Exploring the Possibilities of Automated Annotation of Classical Music with Abrupt Tempo Changes

M. Ištvánek, Š. Miklánek

Department of Telecommunications, Faculty of Electrical Engineering and Communication,
Brno University of Technology, Technicka 12, 616 00 Brno, Czech Republic

E-mail: xistva02@vut.cz, xmikla12@vut.cz

Abstract—In this paper, we introduce options for automatic measure detection based on synchronization, beat detection, and downbeat detection strategy. We evaluate proposed methods on two motifs from the dataset of Leoš Janáček's string quartet music. We use specific user-driven metrics to capture annotation efficiency simulating a scenario where a musicologist has to use the output of an automated system to create ground-truth annotations on given recordings. In the case of the first motif, synchronization outperformed other methods by detecting most of the measure positions correctly. This procedure was also the most suitable for the second motif—it achieved a low number of correct detections, but the vast majority of transferred time positions belonged within the outer tolerance window. Therefore, in most cases, only shifting operations were needed resulting in higher annotation efficiency. Results suggest that the state-of-the-art downbeat tracking is not yet efficient for expressive music.

Keywords—beat tracking, classical music, downbeat detection, DTW, music information retrieval, music performance analysis, synchronization

1. INTRODUCTION

Music Information Retrieval (MIR) is a well-established interdisciplinary area that combines technical approaches and methods with musical analysis. The MIR researchers deal with many music-driven tasks, such as automatic detection of musical features and high-level parameters, user-centric semantic retrieval, recommendation systems, or transcription of audio recordings into symbolic representations [1]. In this paper, we focus on the automatic identification or detection of measure (bar) positions in string quartet recordings, which is closely related to the challenges of Musical Performance Analysis (MPA).

Measures are musically meaningful segments with defined metric patterns. Regarding western music notation, information about their exact position in a given musical hierarchy is automatically encoded in the corresponding score (sheet music). To obtain measure positions in structurally complex music such as string quartets, one needs to have a score available. Manual labeling and annotation is a time-consuming procedure but it is a common approach to obtain ground-truth data. However, recent developments of machine learning methods may change this workflow.

One of the most established topics in MIR is beat tracking or beat detection[1]. A standard beat tracking system outputs a vector of time positions that correspond to individual beats in a given music recording. In our case, we want to obtain only the first beat of each measure—such detectors do not usually distinguish the beat index within measures. Therefore, downbeat tracking systems have been developed which, together with the time position of beats[2], also estimate their position in a musical structure. The second option is a strategy based on the synchronization procedure. The general goal of music synchronization is to establish an alignment between musically corresponding time positions (measures in this case) of the same piece (e.g., audio-to-score or audio-to-audio alignment) [2].

In this paper, we focus on computer-generated annotations and test the state-of-the-art offline beat and downbeat tracking for measure detection on chamber music. We compare the detectors with the music synchronization technique and evaluate all methods by a user-driven metric.

[1] In the context of this paper, we use the terms beat tracking and beat detection interchangeably.
[2] The system outputs the probability of beats and downbeats separately.

2. METHODS

2.1. Dataset

First, we introduce our dataset, which consists of two separate motifs from Janáček's *String Quartet No. 1 "Kreutzer Sonata", JW 7 No. 8* and *String Quartet No. 2 "Intimate Letters", JW 7 No. 13*, respectively. Figure 1 shows a score for the first motif. This motif contains 11 measures of the first movement. At the beginning, all strings except violoncello play *con sordino*[3] and the second violin uses *finger tremolo*[4] technique which may blur the starting point of the second bar. Furthermore, the overall dynamics is higher for the upbeat than for the downbeat. After *subito forte* (suddenly loud), the tempo changes rapidly with possible local deviations based on individual interpretation. The second motif contains 10 measures of the second movement. It is even more complex within the metric structure with many accents in the middle of measures. We selected these excerpts for their various tempo, challenging structure, and expressive nature. We gathered 17 different interpretations for each motif and carefully annotated all ground-truth measure positions. In our experiments, the first recording from both motifs by the Belcea Quartet (year of recording 2018) was selected as a reference. The remaining recordings were used for testing purposes.

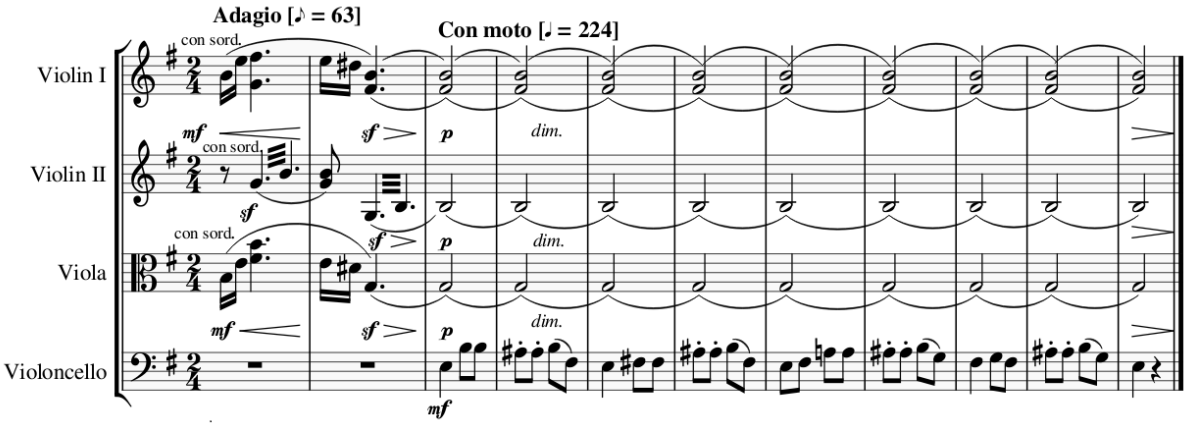

Figure 1: The score for the first motif of our dataset.

2.2. Beat and downbeat detection

Beat tracking systems provide time positions of computed beats for any given music recording. In the case of neural network-based approaches, their output is usually an activation function—its value within a specified feature rate is related to the novelty function or confidence of beat occurrence. Then, peak-picking methods or probabilistic and statistical methods, such as Conditional Random Fields or Dynamic Bayesian Networks (DBN), are often used.

In this paper, we use a beat detector based on the variant of Recurrent Neural Network (RNN) [3] in combination with DBN [4] and a downbeat detector also based on RNN [5] and DBN but with different settings and functionality. We kept the default settings for the DBN with a range of possible tempo detection between 55 and 215 BPM (Beats Per Minute). This system will demonstrate the problematic part of beat tracking when applied to expressive chamber music.

The DBN estimator of downbeats outputs two vectors of data. The first one contains time positions of beats and the second their index within a measure—e.g., output vector $B = [2.5, 3]$ shows the third beat of a measure in the time of 2.5 s. Thus, we have selected only those beats that correspond to the first position of each measure creating a downbeat sequence. Ideally, the output of this modified detector should produce only the beginning of each measure and follow the ground-truth data structure. We also added the prior knowledge (2 or 3 beats per bar) about the metric structure of selected motifs into the detector.

2.3. Synchronization method

The second option to obtain time positions of measures is a synchronization procedure. This is a common approach in MPA due to its advantages. In our experiment, we use an alignment method based on

[3] A technique that uses a "mute pad" to soften the produced sound.

[4] The player uses fingers to alternate rapidly between two notes.

a variant of Dynamic Time Warping (DTW), called Memory-restricted Multiscale DTW (MrMsDTW) that is faster and may provide a better synchronization accuracy [6]. First, we compute chroma energy normalized statistic (CENS) features [7] of reference and target recording with a resolution of 50 features per second. MrMsDTW is applied to compute a cost-minimizing alignment between both CENS matrices and the resulting warping path is limited to be strictly monotonic by postprocessing. The ground-truth annotations are then transferred from the reference to the target recording by the resulting warping path.

This strategy has an advantage over the automated detectors—there will be always the right number of measures detected. The question is whether chroma features contain enough information for alignment to work accurately e.g. in music structures, where there is almost no new information present, but measure number increases.

2.4. User-driven metric

Each machine annotation of musical content usually ends with a certain number of mislabeled time positions. Either the desired time point may not appear in the machine annotation at all, or it is misplaced. In [8], the authors introduced the *annotation efficiency* (*ae*) metric, which is based on how much effort a user has to exert to manually correct detections by shifting, deleting, or inserting time positions. The insert and delete operations correspond to the counts of false negatives and false positives, respectively. The shifting should theoretically be counted twice, once as a false positive and the second time as a false negative. In practice, however, it is more sensible to count this operation separately and prioritize it over deletion and insertion, since it is the most common correction performed by the user.

The process of calculating the *ae* metric is as follows. First, an inner tolerance window of ±70 ms is created around each ground truth annotation. Then, the true positives (unique detections), t^+, are counted. Detections that match ground truth annotations are removed from further calculations and incorrect detections are marked as candidates to be shifted or removed. For each remaining annotation, an outer tolerance window of ±1 s is then created to search for the closest detection that does not match the ground truth. If there is a detection in this window, it is marked as shift. After the analysis of unaccounted detections, the number of shifts s is calculated. The remaining annotations correspond to false negatives, f^-, with leftover detections marked for deletion and counted as false positives, f^+. The *ae* metric is defined by the following equation:

$$ae = t^+/(t^+ + s + f^+ + f^-). \tag{1}$$

3. RESULTS

First, we transferred the ground-truth annotations based on the DTW alignment method, then calculated beats and downbeats as described in section 2. Figure 2 shows the user-driven metric computation and the pipeline with all possible operations for one of the recordings. Operations are marked with different colors to increase the readability. We kept the same inner and outer tolerance window as the original beat tracking evaluation in [8].

In this case, the synchronization procedure outperforms all other methods. The final synchronized positions are not in the exact time positions, however, they mostly fit into the inner tolerance window. The value of beat confidence for the downbeat tracker was in the first motif too low—measures that have ambiguous nature were not detected at all, measures of a faster-paced segment with an abrupt change of rhythmic structure were partially omitted. On the other hand, the RNN beat tracker detected many false positives. The DBN method fills the positions between confident output beats based on their past and future occurrence—this method can work well with small deviations of tempo but fails when the rhythmic and metric structure is unpredictable and highly changing.

Table I shows the sum of all operations and annotation efficiency, recall, and precision for both motifs and each method. Synchronization outperformed other methods for the first motif with 142 correct detections and only 36 additional operations. In the second scenario, however, the beat tracking captured the highest number of correct measure positions. Although the synchronization method achieved the lowest number of all corrections and the best annotation accuracy, recall and precision remained low. Recall and precision scores may give the impression that beat and downbeat detection are more suitable tools for automatic detection of measure positions in a complex structure, but a number of deletion operations reveal that they are in fact counterproductive in this scenario. None of the proposed methods was successful considering only the second motif.

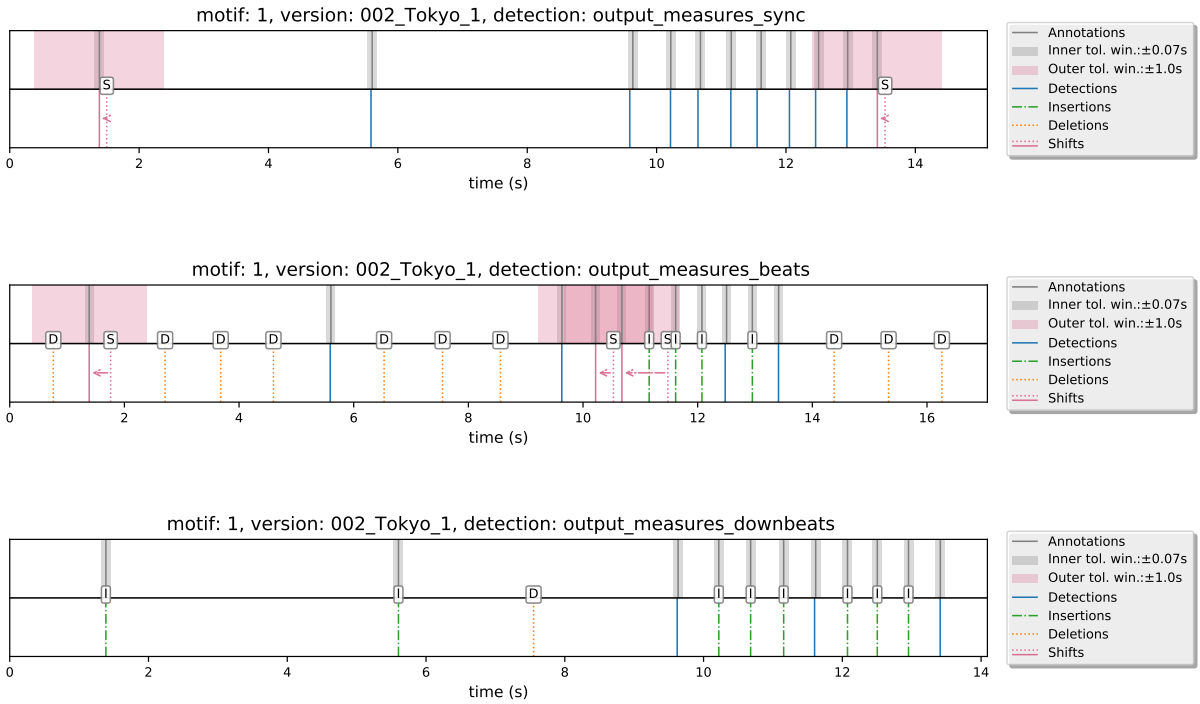

Figure 2: The user-driven metric for synchronization, beat tracking, and downbeat tracking strategy; evaluation of the Tokyo Quartet recording, first motif.

motif 1 (176 measures in total)								
method	∑ det	∑ ins	∑del	∑ shf	∑ ops	*ae* (mean)	*R* (mean)	*P* (mean)
beat tracking	70	42	167	64	273	0.208	0.375	0.225
downbeat tracking	33	116	35	27	178	0.155	0.176	0.329
synchronization	**142**	2	2	32	**36**	**0.799**	**0.727**	**0.727**
motif 2 (160 measures in total)								
method	∑ det	∑ ins	∑del	∑ shf	∑ ops	*ae* (mean)	*R* (mean)	*P* (mean)
beat tracking	**67**	0	516	93	609	0.101	**0.356**	0.087
downbeat tracking	38	61	54	61	176	0.196	0.194	**0.219**
synchronization	38	11	11	111	**133**	**0.224**	0.156	0.156

Table I: The number of detections, insertions, deletions, shifts, and total corrections, annotation efficiency, recall, and precision for each motif and method.

4. DISCUSSION

The synchronization procedure, even if it always detects the correct number of measures, relies only on chroma features, their resolution, and DTW accuracy. The ground-truth annotations may not be always precise—the resulting warping path can transfer reference time positions with some deviations. It depends, e.g, on the harmonic structure, occurrence of onsets, or the ADSR envelope of given instruments. If we tolerate larger deviance (such as 100 ms), almost all annotations will be transferred correctly.

The beat detector has shown an experimental role in illustrating the function of predicting the rhythmic structure and beat occurrence. In the second motif, it achieved the best recall and number of correct measure positions. However, it also detected too many false positives; that would be true even if ground-truth annotations were based on beat positions. The method of filling in beats, even in places where there is no underlying information, can work well in simpler musical structures without significant changes in rhythm and meter. Furthermore, detectors are usually trained on specific audio datasets, for which there are manual ground-truth annotations available—string quartet music is not among them.

The downbeat detector was not sensitive enough or predicted false beat indexes, although it contained prior knowledge about the musical structure (see section 2.2). Table I shows that so far, the only valid option for expressive string quartet music with many abrupt tempo changes, local tempo deviations, and weak onset and beat positions, is the synchronization strategy. Its accuracy can be improved by the choice of additional features for the alignment procedure. In this case, however, the ground-truth annotations are always needed.

5. CONCLUSION

In this contribution, we proposed and evaluated different methods of obtaining measure positions in string quartet music. We first created reference ground-truth data and then compared the synchronization method, beat tracking, and downbeat tracking based on a specific user-driven metric. This metric allows us to calculate the number of operations that one needs to make to obtain the ground-truth annotation of measure positions. We tested different strategies on two carefully selected string quartet motifs from Leoš Janáček's compositions. Both proposed segments are musically challenging, they contain many weak onset positions, ambiguous beats, and abrupt tempo and rhythm changes. Results suggest that the synchronization method is superior to all other possible options. Beat and downbeat tracking approaches are not yet efficient on very expressive pieces of classical music.

ACKNOWLEDGMENT

This work was supported by the "Identification of the Czech origin of digital music recordings using machine learning" grant, which is realized within the project Quality Internal Grants of BUT (KInG BUT), Reg. No. CZ.02.2.69 / 0.0 / 0.0 / 19_073 / 0016948 and financed from the OP RDE.

REFERENCES

[1] M. Schedl, E. Gómez, and J. Urbano, "Music Information Retrieval: Recent Developments and Applications," *Foundations and Trends® in Inform. Retr.*, vol. 8, no. 2-3, pp. 127–261, 2014, doi: 10.1561/1500000042. [Online]. Available: https://ieeexplore.ieee.org/document/8187204

[2] C. Weiß, V. Arifi-Muller, T. Prätzlich, R. Kleinertz, and M. Muller, "Analyzing Measure Annotations for Western Classical Music Recordings," in *Proc. 17th Int. Society for Music Information Retrieval Conf. (ISMIR 2016)*, 2016, pp. 517–523. [Online]. Available: https://archives.ismir.net/ismir2016/paper/000079.pdf

[3] S. Böck and M. Schedl, "Enhanced Beat Tracking With Context-Aware Neural Networks," in *Proc. 14th Int. Conf. on Digital Audio Effects (DAFx-11)*, Sep. 19–23, 2011, pp. 135–139. [Online]. Available: http://www.dafx.de/paper-archive/2011/Papers/31_e.pdf

[4] F. Krebs, S. Böck, and G. Widmer, "An Efficient State-Space Model for Joint Tempo and Meter Tracking," in *Proc. 16th Int. Society for Music Information Retrieval Conf. (ISMIR 2015)*, 2015, pp. 72–78. [Online]. Available: https://archives.ismir.net/ismir2015/paper/000239.pdf

[5] F. Krebs, S. Böck, and G. Widmer, "Joint Beat and Downbeat Tracking with Recurrent Neural Networks," in *Proc. 17th Int. Society for Music Information Retrieval Conf. (ISMIR 2016)*, 2016, pp. 255–261. [Online]. Available: https://archives.ismir.net/ismir2016/paper/000186.pdf

[6] T. Prätzlich, J. Driedger, and M. Müller. "Memory-Restricted Multiscale Dynamic Time Warping," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 569–573. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/mueller/publications/2016_PraetzlichDriedgerMueller_MrMsDTW_ICASSP.pdf

[7] M. Müller, F. Kurth, and M. Clausen, "Audio Matching Via Chroma-based Statistical Features," in *Proc. 6th Int. Conf. on Music Information Retrieval (ISMIR 2005)*, 2005, pp. 288–295. [Online]. Available: https://archives.ismir.net/ismir2005/paper/000019.pdf

[8] A. S. Pinto, I. Domingues, and M. E. P. Davies, "Shift If You Can: Counting and Visualising Correction Operations for Beat Tracking Evaluation," in *Extended Abstracts for the Late-Breaking Demo Session of the 21st Int. Society for Music Information Retrieval Conf. (ISMIR 2020)*, 2020. [Online]. Available: https://program.ismir2020.net/lbd_421.html