

# A MID-LEVEL REPRESENTATION FOR CAPTURING DOMINANT TEMPO AND PULSE INFORMATION IN MUSIC RECORDINGS

Peter Grosche and Meinard Müller

Saarland University and MPI Informatik, Saarbrücken, Germany

{pgrosche, meinard}@mpi-inf.mpg.de

## ABSTRACT

Automated beat tracking and tempo estimation from music recordings become challenging tasks in the case of non-percussive music with soft note onsets and time-varying tempo. In this paper, we introduce a novel mid-level representation which captures predominant local pulse information. To this end, we first derive a tempogram by performing a local spectral analysis on a previously extracted, possibly very noisy onset representation. From this, we derive for each time position the predominant tempo as well as a sinusoidal kernel that best explains the local periodic nature of the onset representation. Then, our main idea is to accumulate the local kernels over time yielding a single function that reveals the predominant local pulse (PLP). We show that this function constitutes a robust mid-level representation from which one can derive musically meaningful tempo and beat information for non-percussive music even in the presence of significant tempo fluctuations. Furthermore, our representation allows for incorporating prior knowledge on the expected tempo range to exhibit information on different pulse levels.

## 1. INTRODUCTION

The automated extraction of tempo and beat information from audio recordings has been a central task in music information retrieval. To accomplish this task, most approaches proceed in two steps. In the first step, positions of note onsets in the music signal are estimated. Here, one typically relies on the fact that note onsets often go along with a sudden change of the signal's energy and spectrum, which particularly holds for instruments such as the piano, guitar, or percussive instruments. This property allows for deriving so-called *novelty curves*, the peaks of which yield good indicators for note onset candidates [1, 15]. In the second step, the novelty curves are analyzed with respect to reoccurring or quasiperiodic patterns. Here, generally spoken, one can roughly distinguish between three different methods. The autocorrelation method allows for detecting periodic self-similarities by comparing a novelty

curve with time-shifted copies [5, 12]. 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 or pulses representing various frequencies and phases [10, 14]. Similarly, one can use a short-time Fourier transform to derive a time-frequency representation of the novelty curve [12]. Here, the novelty curve is compared with templates consisting of sinusoidal kernels each representing a specific frequency. Each of the methods reveals periodicity properties of the underlying novelty curve, from which one can estimate the tempo or beat structure. The intensities of the estimated periodicity, tempo, or beat properties typically change over time and are often visualized by means of spectrogram-like representations referred to as *tempogram* [3], *rhythmogram* [9], or *beat spectrogram* [6].

Relying on previously extracted note onset indicators, tempo and beat tracking tasks become much harder for non-percussive music, where one often has to deal with soft onsets or blurred note transitions. This results in rather noisy novelty curves, exhibiting many spurious peaks. As a consequence, more refined methods have to be used for computing the novelty curves, e. g., by analyzing the signal's spectral content, pitch, or phase [1, 8, 15]. Even more challenging becomes the detection of locally periodic patterns in the case that the music recording reveals significant tempo changes, which typically occur in expressive performances of classical music as a result of *ritardandi*, *accelerandi*, *fermatas*, and so on [4]. Finally, 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 [2]. For example, there are various levels that are presumed to contribute to the human perception of tempo and beat. Most of the previous work focuses on determining musical pulses on the *tactus* (the foot tapping rate or beat [10]) or *measure* level, but only few approaches exist for analyzing the signal on the finer *tatum* level [13]. Here, a *tatum* or *temporal atom* refers to the fastest repetition rate of musically meaningful accents occurring in the signal.

In this paper, we introduce a novel mid-level representation that unfolds predominant local pulse (PLP) information from music signals even for non-percussive music with soft note onsets and changing tempo. Avoiding the explicit determination of note onsets, we derive a tempogram by performing a local spectral analysis on a possibly very noisy novelty curve. From this, we estimate for

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

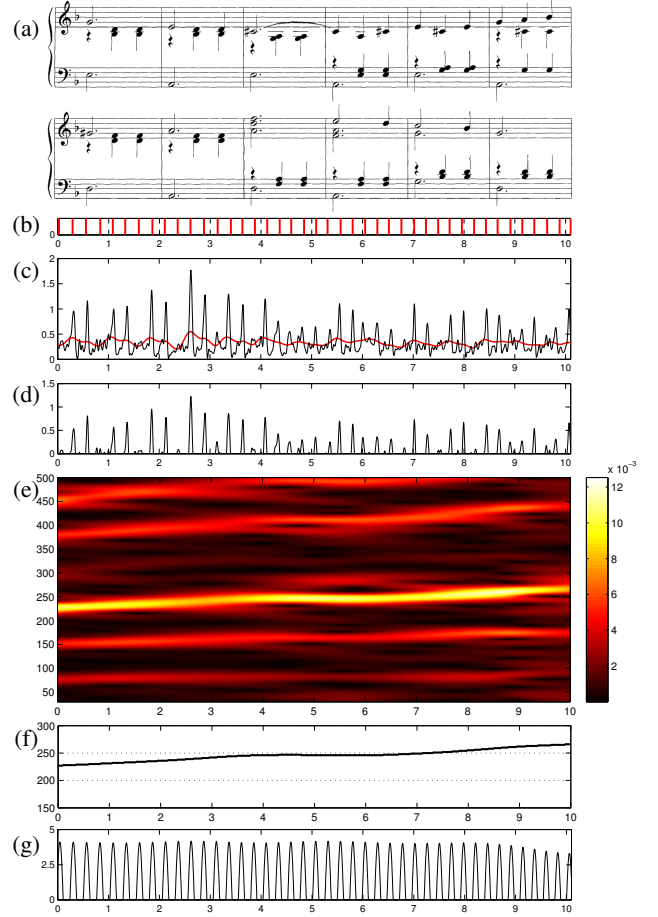
© 2009 International Society for Music Information Retrieval.

each time position a sinusoidal kernel that best explains the local periodic nature of the novelty curve. 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. Our idea is to accumulate all these kernels over time to obtain a mid-level representation, which we refer to as *predominant local pulse* (PLP) curve. As it turns out, PLP curves are robust to outliers and reveal musically meaningful periodicity information even in the case of poor onset information. Note that it is not the objective of our mid-level representation to directly reveal musically meaningful high-level information such as tempo, beat level, or exact onset positions. Instead, our representation constitutes a flexible tool for revealing locally *predominant* information, which may then be used for tasks such as beat tracking, tempo and meter estimation, or music synchronization [10, 11, 14]. In particular, our representation allows for incorporating prior knowledge, e. g., on the expected tempo range, to exhibit information on different pulse levels. In the following sections, we give various examples to illustrate our concept.

The remainder of this paper is organized as follows. In Sect. 2, we review the concept of novelty curves while introducing a variant used in the subsequent sections. Sect. 3 constitutes the main contribution of this paper, where we introduce the tempogram and the PLP mid-level representation. Examples and experiments are described in Sect. 4 and prospects of future work are sketched in Sect. 5.

## 2. NOVELTY CURVE

Combining various ideas from [1, 10, 15], we now exemplarily describe an approach for computing novelty curves that indicate note onset candidates. Note that the particular design of the novelty curve is not in the focus of this paper. Our mid-level representation as introduced in Sect. 3 is designed to work even for noisy novelty curves with a poor pulse structure. Naturally, the overall result may be improved by employing more refined novelty curves as suggested in [15]. Given a music recording, a short-time Fourier transform is used to obtain a spectrogram  $X = (X(k, t))_{k,t}$  with  $k \in [1 : K] := \{1, 2, \dots, K\}$  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  $k^{\text{th}}$  Fourier coefficient for time frame  $t$ . In our implementation, each time parameter  $t$  corresponds to 23 milliseconds of the audio. 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 [10]. 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  $C = 1000$ . To obtain a novelty curve, we basically compute the discrete derivative of the compressed spectrum  $Y$ . More precisely, we sum up only positive intensity changes to emphasize onsets while discarding offsets to obtain the



**Figure 1:** Excerpt of Shostakovich’s second Waltz from Jazz Suite No. 2. The audio recording is a temporally warped orchestral version conducted by Yablonsky with a linear tempo increase (216 – 265 BPM). (a) Piano-reduced score of measures 13 – 24. (b) Ground truth onsets. (c) Novelty curve  $\Delta$  with local mean. (d) Novelty curve  $\bar{\Delta}$ . (e) Magnitude tempogram  $|T|$  for  $KS = 4$  sec. (f) Estimated tempo  $\tau_t$ . (g) PLP curve  $\Gamma$ .

novelty function  $\Delta : [1 : T - 1] \rightarrow \mathbb{R}$ :

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

for  $t \in [1 : T - 1]$ , where  $|x|_{\geq 0} := x$  for a non-negative real number  $x$  and  $|x|_{\geq 0} := 0$  for a negative real number  $x$ . Fig. 1c shows the resulting curve for a music recording of an excerpt of Shostakovich’s second Waltz from the Jazz Suite No. 2. To obtain our final novelty function  $\bar{\Delta}$ , we subtract the local average and only keep the positive part (half-wave rectification), see Fig. 1d. In our implementation, we actually use a higher-order smoothed differentiator. Furthermore, we process the spectrum in a band-wise fashion [14] using 5 bands. The resulting 5 novelty curves are weighted and summed up to yield the final novelty function. For details, we refer to the quoted literature.

## 3. TEMPOGRAM AND PLP CURVE

We now analyze the novelty curve with respect to local periodic patterns. Note that the novelty curve as introduced above typically reveals the note onset candidates in form of impulse-like spikes. Due to extraction errors and local tempo variations, the spikes may be noisy and

irregularly spaced over time. Dealing with spiky novelty curves, autocorrelation methods [5] as well as comb filter techniques [14] encounter difficulties in capturing the quasiperiodic 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 much better suited to detect locally distorted quasiperiodic patterns. Therefore, similar to [12], we use a short-time Fourier transform to analyze the novelty curves. More precisely, let  $\bar{\Delta}$  be the novelty curve as described in Sect. 2. To avoid boundary problems, we assume that  $\bar{\Delta}$  is defined on  $\mathbb{Z}$  by setting  $\bar{\Delta}(t) := 0$  for  $t \in \mathbb{Z} \setminus [1 : T - 1]$ . Furthermore, we fix a window function  $W : \mathbb{Z} \rightarrow \mathbb{R}$  centered at  $t = 0$  with support  $[-N : N]$ . In our experiments, we use a Hann window of size  $2N + 1$ . Then, for a frequency parameter  $\omega \in \mathbb{R}_{\geq 0}$ , the complex Fourier coefficient  $\mathcal{F}(t, \omega)$  is defined by

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

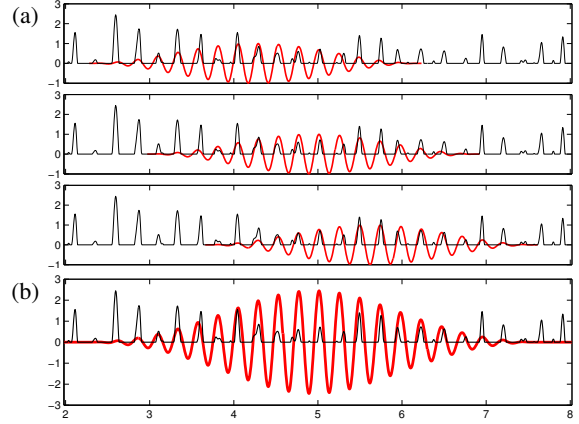
Note that the frequency  $\omega$  corresponds to the period  $1/\omega$ . In the context of beat tracking, we rather think of tempo measured in beats per minutes (BPM) than of frequency measured in Hertz (Hz). Therefore, we use a tempo parameter  $\tau$  satisfying the equation  $\tau = 60 \cdot \omega$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

Similar to a spectrogram, we define a *tempogram* which can be seen as a two-dimensional *time-pulse representation* indicating the strength of the local pulse over time. 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). The semantic level of a pulse is not specified and may refer to the tatum, the tactus, or measure level. Now, let  $\Theta \subset \mathbb{R}_{>0}$  be a finite set of tempo parameters. In our experiments, we mostly use the set  $\Theta = [30 : 500]$ , covering the (integer) musical tempi between 30 and 500 BPM. Here, the bounds are motivated by the assumption that only events showing a temporal separation between 120 milliseconds and 2 seconds contribute to the perception of rhythm [2]. Then, the tempogram is a function  $\mathcal{T} : [1 : T] \times \Theta \rightarrow \mathbb{C}$  defined by

$$\mathcal{T}(t, \tau) = \mathcal{F}(t, \tau/60). \quad (3)$$

For an example, we refer to Fig. 1e, which shows the magnitude tempogram  $|\mathcal{T}|$  for our Shostakovich example. Note that the complex-valued tempogram contains magnitude as well as phase information. 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 accents in 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, see also Fig. 1f. Also, the semantic level of the pulse may change over time, see Fig. 4a. This will be discussed in detail in Sect. 4.

To compute our mid-level representation, we determine for each time position  $t \in [1 : T]$  the tempo parameter



**Figure 2:** (a) Optimal sinusoidal kernel  $\kappa_t$  for various time parameters  $t$  using a kernel size of 4 seconds for the novelty curve shown in Fig. 1d. (b) Accumulation of all kernels. From this, the PLP curve  $\Gamma$  (see Fig. 1f) is obtained by half-wave rectification.

$\tau_t \in \Theta$  that maximizes the magnitude of  $\mathcal{T}(t, \tau)$ :

$$\tau_t := \operatorname{argmax}_{\tau \in \Theta} |\mathcal{T}(t, \tau)|. \quad (4)$$

The corresponding phase  $\varphi_t$  is defined by [11]:

$$\varphi_t := \frac{1}{2\pi} \arccos \left( \frac{\operatorname{Re}(\mathcal{T}(t, \tau_t))}{|\mathcal{T}(t, \tau_t)|} \right). \quad (5)$$

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

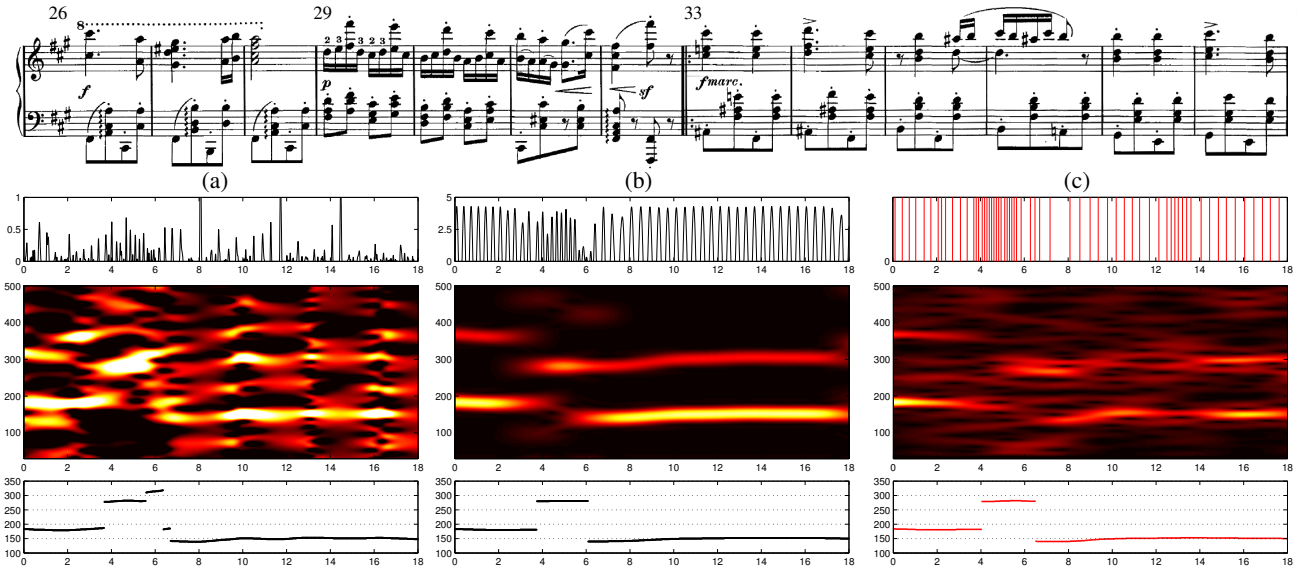
$$\kappa_t(n) := W(n - t) \cos(2\pi(\tau_t/60 \cdot n - \varphi_t)) \quad (6)$$

for  $n \in \mathbb{Z}$ . Fig. 2a shows various optimal sinusoidal kernels for our Shostakovich example. Intuitively, the sinusoid  $\kappa_t$  best explains the local periodic nature of the novelty curve at time position  $t$  with respect to the set  $\Theta$ . The period  $60/\tau_t$  corresponds to the predominant periodicity of the novelty curve and the phase information  $\varphi_t$  takes care of accurately aligning the maxima of  $\kappa_t$  and the peaks of the novelty curve. The properties of the kernels  $\kappa_t$  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  $\Theta$ . Increasing the parameter  $N$  yields more robust estimates for  $\tau_t$  at the cost of temporal flexibility. In our experiments, we chose a window length of 4 to 12 seconds. In the following, this duration is referred to as *kernel size* (KS).

The estimation of optimal sinusoidal kernels for novelty curves with a strongly corrupted pulse structure is still problematic. This particularly holds in the case of small kernel sizes. To make the periodicity estimation more robust, our idea is to accumulate these kernels over all time positions to form a single function instead of looking at the kernels in a one-by-one fashion. More precisely, we define a function  $\Gamma : [1 : T] \rightarrow \mathbb{R}_{\geq 0}$  as follows:

$$\Gamma(n) = \sum_{t \in [1:T]} |\kappa_t(n)|_{\geq 0} \quad (7)$$

for  $n \in [1 : T]$ , see Fig. 2b. The resulting function is our mid-level representation referred to as *PLP curve*. Fig. 1g shows the PLP curve for our Shostakovich example. 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.



**Figure 3:** Excerpt of an orchestral version conducted by Ormandy of Brahms’s Hungarian Dance No. 5. The score shows measures 26 to 38 in a piano reduced version. (a) Novelty curve  $\Delta$ , tempogram derived from  $\Delta$ , and estimated tempo. (b) PLP curve  $\Gamma$ , tempogram derived from  $\Gamma$ , and estimated tempo. (c) Ground-truth pulses, tempogram derived from these pulses, and estimated tempo.  $KS = 4$  sec.

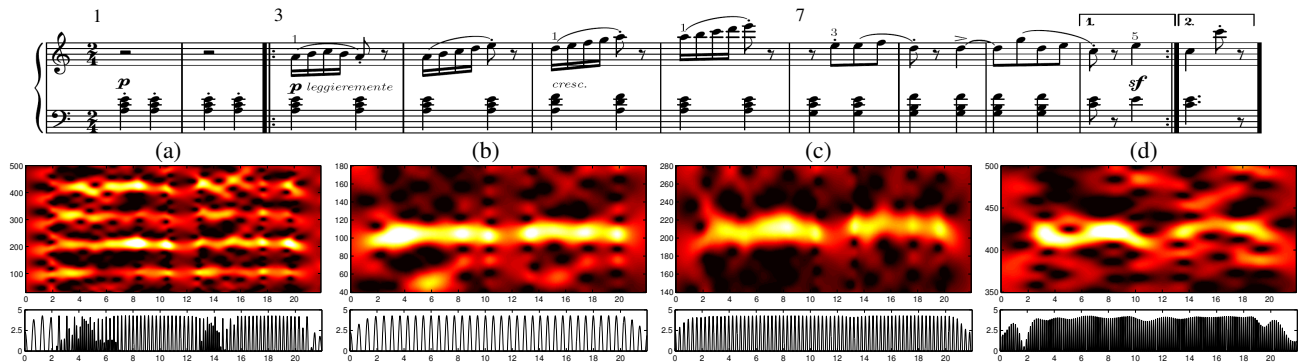
#### 4. DISCUSSION AND EXPERIMENTS

In this section, we discuss various properties of our PLP concept and sketch a number of application scenarios by means of some representative real-world examples. We then give a quantitative evaluation on strongly distorted audio material to indicate the potential of PLP curves for accurately capturing local tempo information.

First, we continue the discussion of our Shostakovich example. Fig. 1a shows a piano-reduced score of the measures 13 – 24. The audio recording (an orchestral version conducted by Yablonsky) has been temporally warped to possess a linearly increasing tempo starting with 216 BPM and ending at 265 BPM at the quarter note level. Firstly, note that the quarter note level has been identified to be the predominant pulse throughout the excerpt, see Fig. 1e. Based on this pulse level, the tempo has been correctly identified as indicated by Fig. 1f. Secondly, first beats in the 3/4 Waltz are played by non-percussive instruments leading to relatively soft and blurred onsets, whereas the second and third beats are played by percussive instruments. This results in some hardly visible peaks in the novelty curve shown in Fig. 1d. However, the beats on the quarter note level are perfectly disclosed by the PLP curve  $\Gamma$  shown in Fig. 1d. In this sense, a PLP curve can be regarded as a periodicity enhancement of the original novelty curve, indicating musically meaningful pulse onset positions. Here, the musical motivation is that the periodic structure of musical events plays a crucial role in the sensation of note changes. In particular, weak note onsets may only be perceptible within a rhythmic context.

As a second example, we consider Brahms’s Hungarian Dance No. 5. Fig. 3 shows a piano reduced version of measures 26 – 38, whereas the audio recording is an orchestral version conducted by Ormandi. This excerpt is very challenging because of several abrupt changes in tempo. Additionally, the novelty curve is rather noisy because of

many weak note onsets played by strings. Fig. 3a shows the extracted novelty curve, the tempogram, and the extracted tempo. Despite of poor note onset information, the tempogram correctly captures the predominant eighth note pulse and the tempo for most time positions. A manual inspection reveals that the excerpt starts with a tempo of 180 BPM (measures 26 – 28, seconds 0 – 4), then abruptly changes to 280 BPM (measures 29 – 32, seconds 4 – 6), and continues with 150 BPM (measures 33 – 38, seconds 6 – 18). Due to the corrupted novelty curve and the rather diffuse tempogram, the extraction of the predominant sinusoidal kernels is problematic. However, accumulating all these kernels smooths out many of the extraction errors. The peaks of the resulting PLP curve  $\Gamma$  (Fig. 3b) correctly indicate the musically relevant eighth note pulse positions in the novelty curve. At this point, we emphasize that all of the sinusoidal kernels have the same unit amplitude independent of the onset strengths. Actually, the amplitude of  $\Gamma$  indicates the confidence in the periodicity estimation. Consistent kernel estimations produce constructive interferences in the accumulation resulting in high values of  $\Gamma$ . Contrary, outliers or inconsistencies in the kernel estimations cause destructive interferences in the accumulation resulting in lower values of  $\Gamma$ . This effect is visible in the PLP curve shown in Fig. 3b, where the amplitude decreases in the region of the sudden tempo change. As noted above, PLP curves can be regarded as a periodicity enhancement of the original novelty curve. Based on this observation, we compute a second tempogram now based on the PLP instead of the original novelty curve. Comparing the resulting tempogram (Fig. 3b) with the original tempogram (Fig. 3a), one can note a significant cleaning effect, where only the tempo information of the dominant pulse (and its harmonics) is maintained. This example shows how our PLP concept can be used in an iterative framework to stabilize local tempo estimations. Finally, Fig. 3c shows the manually generated ground truth onsets



**Figure 4:** Beginning of the Piano Etude Op. 100 No. 2 by Burgmüller. Tempograms and PLP curves ( $KS = 4$  sec) are shown for various sets  $\Theta$  specifying the used tempo range (given in BPM). (a)  $\Theta = [30 : 500]$  (full tempo range). (b)  $\Theta = [40 : 180]$  (quarter note tempo range). (c)  $\Theta = [140 : 280]$  (eighth note tempo range). (d)  $\Theta = [350 : 500]$  (sixteenth note tempo range).

as well as the resulting tempogram (using the onsets as idealized novelty curve). Comparing the three tempograms of Fig. 3 again indicates the robustness of PLP curves to noisy input data and outliers.

In our final example, we look at the beginning of the Piano Etude Op. 100 No. 2 by Burgmüller, see Fig. 4. The audio recording includes the repetition and is played in a rather constant tempo. However, the predominant pulse level changes several times within the excerpt. The piece begins with four quarter note chords (measures 1 – 2), then there are some dominating sixteenth note motives (measures 3 – 6) followed by an eighth note pulse (measures 7 – 10). The change of the predominant pulse level is captured by the PLP curve as shown by Fig. 4a. We now indicate how our PLP concept 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  $\Theta$  of tempo parameters in the maximization (4) of Sect. 3. For example, using a constrained set  $\Theta = [40 : 180]$  instead of the original set  $\Theta = [30 : 500]$ , one obtains the tempogram and PLP curve shown in Fig. 4b. In this case, the PLP curve correctly reveals the quarter note pulse positions as well as the quarter note tempo of 100 BPM. Similarly, using the set  $\Theta = [140 : 280]$  ( $\Theta = [350 : 500]$ ) reveals the eighth (sixteenth) note pulse positions and the corresponding tempos, see Fig. 4c (Fig. 4d). In other words, in the case there is a dominant pulse of (possibly varying) tempo within the specified tempo range  $\Theta$ , the PLP curve yields a good pulse tracking on the corresponding pulse level.

In view of a quantitative evaluation of the PLP concept, we conducted a systematic experiment in the context of tempo estimation. To this end, we used a representative set of ten pieces from the RWC music database [7] consisting of five classical pieces, three jazz, and two popular pieces, see Table 1 (first column). The pieces have different instrumentations containing percussive as well as non-percussive passages of high rhythmic complexity. In this experiment, we investigated to what extent our PLP concept is capable of capturing local tempo deviations. Using the MIDI files supplied by [7], we manually determined the pulse level that dominates the piece. Then, for each MIDI file, we set the tempo to a constant value with regard

to the respective dominant pulse level,<sup>1</sup> see Table 1 (second and third columns). The resulting MIDI files are referred to as *original MIDI*s. We then temporally distorted the MIDI files by simulating strong local tempo changes such as ritardandi, accelerandi, and fermatas. To this end, we divided the original MIDI files into 20-second segments and then 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 MIDI*s. Finally, audio files were generated from the original and distorted MIDI files using a high-quality synthesizer.

To evaluate the tempo extraction capability of our PLP concept, we proceed as follows. Given an original MIDI, let  $\tau$  denote the tempo and let  $\Theta$  be the set of integer tempo parameters covering the tempo range of  $\pm 40\%$  of the original tempo  $\tau$ . 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  $\Theta$ , we compute for each time position  $t$  the maximizing tempo parameter  $\tau_t \in \Theta$  as defined in (4) of Sect. 3 for the original MIDI using various kernel sizes. We consider the local tempo estimate  $\tau_t$  correct, if it falls within a 2% deviation of the original tempo  $\tau$ . The left part of Table 1 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 C022 (Brahms’s Hungarian Dance No. 5), the tempo estimation is correct for 74.5% of the time parameters when using a kernel size (KS) of 4 sec. Assuming a constant tempo, it is not surprising that the tempo estimation stabilizes when using a longer kernel. In case of C022, the percentage increases to 85.4% for  $KS = 12$  sec.

<sup>1</sup> In this experiment, we make the simplistic assumption that the predominant pulse does not change throughout the piece. Actually, this is not true for most pieces such as C003 (Beethoven’s Fifth), C022 (Brahms’s Hungarian Dance No. 5), or J001 (Nakamura’s Jive).

Piece	Tempo	Level	original MIDI				distorted MIDI			
			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
average (after iteration):			89.2	92.0	93.0	95.2	86.0	88.8	88.5	83.1

**Table 1:** Percentage of correctly estimated local tempi for the experiment based on original MIDI files (constant tempo) and distorted MIDI files for kernel sizes  $KS = 4, 6, 8, 12$  sec.

Anyway, 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 1 shows the percentage of correctly estimated local tempi for the distorted case. The crucial point is that even when using strongly distorted MIDIs, the quality of the tempo estimations only slightly decreases. For C022, the tempo estimation is correct for 73.9% of the time parameters when using a kernel size of 4 sec (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  $KS = 4$  sec. 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 sec kernel) to 59.8% (12 sec kernel).

Furthermore, we investigated the iterative approach already sketched for the Brahms example, see Fig 3b. Here, we use the PLP curve as basis for computing a second tempogram from which the tempo estimation is derived. As indicated by the last line of Table 1, this iteration indeed yields an improvement for the tempo estimation for the original as well as the distorted MIDI files. For example, in the distorted case with  $KS = 4$  sec the estimation rate raises from 83.5% (tempogram based on  $\bar{\Delta}$ ) to 86.0% (tempogram based on  $\Gamma$ ).

## 5. CONCLUSIONS

In this paper, we introduced a novel concept for extracting the predominant local pulse even from music with weak non-percussive note onsets and strongly fluctuating tempo. We indicated and discussed various application scenarios ranging from pulse tracking, periodicity enhancement of novelty curves, and tempo tracking, where our mid-level representation yields robust estimations. Furthermore, our

representation allows for incorporating prior knowledge on the expected tempo range to adjust to different pulse levels. In the future, we will use our PLP concept for supporting higher-level music tasks such as music synchronization, tempo and meter estimation, onset detection, as well as rhythm-based audio segmentation. In particular the sketched iterative approach, as first experiments show, constitutes a powerful concept for such applications.

**Acknowledgements:** The research is funded by the *Cluster of Excellence on Multimodal Computing and Interaction* at Saarland University.

## 6. REFERENCES

- [1] 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. on Speech and Audio Processing*, Vol. 13(5), 1035–1047, 2005.
- [2] J. Bilmes: "A Model for Musical Rhythm," in *Proc. ICMC*, San Francisco, USA, 1992.
- [3] A. T. Cemgil, B. Kappen, P. Desain, and H. Honing: "On Tempo Tracking: Tempogram Representation and Kalman Filtering," *Journal of New Music Research*, Vol. 28(4), 259–273, 2001.
- [4] S. Dixon: "Automatic Extraction of Tempo and Beat from Expressive Performances," *Journal of New Music Research*, Vol. 30(1), 39–58, 2001.
- [5] D. P. W. Ellis: "Beat Tracking by Dynamic Programming," *Journal of New Music Research*, Vol. 36(1), 51–60, 2007.
- [6] J. Foote and S. Uchihashi: "The Beat Spectrum: A New Approach to Rhythm Analysis," in *Proc. ICME*, Los Alamitos, USA, 2001.
- [7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka: "RWC Music Database: Popular, Classical and Jazz Music Databases," in *Proc. ISMIR*, Paris, France, 2002.
- [8] A. Holzapfel and Y. Stylianou: "Beat Tracking Using Group Delay Based Onset Detection," in *Proc. ISMIR*, Philadelphia, USA, 2008.
- [9] K. Jensen, J. Xu, and M. Zachariassen: "Rhythm-Based Segmentation of Popular Chinese Music," in *Proc. ISMIR*, London, UK, 2005.
- [10] A. P. Klapuri, A. J. Eronen, and J. Astola: "Analysis of the meter of acoustic musical signals," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 14(1), 342–355, 2006.
- [11] M. Müller: "Information Retrieval for Music and Motion," Springer, 2007.
- [12] G. Peeters: "Template-based estimation of time-varying tempo," *EURASIP Journal on Advances in Signal Processing*, Vol. 2007, 158–171, 2007.
- [13] J. Seppänen: "Tatum grid analysis of musical signals," in *Proc. IEEE WASPAA*, New Paltz, USA, 2001.
- [14] E. D. Scheirer: "Tempo and beat analysis of acoustical musical signals," *Journal of the Acoustical Society of America*, Vol. 103(1), 588–601, 1998.
- [15] R. Zhou, M. Mattavelli, and G. Zoia: "Music Onset Detection Based on Resonator Time Frequency Image," *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 16(8), 1685–1695, 2008.