the same experiments as described in the last section. Now, the weaknesses of a simple peak picking strategy based on Δ or Γ become obvious. Compared to the previous F-measure (cf. Table V), the F^L-measures drop significantly for all datasets. In particular for popular music, these measures are close to zero (e.g., $\mathrm{F}^L = 0.015$ for RWC-POP and Δ), which indicates that basically no four consecutive beats are detected without any intervening spurious peaks. Actually, the situation already improves significantly when using the constrained PLP curve $\Gamma^{\pm 40}$ (e.g., $\mathrm{F}^L = 0.486$ for RWC-POP). This shows that the PLP curve captures meaningful local beat information when restricted to the desired pulse level. Γ^{DP} once again obtains similar results. For example, in the case of BEATLES $\mathrm{F}^L = 0.554$ for $\Gamma^{\pm 40}$ and $\mathrm{F}^L = 0.555$ for Γ^{DP}. For MAZURKA; however, exhibiting many abrupt tempo changes, Γ^{DP} leads to lower F^L-measures ($\mathrm{F}^L = 0.484$) than $\Gamma^{\pm 40}$ ($\mathrm{F}^L = 0.539$). Here, simply choosing the local maximum from the constrained tempo set allows for locally adapting to the strongly varying tempo. Actually, peak picking on $\Gamma^{\pm 40}$ leads to the best results for this dataset. For all other datasets, however, employing the DP beat tracking procedure improves the results. In particular, for popular music with only moderate tempo changes, the stricter F^L-measures come close to the simple F-measures (e.g., $\mathrm{F}^L = 0.824$ compared to $\mathrm{F} = 0.861$ for BEATLES and Ψ).

To investigate the role of half-beat shifts as discussed in Section V-F, we make the evaluation measure invariant to such errors. To this end, we shift the sequence R of reference beats by one half-beat to the right (replacing r_m by $\tilde{r}_m := (r_{m+1} + r_m)/2$) to obtain a sequence $\tilde{R}$. Then the reference beat r_m is considered correct if r_m is L-correct w.r.t. R or if $\tilde{r}_m$ is L-correct w.r.t. $\tilde{R}$. As before, we define recall and precision to obtain a half-shift invariant F-measure denoted by $\tilde{\mathrm{F}}^L$. Table VI(b) shows the corresponding evaluation results. In

TABLE VI
BEAT TRACKING RESULTS BASED ON CONTEXT-SENSITIVE EVALUATION MEASURES ($L = 4$). (a) F^L-MEASURES. (b) HALF-SHIFT INVARIANT $\bar{\mathrm{F}}^L$-MEASURES

(a)

	Peak Picking				DP Beat Tracking		
Dataset	Δ	Γ	$\Gamma^{\pm 40}$	Γ^{DP}	Δ	Γ	Ψ
BEATLES	0.050	0.044	0.554	0.555	0.789	0.708	**0.824**
RWC-POP	0.015	0.005	0.486	0.444	0.757	0.743	**0.808**
RWC-JAZZ	0.014	0.003	0.253	0.231	0.414	**0.535**	0.493
RWC-CLASSIC	0.124	0.118	0.381	0.393	0.536	0.528	**0.560**
MAZURKA	0.238	0.225	**0.539**	0.484	0.451	0.479	0.508

(b)

	Peak Picking				DP Beat Tracking		
Dataset	Δ	Γ	$\Gamma^{\pm 40}$	Γ^{DP}	Δ	Γ	Ψ
BEATLES	0.050	0.044	0.597	0.592	0.902	0.909	**0.926**
RWC-POP	0.016	0.005	0.528	0.494	0.881	0.917	**0.923**
RWC-JAZZ	0.014	0.003	0.282	0.259	0.564	0.705	**0.708**
RWC-CLASSIC	0.125	0.119	0.404	0.420	0.638	0.633	**0.661**
MAZURKA	0.238	0.226	**0.540**	0.486	0.466	0.528	0.527

particular for the DP tracking approach, one obtains a significant increase in the evaluation measures for all datasets. For example, for BEATLES and Ψ, one has $\tilde{\mathrm{F}}^L = 0.926$ opposed to $\mathrm{F}^L = 0.824$, which shows that half-beat shifts are a common problem in beat tracking. Actually, even humans sometimes perceive beats on off-beat positions, in particular for syncopal passages with strong off-beat events. This also explains the strong increase in the case of Jazz music ($\tilde{\mathrm{F}}^L = 0.708$ opposed to $\mathrm{F}^L = 0.493$ for RWC-JAZZ and Ψ), where one often encounters syncopal elements. For DP tracking based on Γ, the improvements are most noticeable over all datasets. As Γ is invariant to dynamics, the half-shift beat confusion is very distinctive; see Fig. 8(b).

Finally, we note that the context-sensitive evaluation measures much better reveal the kind of improvements introduced by our PLP-concept, which tends to suppress spurious peaks. For both approaches, peak picking and DP beat tracking, one obtains the best results when using a PLP-based enhancement.

VI. CONCLUSION

In this paper, we introduced a novel concept for deriving musically meaningful local pulse information from possibly noisy onset information. Here, opposed to previous approaches that assume constant tempo, the main benefit of our PLP mid-level representation is that it can locally adjust to changes in tempo as long as the underlying music signal possesses some quasi-periodicity. In our representation, we do not aim at extracting pulses at a specific level. Instead, a PLP curve is able to locally switch to the dominating pulse level, which typically is the tatum level. Furthermore, our concept allows for integrating additional knowledge in form of a tempo range to enforce pulse detection on a specific level. Conducting extensive experiments based on well-known datasets of different genres, we have shown that our PLP concept constitutes a powerful tool for tempo estimation and beat tracking. Even for classical music with soft onsets, we were able to extract useful tempo and beat information. However, for highly expressive interpretations of romantic music, the assumption of local quasi-periodicity is often violated leading to poor results. At least, our PLP concept yields a confidence measure to reveal such problematic passages.

Highly-expressive music also reveals the limits of purely onset-oriented tempo and beat tracking procedures. Here, future work is concerned with jointly considering additional musical aspects regarding meter, harmony, polyphony, or structure in order to support and stabilize tempo and beat tracking; see [15], [22], [42], [43] for first approaches towards this direction. For example, in case of the Chopin Mazurkas the tempo and beat is often revealed only by the left hand, whereas the right hand often has an improvisatory character. Here, one may achieve improvements when separating the recording into melody (right hand) and accompaniment (left hand) and analyzing the voices individually. Furthermore, first experiments have shown that PLP curves are suitable for supporting higher level music processing tasks such as music synchronization [44], meter estimation [25], as well as pulse-adaptive feature design [45].

REFERENCES

[1] R. Parncutt, "A perceptual model of pulse salience and metrical accent in musical rhythms," *Music Perception*, vol. 11, pp. 409–464, 1994.
[2] W. A. Sethares, *Rhythm and Transforms*. New York: Springer, 2007.
[3] F. Lerdahl and R. Jackendoff, *Generative Theory of Tonal Music.*. Cambridge, MA: MIT Press, 1983.
[4] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.
[5] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *Proc. AES Conv. 118*, Barcelona, Spain, 2005.
[6] R. Zhou, M. Mattavelli, and G. Zoia, "Music onset detection based on resonator time frequency image," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1685–1695, Nov. 2008.
[7] E. D. Scheirer, "Tempo and beat analysis of acoustical musical signals," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 588–601, 1998.
[8] G. Peeters, "Template-based estimation of time-varying tempo," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, pp. 158–158, 2007.
[9] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1009–1020, Mar. 2007.
[10] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Res.*, vol. 36, no. 1, pp. 51–60, 2007.
[11] P. Grosche and M. Müller, "Computing predominant local periodicity information in music recordings," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, 2009, pp. 33–36.
[12] P. Grosche and M. Müller, "A mid-level representation for capturing dominant tempo and pulse information in music recordings," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Kobe, Japan, 2009, pp. 189–194.
[13] S. Dixon, W. Goebl, and E. Cambouropoulos, "Perceptual smoothness of tempo in expressively performed music," *Music Perception*, vol. 23, no. 3, pp. 195–214, 2006.
[14] A. J. Eronen and A. P. Klapuri, "Music tempo estimation with k-NN regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 50–57, 2010.
[15] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *J. New Music Res.*, vol. 30, no. 2, pp. 159–171, 2001.
[16] A. Holzapfel and Y. Stylianou, "Beat tracking using group delay based onset detection," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Sep. 2008.
[17] C. C. Toh, B. Zhang, and Y. Wang, "Multiple-feature fusion based onset detection for solo singing voice," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Kobe, Japan, 2009.
[18] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Washington, DC, 1999, pp. 3089–3092.
[19] P. Masri and A. Bateman, "Improved modeling of attack transients in music analysis-resynthesis," in *Proc. Int. Comput. Music Conf. (ICMC)*, Hong Kong, 1996, pp. 100–103.

[20] D. Stowell and M. Plumbley, "Adaptive whitening for improved real-time audio onset detection," in *Proc. Int. Comput. Music Conf. (ICMC)*, Copenhagen, Denmark, 2007.
[21] S. Dixon, "Evaluation of the audio beat tracking system beatroot," *J. New Music Res.*, vol. 36, pp. 39–50, 2007.
[22] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *J. New Music Res.*, vol. 30, pp. 39–58, 2001.
[23] J. Seppänen, "Tatum grid analysis of musical signals," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2001, pp. 131–134.
[24] J. Seppänen, A. Eronen, and J. Hiipakka, "Joint beat & tatum tracking from music signals," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Victoria, BC, Canada, 2006.
[25] A. P. Klapuri, A. J. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 342–355, Jan. 2006.
[26] A. T. Cemgil, B. Kappen, P. Desain, and H. Honing, "On tempo tracking: Tempogram representation and Kalman filtering," *J. New Music Res.*, vol. 28, no. 4, pp. 259–273, 2001.
[27] K. Jensen, J. Xu, and M. Zachariasen, "Rhythm-based segmentation of popular chinese music," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005.
[28] J. Foote and S. Uchihashi, "The beat spectrum: A new approach to rhythm analysis," in *Proc. Int. Conf. Multimedia Expo. (ICME)*, Los Alamitos, CA, 2001.
[29] J. Paulus and A. Klapuri, "Measuring the similarity of rhythmic patterns," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Paris, France, 2002, pp. 150–156.
[30] N. Degara, A. Pena, M. E. P. Davies, and M. D. Plumbley, "Note onset detection using rhythmic structure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, 2010, pp. 5526–5529.
[31] J. Bilmes, "Techniques to foster drum machine expressivity," in *Proc. Int. Comput. Music Conf.*, Tokyo, Japan, 1993.
[32] F. Gouyon and P. Herrera, "Pulse-dependent analysis of percussive music," in *Proc. AES 22nd Int. Conf. Virtual, Synth., Entertainment Audio*, Espoo, Finland, 2002.
[33] M. Müller, *Information Retrieval for Music and Motion.*. New York: Springer, 2007.
[34] M. Alonso, G. Richard, and B. David, "Accurate tempo estimation based on harmonic+noise decomposition," *EURASIP J. Adv. Signal Process.*, vol. 2007, p. 14, 2007, pp. Article ID 82 795.
[35] M. Alonso, B. David, and G. Richard, "Tempo and beat estimation of musical signals," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Barcelona, Spain, 2004, pp. 158–163.
[36] M. F. McKinney, D. Moelants, M. E. P. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," *J. New Music Res.*, vol. 36, no. 1, pp. 1–16, 2007.
[37] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Paris, France, 2002.
[38] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler, "OMRAS2 metadata project 2009," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Kobe, Japan, 2009.
[39] C. S. Sapp, "Comparative analysis of multiple musical performances," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 497–500.
[40] P. Grosche, M. Müller, and F. Kurth, "Cyclic tempogram—A midlevel tempo representation for music signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, 2010, pp. 5522–5525.
[41] M. E. P. Davies, N. Degara, and M. D. Plumbley, Evaluation Methods for Musical Audio Beat Tracking Algorithms Queen Mary Univ., London, U.K., Tech. Rep. C4DM-TR-09-06, 2009, Centre for Digital Music.
[42] H. Papadopoulos and G. Peeters, "Simultaneous estimation of chord progression and downbeats from an audio file," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 121–124.
[43] R. B. Dannenberg, "Toward automated holistic beat tracking, music analysis and understanding," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005, pp. 366–373.
[44] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, 2009, pp. 1869–1872.
[45] D. P. W. Ellis, C. V. Cotton, and M. I. Mandel, "Cross-correlation of beat-synchronous representations for music similarity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, 2008, pp. 57–60.

Peter Grosche (S'09) received the B.S. and M.Sc. (Diplom) degrees in electrical engineering and information technology from Technical University of Munich (TUM), Munich, Germany, in 2006 and 2008, respectively. He is currently pursuing the Ph.D. degree in the Multimedia Information Retrieval and Music Processing Group at Saarland University and Max-Planck Institut für Informatik, Saarbrücken, Germany, under the supervision of M. Müller.

Working in the field of music signal processing and music information retrieval, his research interests cover beat tracking, tempo estimation, music segmentation, and music transcription.

Meinard Müller (M'09) received the Diplom degree in mathematics and the Ph.D. degree in computer science from Bonn University, Bonn, Germany.

In 2002/2003, he conducted postdoctoral research in combinatorics in the Mathematical Department of Keio University, Tokyo, Japan. In 2007, he finished his Habilitation at Bonn University in the field of multimedia retrieval writing a book titled *Information Retrieval for Music and Motion*, which appeared as Springer monograph. Currently, he is a member of the Saarland University and the Max-Planck Institut für Informatik, Saarbrücken, Germany, where he leads the research group Multimedia Information Retrieval and Music Processing within the Cluster of Excellence on Multimodal Computing and Interaction. His recent research interests include content-based multimedia retrieval, audio signal processing, music processing, music information retrieval, and motion processing.