WHAT MAKES BEAT TRACKING DIFFICULT? A CASE STUDY ON CHOPIN MAZURKAS

Peter Grosche
Saarland University
and MPI Informatik
pgrosche@mpi-inf.mpg.de

Meinard Müller
Saarland University
and MPI Informatik
meinard@mpi-inf.mpg.de

Craig Stuart Sapp
Standford University
CCRMA / CCARH
craig@ccrma.stanford.edu

ABSTRACT

The automated extraction of tempo and beat information from music recordings is a challenging task. Especially in the case of expressive performances, current beat tracking approaches still have significant problems to accurately capture local tempo deviations and beat positions. In this paper, we introduce a novel evaluation framework for detecting critical passages in a piece of music that are prone to tracking errors. Our idea is to look for consistencies in the beat tracking results over multiple performances of the same underlying piece. As another contribution, we further classify the critical passages by specifying musical properties of certain beats that frequently evoke tracking errors. Finally, considering three conceptually different beat tracking procedures, we conduct a case study on the basis of a challenging test set that consists of a variety of piano performances of Chopin Mazurkas. Our experimental results not only make the limitations of state-of-the-art beat trackers explicit but also deepens the understanding of the underlying music material.

1. INTRODUCTION

When listening to a piece of music, most humans are able to tap to the musical beat without difficulty. In recent years, various different algorithmic solutions for automatically extracting beat position from audio recordings have been proposed. However, transferring this cognitive process into an automated system that reliably works for the large variety of musical styles is still not possible. Modern pop and rock music with a strong beat and steady tempo can be handled by many methods well, but extracting the beat locations from highly expressive performances of, *e.g.*, romantic piano music, is a challenging task.

To better understand the shortcomings of recent beat tracking methods, significant efforts have been made to compare and investigate the performance of different strategies on common datasets [6, 10, 13]. However, most approaches were limited to comparing the different methods by specifying evaluation measures that refer to an entire recording or even an entire collection of recordings. Such globally oriented evaluations do not provide any information on the critical passages within a piece where the tracking errors occur. Thus, no conclusions can be drawn from these experiments about possible *musical reasons* that lie behind the beat tracking errors. A first analysis of *musical properties* influencing the beat tracking quality was conducted by Dixon [6], who proposed quantitative measures for the rhythmic complexity and for variations in tempo and timings. However, no larger evaluations were carried out to show a correlation between these theoretical measures and the actual beat tracking quality.

In this paper, we continue this strand of research by analyzing the tracking results obtained by different beat tracking procedures. As one main idea of this paper, we introduce a novel evaluation framework that exploits the existence of different performances available for a given piece of music. For example, in our case study we revert to a collection of recordings for the Chopin Mazurkas containing in average over 50 performances for each piece. Based on a local, beat-wise histogram, we simultaneously determine consistencies of beat tracking errors over many performances. The underlying assumption is, that tracking errors consistently occurring in many performances of a piece are likely caused by musical properties of the piece, rather than physical properties of a specific performance. As a further contribution, we classify the beats of the critical passages by introducing various types of beats such as non-event beats, ornamented beats, weak bass beats, or constant harmony beats. Each such beat class stands for a musical performance-independent property that frequently evokes beat tracking errors. In our experiments, we evaluated three conceptually different beat tracking procedures on a corpus consisting of 300 audio recordings corresponding to five different Mazurkas. For each recording, the tracking results were compared with manually annotated ground-truth beat positions. Our local evaluation framework and detailed analysis explicitly indicates various limitations of current state-of-the-art beat trackers, thus laying the basis for future improvements and research directions.

This paper is organized as follows: In Sect. 2, we formalize and discuss the beat tracking problem. In Sect. 3, we describe the underlying music material and specify various beat classes. After summarizing the three beat tracking strategies (Sect. 4) and introducing the evaluation measure (Sect. 5) used in our case study, we report on the experimental results in Sect. 6. Finally, we conclude in

ID	Composer	Piece	#(Meas.)	#(Beats)	#(Perf.)
M17-4	Chopin	Op. 17, No. 4	132	396	62
M24-2	Chopin	Op. 24, No. 2	120	360	64
M30-2	Chopin	Op. 30, No. 2	65	193	34
M63-3	Chopin	Op. 63, No. 3	77	229	88
M68-3	Chopin	Op. 68, No. 3	61	181	50

Table 1: The five Chopin Mazurkas and their identifiers used in our study. The last three columns indicate the number of measures, beats, and performances available for the respective piece.

Sect. 7 with a discussion of future research directions. Further related work is discussed in the respective sections.

2. PROBLEM SPECIFICATION

For a given piece of music, let N denote the number of *musical beats*. Enumerating all beats, we identify the set of musical beats with the set $\mathcal{B} = [1:N] := \{1, 2, \ldots, N\}$. Given a performance of the piece in the form of an audio recording, the musical beats correspond to specific physical time positions within the audio file. Let $\pi : \mathcal{B} \to \mathbb{R}$ be the mapping that assigns each musical beat $b \in \mathcal{B}$ to the time position $\pi(b)$ of its occurrence in the performance. In the following, a time position $\pi(b)$ is referred to as *physical beat* or simply as *beat* of the performance. Then, the task of *beat tracking* is to recover the set $\{\pi(b) \,|\, b \in \mathcal{B}\}$ of all beats from a given audio recording.

Note that this specification of the beat tracking problem is somewhat simplistic, as we only consider physical beats that are defined by onset events. More generally, a beat is a perceptual phenomenon and perceptual beat times do not necessarily coincide with physical beat times [7]. Furthermore, the perception of beats varies between listeners.

For determining physical beat times, we now discuss some of the problems, one has to deal with in practice. Typically, a beat goes along with a note onset revealed by an increase of the signal's energy or a change in the spectral content. However, in particular for non-percussive music, one often has soft note onsets, which lead to blurred note transitions rather than sharp note onset positions. In such cases, there are no precise timings of note events within the audio recording, and the assignment of exact physical beat positions becomes problematic. This issue is aggravated in the presence of tempo changes and expressive tempo nuances (*e.g.*, ritardando and accelerando).

Besides such physical reasons, there may also be a number of musical reasons for beat tracking becoming a challenging task. For example, there may be beats with no note event going along with them. Here, a human may still perceive a steady beat, but the automatic specification of physical beat positions is quite problematic, in particular in passages of varying tempo where interpolation is not straightforward. Furthermore, auxiliary note onsets can cause difficulty or ambiguity in defining a specific physical beat time. In music such as the Chopin Mazurkas, the main melody is often embellished by ornamented notes such as trills, grace notes, or arpeggios. Also, for the sake of expressiveness, the notes of a chord need not be played at the same time, but slightly displaced in time. This renders a precise definition of a physical beat position impossible.

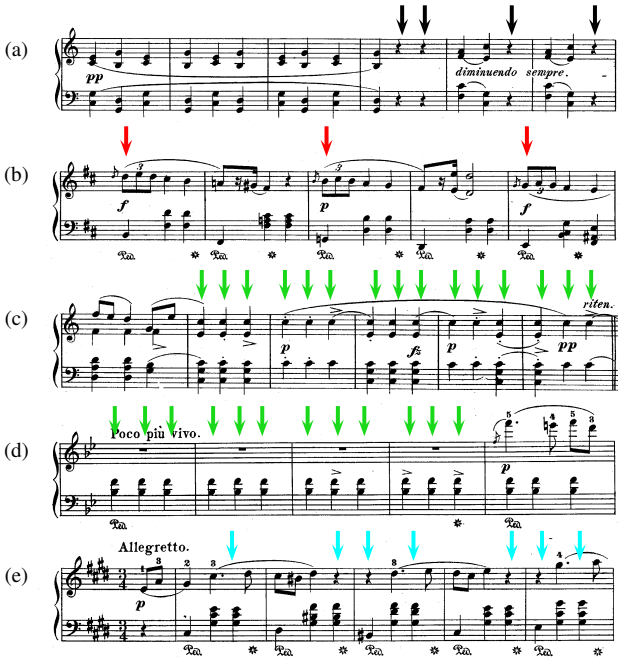

Figure 1: Scores of example passages for the different beat classes introduced in Sect. 3. **(a)** Non-event beats ($\mathcal{B}_1$) in M24-2, **(b)** Ornamented beats ($\mathcal{B}_3$) in M30-2, **(c)** Constant harmony beats ($\mathcal{B}_5$) in M24-2, **(d)** Constant harmony beats ($\mathcal{B}_5$) in M68-3, and **(e)** Weak bass beats ($\mathcal{B}_4$) in M63-3.

3. DATA AND ANNOTATIONS

The Mazurka Project [1] has collected over 2700 recorded performances for 49 Mazurkas by Frédéric Chopin, ranging from the early stages of music recording (Grünfeld 1902) until today [15]. In our case study, we use 298 recordings corresponding to five of the 49 Mazurkas, see Table 1. For each of theses recordings the beat positions were annotated manually [15]. These annotations are used as ground truth in our experiments. Furthermore, Humdrum and MIDI files of the underlying musical scores for each performance are provided, representing the pieces in an uninterpreted symbolic format.

In addition to the physical beat annotations of the performances, we created musical annotations by grouping the musical beats $\mathcal{B}$ in five different beat classes $\mathcal{B}_1$ to $\mathcal{B}_5$. Each of these classes represents a musical property that typically constitutes a problem for determining the beat positions. The colors refer to Fig. 4 and Fig. 5.

- **Non-event beats $\mathcal{B}_1$ (black)**: Beats that do not coincide with any note events, see Fig. 1(a).
- **Boundary beats $\mathcal{B}_2$ (blue)**: Beats of the first measure and last measure of the piece.
- **Ornamented beats $\mathcal{B}_3$ (red)**: Beats that coincide with ornaments such as trills, grace notes, or arpeggios, see Fig. 1(b).
- **Weak bass beats $\mathcal{B}_4$ (cyan)**: Beats where only the left hand is played, see Fig. 1(e).
- **Constant harmony beats $\mathcal{B}_5$ (green)**: Beats that correspond to consecutive repetitions of the same chord, see Fig. 1(c-d).

Furthermore, let $\mathcal{B}_* := \cup_{k=1}^{5} \mathcal{B}_k$ denote the union of the five beat classes. Table 2 details for each Mazurka the number of beats assigned to the respective beat classes.

ID	$\|\mathcal{B}\|$	$\|\mathcal{B}_1\|$	$\|\mathcal{B}_2\|$	$\|\mathcal{B}_3\|$	$\|\mathcal{B}_4\|$	$\|\mathcal{B}_5\|$	$\|\mathcal{B}_*\|$
M17-4	396	9	8	51	88	0	154
M24-2	360	10	8	22	4	12	55
M30-2	193	2	8	13	65	0	82
M63-3	229	1	7	9	36	0	47
M68-3	181	17	7	0	14	12	37

Table 2: The number of musical beats in each of the different beat classes defined in Sect. 3. Each beat may be a member of more than one class.

Note that the beat classes need not be disjoint, *i.e.*, each beat may be assigned to more than one class. In Sect. 6, we discuss the beat classes and their implications on the beat tracking results in more detail.

4. BEAT TRACKING STRATEGIES

Beat tracking algorithms working on audio recordings typically proceed in three steps: In the first step, note onset candidates are extracted from the signal. More precisely, a *novelty curve* is computed that captures changes of the signal's energy, pitch or spectral content [3, 5, 8, 12]. The peaks of this curve indicate likely note onset candidates. Fig. 2(c) shows a novelty curve for an excerpt of M17-4 (identifier explained in Table 1). Using a peak picking strategy [3] note onsets can be extracted from this curve. In the second step, the local tempo of the piece is estimated. Therefore, the onset candidates are analyzed with respect to locally periodic or reoccurring patterns [5, 12, 14]. The underlying assumption is that the tempo of the piece does not change within the analysis window. The choice of the window size constitutes a trade-off between the robustness of the tempo estimates and the capability to capture tempo changes. In the third step, the sequence of beat positions is determined that best explains the locally periodic structure of the piece, in terms of frequency (tempo) and phase (timing) [5, 12], see Fig. 2(d).

In our experiments we use three different beat trackers. First, we directly use the onset candidates extracted from a novelty curve capturing spectral differences [11] as indicated by Fig. 2(c). In this method, referred to as ONSET in the following sections, each detected note onset is considered as a beat position. Second, as a representative of the beat tracking algorithms that transform the novelty curve into the frequency (tempo) or periodicity domain [5, 12, 14], we employ the predominant local periodicity estimation [11], referred to as PLP in the following. We use a window size of three seconds and initialize the tempo estimation with the mean of the annotated tempo. More precisely, we define the global tempo range for each performance covering one octave around the mean tempo, *e.g.*, for a mean tempo of 120 BPM, tempo estimates in the range $[90 : 180]$ are valid. This prevents tempo doubling or halving errors and robustly allows for investigating beat tracking errors, rather than tempo estimation errors. The third beat tracking method (SYNC) we use in our experiments employs the MIDI file available for each piece. This MIDI file can be regarded as additional knowledge, including the pitch, onset time and duration of each note. Using suitable synchronization techniques [9] on the basis of coarse harmonic and very precise onset information,

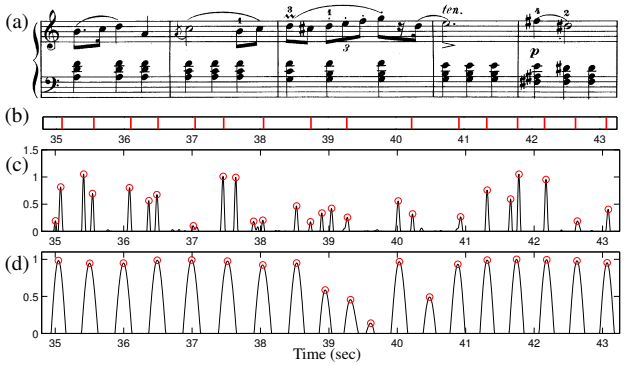

Figure 2: Representations for an excerpt of M17-4. **(a)** Score representation of beats 60 to 74. **(b)** Annotated ground truth beats for the performance pid50534-05 by Horowitz (1985), see [1]. **(c)** Novelty curve (note onset candidates indicated by circles). **(d)** PLP curve (beat candidates indicated by circles).

we identify for each musical event of the piece (given by the MIDI file) the corresponding physical position within a performance. This coordination of MIDI events to the audio is then used to determine the beat positions in a performance and simplifies the beat tracking task to an alignment problem, where the number of beats and the sequence of note events is given as prior knowledge.

5. EVALUATION MEASURES

Many evaluation measures have been proposed to quantify the performance of beat tracking systems [4] by comparing the beat positions determined by a beat tracking algorithm and annotated ground truth beats. These measures can be divided into two groups. Firstly, measures that analyze each beat position separately and secondly, measures that take the tempo and metrical levels into account [5, 12, 13]. While the latter gives a better estimate of how well a *sequence* of retrieved beats correlates with the manual annotation, it does not give any insight into the beat tracking performance at a specific beat of the piece.

In this paper, we evaluate the beat tracking quality on the beat-level of a piece and combine the results of all performances available for this piece. This allows for detecting beats that are prone to errors in many performances. For a given performance, let $\Pi := \{\pi(b) \,|\, b \in \mathcal{B}\}$ be the set of manually determined physical beats, which are used as ground truth. Furthermore, let $\Phi \subset \mathbb{R}$ be the set of beat candidates obtained from a beat tracking procedure. Given a tolerance parameter $\tau > 0$, we define the τ-neighborhood $I_\tau(p) \subset \mathbb{R}$ of a beat $p \in \Pi$ to be the interval of length 2τ centered at p, see Fig. 3. We say that a beat p has been *identified* if there is a beat candidate $q \in \Phi$ in the τ-neighborhood of p, *i.e.*, $q \in \Phi \cap I_\tau(p)$. Let $\Pi_{\mathrm{id}} \subset \Pi$ be the set of all identified beats. Furthermore, we say that a beat candidate $q \in \Phi$ is *correct* if q lies in the τ-neighborhood $I_\tau(p)$ of some beat $p \in \Pi$ and there is no other beat candidate lying in $I_\tau(p)$ that is closer to p than q. Let $\Phi_{\mathrm{co}} \subset \Phi$ be the set of all correct beat candidates. We then define the precision $\mathrm{P} = \mathrm{P}_\tau$, the recall $\mathrm{R} = \mathrm{R}_\tau$, and F-measure $\mathrm{F} = \mathrm{F}_\tau$ as [4]

$$\mathrm{P} = \frac{|\Phi_{\mathrm{co}}|}{|\Phi|}, \quad \mathrm{R} = \frac{|\Pi_{\mathrm{id}}|}{|\Pi|}, \quad \mathrm{F} = \frac{2 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}. \qquad (1)$$

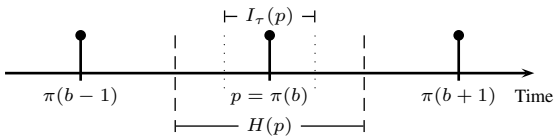

Figure 3: Illustration of the τ-neighborhood $I_\tau(p)$ and the half-beat neighborhood $H(p)$ of a beat $p = \pi(b)$, $b \in \mathcal{B}$.

Table 3 shows the results of various beat tracking procedures on the Mazurka data. As it turns out, the F-measure is a relatively soft evaluation measure that only moderately punishes additional, non-correct beat candidates. As a consequence, the simple onset-based beat tracker seems to outperform most other beat trackers. As for the Mazurka data, many note onsets coincide with beats, the onset detection leads to a high recall, while having only a moderate deduction in the precision.

We now introduce a novel evaluation measure that punishes non-correct beat candidates, which are often musically meaningless, more heavily. To this end, we define a *half-beat neighborhood* $H(p)$ of a beat $p = \pi(b) \in \Pi$ to be the interval ranging from $\frac{\pi(b-1)-\pi(b)}{2}$ (or $\pi(b)$ for $b = 1$) to $\frac{\pi(b+1)-\pi(b)}{2}$ (or $\pi(b)$ for $b = N$), see Fig. 3. Then, we say that a beat $b \in \mathcal{B}$ has been *strongly identified* if there is a beat candidate $q \in \Phi$ with $q \in \Phi \cap I_\tau(p)$ and if $H(p) \cap \Phi = \{q\}$ for $p = \pi(b)$. In other words, q is the only beat candidate in the half-beat neighborhood of p. Let $\Pi_{\text{stid}} \subset \Pi$ be the set of all strongly identified beats, then we define the *beat accuracy* $A = A_\tau$ to be

$$A = \frac{|\Pi_{\text{stid}}|}{|\Pi|}. \quad (2)$$

6. EXPERIMENTS

We now discuss the experimental results obtained using our evaluation framework and explain the relations between the beat tracking results and the beat classes introduced in Sect. 3.

We start with discussing Table 3. Here, the results of the different beat tracking approaches for all performances of the five Mazurkas are summarized, together with some results from the MIREX 2009 beat tracking task [2]. All beat trackers used in our evaluation yield better results for the Mazurkas than all trackers used in the MIREX evaluation. As noted before, the F-measure only moderately punishes additional beats. In consequence, ONSET ($\mathrm{F} = 0.754$) seems to outperform all other methods, except SYNC ($\mathrm{F} = 0.890$). In contrast, the introduced beat accuracy A punishes false positives more heavily, leading to $A = 0.535$ for ONSET, which is significantly lower than for PLP ($A = 0.729$) and SYNC ($A = 0.890$). For SYNC, the evaluation metrics P, R, F, and A are equivalent because the number of detected beats is always correct. Furthermore, SYNC is able to considerably outperform the other strategies. This is not surprising, as it is equipped with additional knowledge in the form of the MIDI file.

There are some obvious differences in the beat tracking results of the individual Mazurkas caused by the musical reasons explained in [6]. First of all, all methods deliver

	SYNC	ONSET				PLP			
ID	P/R/F/A	P	R	F	A	P	R	F	A
M17-4	0.837	0.552	0.958	0.697	0.479	0.615	0.743	0.672	0.639
M24-2	0.931	0.758	0.956	0.845	0.703	0.798	0.940	0.862	0.854
M30-2	0.900	0.692	0.975	0.809	0.623	0.726	0.900	0.803	0.788
M63-3	0.890	0.560	0.975	0.706	0.414	0.597	0.744	0.661	0.631
M68-3	0.875	0.671	0.885	0.758	0.507	0.634	0.755	0.689	0.674
Mean:	**0.890**	0.634	0.952	**0.754**	0.535	0.665	0.806	**0.728**	0.729

	MIREX				Our Methods		
Method	DRP3	GP2	OGM2	TL	SYNC	ONSET	PLP
F	0.678	0.547	0.321	0.449	**0.890**	**0.754**	**0.728**

Table 3: Comparison of the beat tracking performance of the three strategies used in this paper and the MIREX 2009 results (see [2] for an explanation) based on the evaluation metrics Precision P, Recall R, F-measure F and the beat accuracy A.

the best result for M24-2. This piece is rather simple, with many quarter notes in the dominant melody line. M17-4 is the most challenging for all three trackers because of a frequent use of ornaments and trills and many beat positions that are not reflected in the dominating melody line. For the ONSET tracker, M63-3 constitutes a challenge ($A = 0.414$), although this piece can be handled well by the SYNC tracker. Here, a large number of notes that do not fall on beat positions provoke many false positives. This also leads to a low accuracy of PLP ($A = 0.631$).

Going beyond this evaluation on a piece-level, Fig. 4 and Fig. 5 illustrate the beat-level beat tracking results of our evaluation framework for the SYNC and PLP strategy, respectively. Here, for each beat $b \in \mathcal{B}$ of a piece, the bar encodes for how many of the performances of this piece the beat was not *strongly identified* (see Sect. 5). High bars indicate beats that are incorrectly identified in many performances, low bars indicate beats that are identified in most performances without problems. As a consequence, this representation allows for investigating the musical properties leading to beat errors. More precisely, beats that are consistently wrong over a large number of performances of the same piece are likely to be caused by musical properties of the piece, rather than physical properties of a specific performance. For example, for both tracking strategies (SYNC and PLP) and all five pieces, the first and last beats are incorrectly identified in almost all performances, as shown by the blue bars ($\mathcal{B}_2$). This is caused by boundary problems and adaption times of the algorithms.

Furthermore, there is a number of significant high bars within all pieces. The SYNC strategy for M68-3 (see Fig. 4) exhibits a number of isolated black bars. These non-event beats do not fall on any note-event ($\mathcal{B}_1$). As stated in Sect. 2, especially when dealing with expressive music, simple interpolation techniques do not work to infer these beat positions automatically. The same beat positions are problematic in the PLP strategy, see Fig. 5. For M30-2 (Fig. 4) most of the high bars within the piece are assigned to $\mathcal{B}_3$ (red). These beats, which coincide with ornaments such as trills, grace notes, or arpeggios are physically not well defined and hard to determine. For the Mazurkas, chords are often played on-beat by the left hand. However, for notes of lower pitch, onset detection is problematic, especially when played softly. As a consequence, beats that only coincide with a bass note or chord, but without any note being played in the main melody, are a frequent source

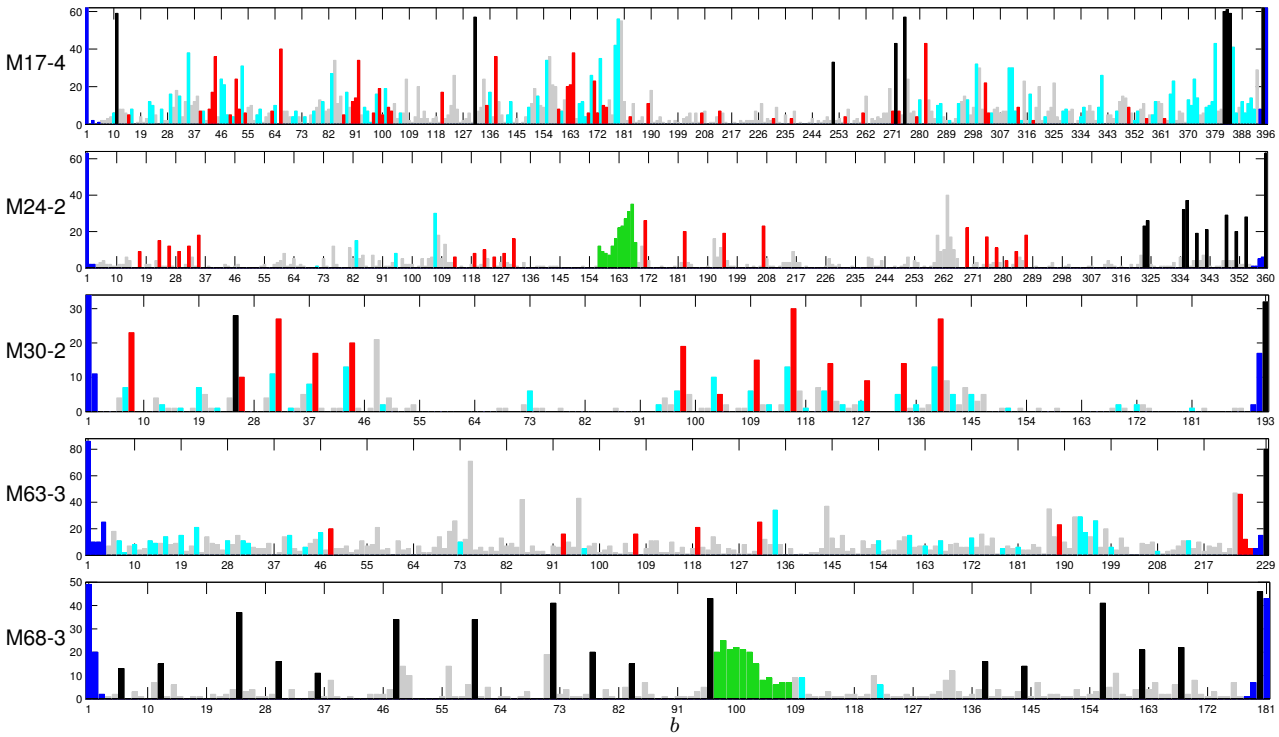

Figure 4: The beat error histogram for the synchronization based beat tracking (SYNC) shows for how many performances of each of the five Mazurkas a beat b is not identified. The different colors of the bars encode the beat class $\mathcal{B}$ a beat is assigned to, see Sect. 3.

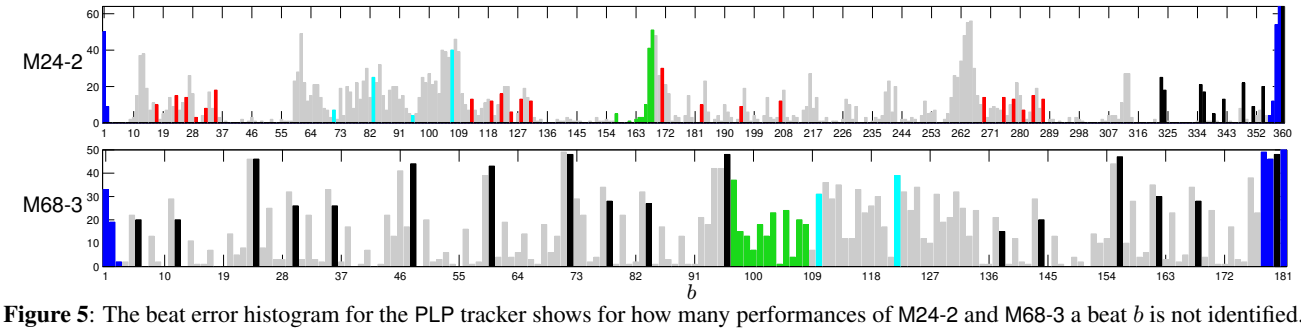

Figure 5: The beat error histogram for the PLP tracker shows for how many performances of M24-2 and M68-3 a beat b is not identified. The different colors of the bars encode the beat class $\mathcal{B}$ a beat is assigned to, see Sect. 3.

for errors. This is reflected by the cyan bars ($\mathcal{B}_3$) frequently occurring in M17-4 (Fig. 4). Finally, $\mathcal{B}_5$ (green) contains beats falling on consecutive repetitions of the same chord. This constitutes a challenge for the onset detection, especially when played softly. Both M24-2 and M68-3 exhibit a region of green bars that are incorrectly tracked by the SYNC (Fig. 4) and PLP (Fig. 5) trackers.

As mentioned in Sect. 4, PLP can not handle tempo changes well. As a consequence, many of the beat errors for PLP that are not assigned to any beat class (*e.g.*, M24-2 in Fig. 5, $b = [260 : 264]$) are caused by sudden tempo changes appearing in many of the performances. However, these are considered a performance-dependent property, rather than a piece-dependent musical property and are not classified in a beat class.

Table 4 summarizes the effect of each beat class on the piece-level results. Here, the mean beat accuracy is reported for each of the five Mazurkas, when excluding the beats of a certain class. For example, M30-2 contains many beats of $\mathcal{B}_3$. Excluding these ornamented beats from the evaluation, the overall beat accuracy increases from $A = 0.900$ to $A = 0.931$ for SYNC (Table 4 (left)) and from 0.788 to 0.814 for PLP (Table 4 (right)). The challenge of M68-3 however, are non-event beats ($\mathcal{B}_1$). Leaving out these beats, the accuracy increases from 0.875 to 0.910 for SYNC and from 0.674 to 0.705 for PLP.

Aside from musical properties of a piece causing beat errors, physical properties of certain performances make beat tracking difficult. In the following, we exemplarily compare the beat tracking results of the performances of M63-3. Fig. 6 shows the beat accuracy A for all 88 performances available for this piece. In case of the SYNC tracker, the beat accuracy for most of the performances is in the range of $0.8 - 0.9$, with only few exceptions that deviate significantly (Fig. 6(a)). In particular, Michalowski's 1933 performance with index 39 (pid9083-16, see [1]) shows a low accuracy of only $A = 0.589$ due to a poor condition of the original recording which contains a low signal-to-noise ratio and many clicks. The low accuracy ($A = 0.716$) of performance 1 (Csalog 1996, pid1263b-12) is caused by a high amount of reverberation, which makes a precise determination of the beat positions hard. The poor result of performance 81 (Zak 1951, pid918713-20) is caused by a detuning of the piano. Compensating

ID	$\mathcal{B}$	$\mathcal{B}\backslash\mathcal{B}_1$	$\mathcal{B}\backslash\mathcal{B}_2$	$\mathcal{B}\backslash\mathcal{B}_3$	$\mathcal{B}\backslash\mathcal{B}_4$	$\mathcal{B}\backslash\mathcal{B}_5$	$\mathcal{B}\backslash\mathcal{B}_*$
M17-4	0.837	0.852	0.842	0.843	0.854	0.837	0.898
M24-2	0.931	0.940	0.936	0.941	0.933	0.939	0.968
M30-2	0.900	0.900	0.903	0.931	0.905	0.900	0.959
M63-3	0.890	0.890	0.898	0.895	0.895	0.890	0.911
M68-3	0.875	0.910	0.889	0.875	0.875	0.887	0.948
Mean:	0.890	0.898	0.894	0.897	0.894	0.892	0.925

ID	$\mathcal{B}$	$\mathcal{B}\backslash\mathcal{B}_1$	$\mathcal{B}\backslash\mathcal{B}_2$	$\mathcal{B}\backslash\mathcal{B}_3$	$\mathcal{B}\backslash\mathcal{B}_4$	$\mathcal{B}\backslash\mathcal{B}_5$	$\mathcal{B}\backslash\mathcal{B}_*$
M17-4	0.639	0.650	0.641	0.671	0.593	0.639	0.649
M24-2	0.854	0.857	0.862	0.857	0.856	0.854	0.873
M30-2	0.788	0.788	0.794	0.814	0.772	0.788	0.822
M63-3	0.631	0.631	0.638	0.639	0.647	0.631	0.668
M68-3	0.674	0.705	0.689	0.674	0.678	0.674	0.733
Mean:	0.729	0.735	0.734	0.739	0.723	0.729	0.751

Table 4: Beat accuracy A results comparing the different beat classes for SYNC **(left)** and PLP **(right)**: For all beats $\mathcal{B}$, excluding non-event beats $\mathcal{B}_1$, boundary beats $\mathcal{B}_2$, ornamented beats $\mathcal{B}_3$, weak bass beats $\mathcal{B}_4$, constant harmony beats $\mathcal{B}_5$, and the union $\mathcal{B}_*$.

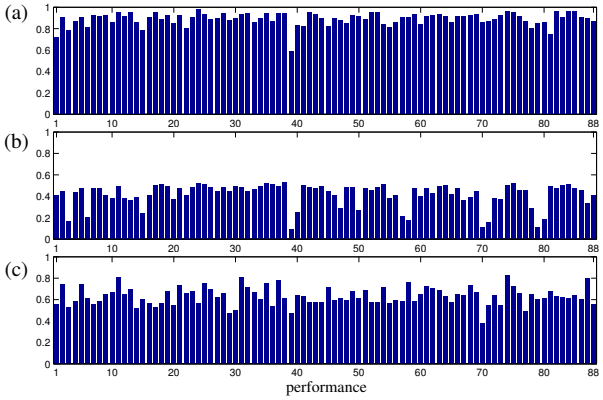

Figure 6: Beat accuracy A for the beat tracker SYNC **(a)**, ONSET **(b)**, and PLP **(c)** of all 88 performances of M63-3.

for this tuning effect, the synchronization results and thus, the beat accuracy improves from $A = 0.767$ to $A = 0.906$. As it turns out, ONSET tends to be even more sensitive to bad recording conditions. Again, performance 39 shows an extremely low accuracy ($A = 0.087$), however, there are more recordings with a very low accuracy (70, 71, 79, 80, 57, and 58). Further inspection shows that all of these recordings contain noise, especially clicks and crackling, which proves devastating for onset detectors and leads to a high number of false positives. Although onset detection is problematic for low quality recordings, the PLP approach shows a different behavior. Here, the periodicity enhancement of the novelty curve [11] provides a cleaning effect and is able to eliminate many spurious peaks caused by recording artifacts and leads to a higher beat accuracy. However, other performances suffer from a low accuracy (performances 29, 30, and 77). As it turns out, these examples exhibit extreme local tempo changes that can not be captured well by the PLP approach, which relies on a constant tempo within the analysis window. On the other hand, some performances show a noticeably higher accuracy (2, 5, 11, 31, 74, and 87). All oft these recordings are played in a rather constant tempo.

7. FUTURE DIRECTIONS

Our experiments indicate that our approach of considering multiple performances simultaneously for a given piece of music for the beat tracking task yields a better understanding not only of the algorithms' behavior but also of the underlying music material. The understanding and consideration of the physical and musical properties that make beat tracking difficult is of essential importance for improving the performance of beat tracking approaches. Exploiting the knowledge of the musical properties leading to beat errors one can design suited audio features. For example, in the case of the Mazurkas, a separation of bass and melody line can enhance the quality of the novelty curve and alleviate the negative effect of the ornamented beats or weak bass beats.

Acknowledgment. The first two authors are supported by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University. The raw evaluation was generated by the third author at AHRC Centre for the History and Analysis of Recorded Music (CHARM), Royal Hollway, University of London.

8. REFERENCES

[1] The Mazurka Project. http://www.mazurka.org.uk, 2010.

[2] MIREX 2009. Audio beat tracking results. http://www.music-ir.org/mirex/2009/index.php/Audio_Beat_Tracking_Results, 2009.

[3] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. E. P. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Trans. on Speech and Audio Processing*, 13(5):1035–1047, 2005.

[4] M. E. P. Davies, N. Degara, and M. D. Plumbley. Evaluation methods for musical audio beat tracking algorithms. Technical Report C4DM-TR-09-06, Queen Mary University, Centre for Digital Music, 2009.

[5] M. E. P. Davies and M. D. Plumbley. Context-dependent beat tracking of musical audio. *IEEE Trans. on Audio, Speech and Language Processing*, 15(3):1009–1020, 2007.

[6] S. Dixon. An empirical comparison of tempo trackers. In *Proc. of Brazilian Symposium on Computer Music*, pages 832–840, 2001.

[7] S. Dixon and W. Goebl. Pinpointing the beat: Tapping to expressive performances. In *Proc. of International Conference on Music Perception and Cognition*, pages 617–620, Sydney, Australia, 2002.

[8] A. Earis. An algorithm to extract expressive timing and dynamics from piano recordings. *Musicae Scientiae*, 11(2), 2007.

[9] S. Ewert, M. Müller, and P. Grosche. High resolution audio synchronization using chroma onset features. In *Proc. of IEEE ICASSP*, Taipei, Taiwan, 2009.

[10] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Trans. on Speech and Audio Processing*, 14, 2006.

[11] P. Grosche and M. Müller. A mid-level representation for capturing dominant tempo and pulse information in music recordings. In *Proc. of ISMIR*, pages 189–194, Kobe, Japan, 2009.

[12] 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*, 14(1):342–355, 2006.

[13] M. F. McKinney, D. Moelants, M. E. P. Davies, and A. Klapuri. Evaluation of audio beat tracking and music tempo extraction algorithms. *Journal of New Music Research*, 36(1):1–16, 2007.

[14] G. Peeters. Template-based estimation of time-varying tempo. *EURASIP Journal on Advances in Signal Processing*, 2007.

[15] C. S. Sapp. Hybrid numeric/rank similarity metrics. In *Proc. of ISMIR*, pages 501–506, Philadelphia, USA, 2008.