MODELING AND ESTIMATING LOCAL TEMPO: A CASE STUDY ON CHOPIN'S MAZURKAS

Hendrik Schreiber, Frank Zalkow, Meinard Müller
International Audio Laboratories Erlangen, Germany
{hendrik.schreiber,frank.zalkow,meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

Even though local tempo estimation promises musicological insights into expressive musical performances, it has never received as much attention in the music information retrieval (MIR) research community as either beat tracking or global tempo estimation. One reason for this may be the lack of a generally accepted definition. In this paper, we discuss how to model and measure local tempo in a musically meaningful way using a cross-version dataset of Frédéric Chopin's Mazurkas as a use case. In particular, we explore how tempo stability can be measured and taken into account during evaluation. Comparing existing and newly trained systems, we find that CNN-based approaches can accurately measure local tempo even for expressive classical music, if trained on the target genre. Furthermore, we show that different training–test splits have a considerable impact on accuracy for difficult segments.

1. INTRODUCTION

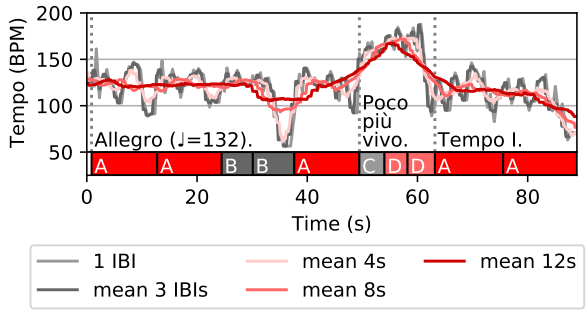
While *global* tempo is well defined for music with little or no tempo variability [1], this is less so the case for *local* tempo, especially for expressive classical music. Composer markings like *rubato* (expressive, local tempo change) or *ritardando* (slow down) indicate continuous or even abrupt tempo changes, leading to one or more segments with stable tempi and segments of tempo instability in between. Figure 1, for example, shows tempo markings for Frédéric Chopin's Mazurka Op. 68, 3 (details are discussed in Section 2). Naïvely, one may model local tempo for such a piece as one of two extremes: at the *micro level*, as an instantaneous value, e.g., as the Inter Beat Interval (IBI) between two consecutive beats, or at the *macro level*, by averaging the number of beats over a longer period of time. For expressive music, both approaches have disadvantages. IBIs exhibit a large variance, and averaging beat counts may underestimate the tempo, because expression leads more often to longer than shorter IBIs [2]. Repp therefore attempts to find a definition for the *basic tempo* [3], i.e., the implied tempo the instantaneous tempo varies around. In [2], he suggests to derive the basic tempo from the first quartile of eighth-note Inter Onset Intervals (IOIs). Similarly, Dixon [4] proposes IOI clustering, using centroids as tempo hypotheses. Grosche and Müller [5] propose yet another approach by defining local tempo as the mean of three consecutive IBIs, which is identical to using Inter Measure Intervals (IMIs) for pieces in ¾ time. The same method is also used by Chew and Callender [6]. In summary, local tempo is usually modeled by aggregating local pulse information, but there appears to be no clear consensus on how. Even though local tempo estimates are popular intermediate features for beat trackers (e.g., [7,8]), few works explicitly estimate and evaluate local tempo estimates. Peeters [9] simply measures whether 75% of the estimated local tempi match the annotated global tempo. In subsequent work [10], he compares the median of local tempi with a global ground truth. A similar approach is taken in [11]—after beat tracking, the median IBI is used as global tempo and then evaluated. Similar to global tempo evaluation, Grosche and Müller [5] compute the accuracy of their IMIs allowing a 4% tolerance and certain integer factors. Schreiber and Müller [12] only provide visualizations for local tempo estimates. To our knowledge, there is no commonly accepted evaluation procedure. Even less researched than local tempo is tempo *stability*, usually only referred to as a precondition for global tempo estimation [1]. Grosche et al. [13] mention that beat trackers tend to have problems with the first and last few beats of Mazurkas due to boundary problems, and observe increased error-levels caused by sudden tempo changes, but as far as we know no measure for local tempo stability has been proposed.

Work	Measures	Beats	Recordings
Op. 17, 4	132	396	62
Op. 24, 2	120	360	64
Op. 30, 2	65	193	34
Op. 63, 3	77	229	88
Op. 68, 3	61	181	50

Table 1: Dataset overview [13]: Number of measures, beats, recordings for five Chopin Mazurkas.

Modeling local tempo, determining its stability, and estimating it automatically from audio are problems at the intersection of music information retrieval (MIR) and computational musicology. We believe that all three problems have to be solved together in order to provide use-

(a) Local tempo depending on selection



(b) Local tempo depending on aggregation

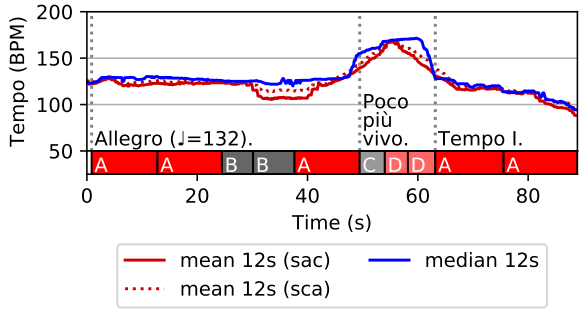


Figure 1: Local reference tempo depending on **(a)** selection and **(b)** aggregation functions for Op. 68, 3 (Cohen, 1997) with section boundaries and score tempo markings.

ful tools for computational music performance analysis (MPA) [14]. Such tools can, for example, be used to determine how well a given performance matches the score—similar to how it has been done for dynamics [15]. Studies like [16], comparing relative local tempo variations within performance collections, could be enhanced by using absolute tempo information.

Working towards this goal, we investigate how to model local tempo (Section 2) and tempo stability (Section 3) for expressive music using Mazurkas by Chopin. As our main contribution, we estimate local tempi using neural network-based approaches, adapt these approaches to our use case, and explore their behavior and potential (Section 4). In our evaluation, we focus on identifying error classes and sources, and in particular the effect of stability. In Section 5, we discuss our findings and draw conclusions.

2. LOCAL TEMPO

Cancino-Chacón et al. [17] see the *global tempo* of a performance as the approximate rate at which musical events happen throughout that performance. In contrast, *local tempo* refers to the rate of events within a smaller time window and can therefore be regarded as local deviation from the global tempo. In accordance with this definition, we are interested in a musically meaningful, single-value description of a segment of limited length. We can define this length musically, e.g., as three consecutive IBIs [5, 6], or physically, e.g., as 6 s or 8 s segments [9,10]. In either case, we first *select* beat events, because they fall into a time span, and then *aggregate* them. For example, we may use the mean or the median of all IBIs falling into a 4 s interval. One purpose of this aggregation is to be able to largely ignore *expressive timing*, which can be defined as deviations of individual beat events from the local tempo [17], e.g., rolled or arpeggiated chords [18]. Note that, in this work, we are not attempting to find the *most* suitable selection and aggregation functions (see [3]), but merely discuss options and aim to establish a framework that can be used for such an endeavor. To illustrate different choices, we use Chopin's Op. 68, 3 (piano: Cohen, 1997) as an example. It is one of over 2,700 recordings of 49 Mazurkas by Chopin collected by the Mazurka Project. [1] Of all collected recordings, 298 recordings of five Mazurkas have been manually beat-annotated [19]. We refer to this subset as the *Mazurka-5* dataset. It contains between 34 and 88 different versions of each of the five Mazurkas (Table 1).

Our example, Op. 68, 3 (Cohen, 1997), consists of four different musical sections A to D (Figure 1). While the score does not contain section markers, [2] it explicitly specifies two tempo changes: at the start of section C from *Allegro, ma non troppo (♩=132)* (fast, but not too fast), to *Poco più vivo* (a little more lively), and back to *Tempo I* after the second D-section. Figure 1a depicts the effects of different selection functions using the mean for aggregation. We see that defining local tempo as individual IBIs leads to very high variance. Using three consecutive IBIs smoothes the tempo curve slightly. The shown tempo curves based on 4 s, 8 s, and 12 s segments progressively lead to less variance. While the 4 s tempo curve still follows the phrasing closely (distinct minima at the end of each musical section), this is less the case for the curves based on longer segments. This is especially obvious at the end of the 2nd B section at 38 s.

Figure 1b shows the differences between using mean and median as aggregation function. The tempo curves for mean show local over-smoothing in transitional sections, leading to a triangular shape in the more lively CDD-section from $50 - 60$ s. Because of the edge-preserving property of median-filtering, the median curve captures sudden tempo changes better. The CDD-section resembles a rectangle, i.e., a high tempo plateau. At the same time, the local minimum at the end of the 2nd B disappears. Thus, the median curve corresponds to the composer's markings.

So far we first selected IBIs, aggregated them, and then converted the result to BPM (selection → aggregation → conversion: sac). As an alternative, we could have first converted IBIs to BPM and then aggregated them (selection → conversion → aggregation: sca). When using mean, the result is not the same. For sections with changing tempo (Figure 1b, $30 - 70$ s), local tempo values are lower when we first average and then convert (sac, solid red line) as opposed to first convert and then average (sca, dotted red line). Note that the median is unaffected by this issue.

[1] http://www.mazurka.org.uk/

[2] Section markers were added by us to allow an easier discussion.

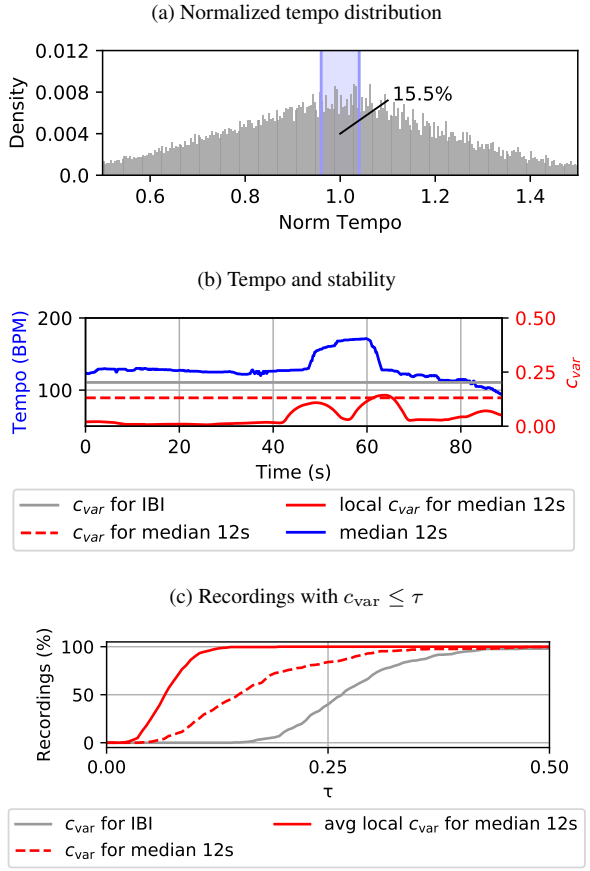


Figure 2: **(a)** Per recording normalized tempo distribution with percentage of values between 0.96 and 1.04 (light-blue area). **(b)** Local tempo (blue line) and stability (c_{var}) for Op. 68, 3 (Cohen, 1997). c_{var} based either on IBIs (gray line), the (sampled) median tempo over 12 s intervals (dashed red line), or the averaged local c_{var} over 12 s segments of median tempi (solid red line). **(c)** Percentage of recordings with $c_{\text{var}} \leq \tau$.

3. TEMPO STABILITY

For the evaluation of global tempo estimation one typically requires recordings with approximately constant tempi [1], i.e., a certain degree of tempo stability. Since local tempo estimation is in fact global tempo estimation for very short segments, we seek to quantify local tempo stability in order to conduct an informed evaluation of our experiments in Section 4. As a first approach to describe tempo stability quantitatively on the intra-track level, we convert all *Mazurka-5* IBIs to tempo values and normalize them by dividing by their respective track's average. Figure 2a depicts the resulting normalized histogram. [3] Only 15.5% of the *Mazurka-5*'s normalized tempi are in the interval between 0.96 and 1.04—the often used ±4% tolerance interval for stable tempi [1]. For comparison, 90.9% of the *Ballroom* [1,20] dataset's normalized tempi are in the same interval. Obviously, the two datasets are very different w.r.t. intra-track tempo stability.

While the ±4% interval is illustrative when categorizing stable vs. unstable, it is a rather arbitrary threshold. Arguably, the standard deviation of a track's normalized tempi is better suited to describe intra-track tempo variability. It is identical to the coefficient of variation, [4] which is defined as the ratio between the standard deviation σ and the mean μ:

$$c_{\text{var}} = \frac{\sigma}{\mu}. \quad (1)$$

We show this IBI-based c_{var}-value for our example Op. 68, 3 (Cohen, 1997) as a horizontal gray line in Figure 2b. As discussed in Section 2, instantaneous tempo values like IBIs tend to overestimate the variance of a musically meaningful local tempo for expressive music. From a musical point of view, it is therefore more appropriate to analyze tempo stability of Mazurkas not based on individual IBIs, but on the basic tempo, which—*for the purpose of this discussion*—we approximate with the median tempo over 12 s segments (Figure 2b, blue line). Sampling the local median tempo allows us to calculate an arguably more appropriate c_{var} (Figure 2b, dashed red line), which lies well below the gray line, indicating higher stability. This however, still ignores the fact that Mazurkas may contain multiple sections with stable but different tempi. We can take this into account by calculating local coefficients of variation for short segments of the median-based tempo curve. The solid red curve in Figure 2b shows the results for overlapping 12 s-segments. For most of the recording it is very low. Only in the transitional regions, at the beginning and end of the CDD-section, we see higher values. Note that by averaging the local c_{var} we can obtain a measure for intra-*segment* stability, while the two track-level c_{var} measures represent intra-*track* stability. Figure 2c depicts how many recordings of our dataset have a c_{var} below a threshold τ for all three ways of calculating it. The comparison shows that for *Mazurka-5* intra-segment variability is far smaller than intra-track variability.

4. EXPERIMENTS

We now investigate how different local tempo estimation systems perform when tested with *Mazurka-5*. We consider the following systems: The RNN-based beat tracking system `Böck` [5] [21] (estimated beats are aggregated identically to the ground truth), the CNN-based tempo estimation system `DeepTemp` [6] [22], and the system `DT-Maz`, which is set up identically to `DeepTemp`, but has been trained on *Mazurka-5* recordings instead of Pop/Rock, EDM, and Ballroom music. Based on our observations in Section 2 and informal experiments with several segment lengths, we model the local tempo with median-aggregated IBIs from 11.9 s segments. [7] As mentioned in Section 2, we do not claim that this is the best possible selection or aggregation, but a reasonable configuration.

[3] The comb pattern is a consequence of the 10 ms resolution of the original annotations.

[4] Also known as CV or relative standard deviation (RSD).

[5] `https://github.com/CPJKU/madmom` with default parameters.

[6] Scaled with model sizing parameter $k = 16$, see [22] for details.

[7] We chose 11.9 s instead of the previously used 12 s for practical reasons. The system `DeepTemp` is already trained on 11.9 s.

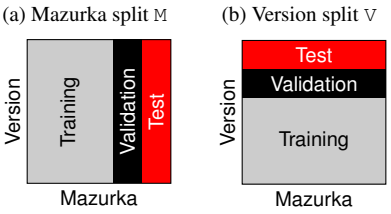


Figure 3: Dataset splitting into training, validation, and test sets.

4.1 Setup

We trained DT-Maz from scratch [8] on *Mazurka-5* recordings using 5-fold cross validation with two different kinds of splits, M for Mazurka and V for version (or performance). For M, each split contains all versions of one Mazurka (Figure 3a). For V, each split consists of a disjoint 5th of all versions of each of the five Mazurkas (Figure 3b). During training, three splits were used as training data and one for validation. The remaining 5th split was used for testing. Each split was used exactly once for validation or testing. We refer to the models trained on M-splits as $\texttt{DT-Maz}_\texttt{M}$ and to the V-split models as $\texttt{DT-Maz}_\texttt{V}$. The employed training procedure was very similar to [12]. Audio is first converted to mel-magnitude-spectrograms. Then samples with the dimensions $F \times T$ are used as network input. $F = 40$ being the number of frequency bins covering the frequency range $20 - 5{,}000$ Hz, and $T = 256$ being the number of time frames with a length of 46 ms per frame, corresponding to 11.9 s. We further use scale & crop data augmentation [12] with time scale factors drawn from $\mathcal{N}(1,\, 0.1)$, but limited to $[0.7, 1.3]$ to avoid extreme distortions. After augmentation, samples are standardized to zero mean and unit variance. Like [12], we use categorical crossentropy as loss, because we cast tempo estimation as a classification problem, predicting tempo as one of 256 linearly spaced classes ranging from 30 to 255 BPM. [9] Adam [23] is used as optimizer with a batch size of 32 and an initial learning rate of 0.001. The rate is halved once the validation loss stops improving and training is continued with the best performing model up to that point (stepwise annealing). We repeat this at most 10 times. If reduction does not lead to a lower validation loss three times in a row, training is stopped. To avoid overfitting to longer recordings, we ensure that samples from all training recordings are presented with the same frequency.

4.2 Evaluation

To evaluate, we estimate the tempo for a sliding segment with length 11.9 s (256 frames) and a hop size of 186 ms (4 frames) over all recordings. As metric we use ACC_1 (tempo accuracy) and ACC_2 (accuracy allowing so-called *octave errors*, i.e., estimates that are wrong by the factor 2, ½, 3 or ⅓) from the global tempo estimation

[8] Transfer learning on the DeepTemp model led to similar results.

[9] For an eventual performance analysis, one may want to rescale estimates logarithmically, as suggested in [6].

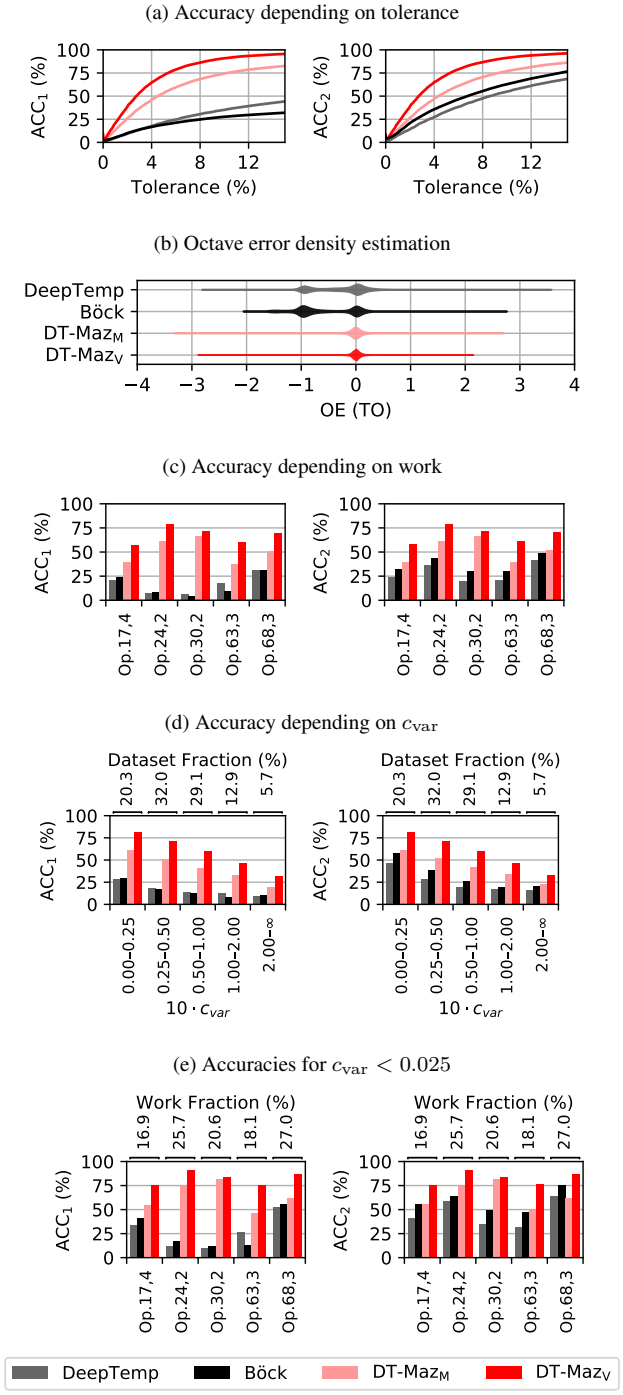


Figure 4: **(a)** Local ACC_1 and ACC_2 depending on accuracy tolerance. **(b)** Density estimation for OE. **(c)** Local ACC_1 and ACC_2 for the five Mazurkas. **(d)** Local ACC_1 and ACC_2 considering c_{var} ranges. **(e)** Accuracies for segments with $c_{\mathrm{var}} < 0.025$.

task [1], which are meant for music with low intra-track tempo variability. This is reasonable, because we apply the metric locally for each segment, so that the tolerance does not have to correspond to intra-track, but to intra-segment variability, and as we have shown in Section 3, intra-segment variability is relatively low. Nevertheless, we consider the typical 4% tolerance an arbitrary threshold and therefore plot accuracy values for the tolerance interval

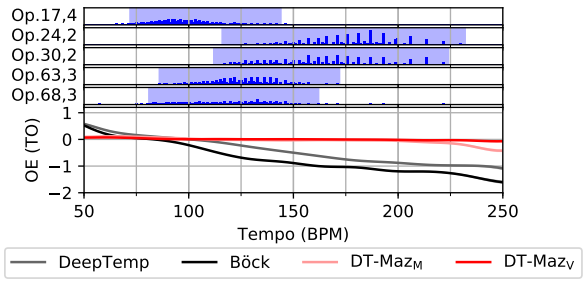


Figure 5: **(top)** Sweet octaves in light-blue. Local tempo histograms in dark-blue. **(bottom)** Estimates of generalized additive models fit to OE/tempo-pairs

$0-15\%$ in Figure 4a. For both variants of DT−Maz, ACC_1 values are higher than for the other systems, regardless of tolerance. Not surprisingly, ACC_1 values are also generally higher for higher tolerances.[10] The best performing system for the tolerances 4%, 8%, and 12% is $\mathrm{DT-Maz_V}$ with remarkable 64.6%, 86.4%, and 93.5%. The worst performing system is Böck, with 16.8%, 24.8%, and 29.7%. For ACC_2 the best performing system is also $\mathrm{DT-Maz_V}$ with 64.8%, 86.8%, and 94.0%, and the worst performing system is DeepTemp with 27.3%, 47.5%, and 61.2%. In the following paragraphs we discuss the most prominent errors, namely octave errors, tempo stability related errors, and problems with specific musical properties.

Tempo Octave. Using violin plots, Figure 4b depicts kernel density estimates (KDE) of the octave error OE defined as

$$\mathrm{OE}(y, \hat{y}) = \log_2 \frac{\hat{y}}{y}, \tag{2}$$

with $y, \hat{y} \in \mathbb{R}_{>0}$ as the ground truth and estimate. Identifiable by the very dense section around -1 Tempo Octaves (TO), DeepTemp and Böck suffer most from underestimating the actual tempo. As Figure 4c shows, octave errors are not evenly distributed among the five Mazurkas. Op. 24, 2 and Op. 30, 2 are much more affected than the other three. This can be partially explained by the fact that on average versions for Op. 24, 2 and Op. 30, 2 are much faster. Their *sweet octaves* [24], i.e., the tempo octave most tempo values are in, are $[116, 232)$ and $[112, 224)$ BPM, while the sweet octaves for Op. 17, 4, Op. 63, 3, and Op. 68, 3 are $[72, 144)$, $[86, 172)$, and $[81, 162)$ BPM (Figure 5, top). A closer investigation shows that for the tested Mazurkas, both DeepTemp and Böck lean towards negative octave errors for higher tempi, revealing an *octave bias* [24]. This is visualized in Figure 5, bottom. It shows the estimates of generalized additive models (GAM) that are fit to measured OE per reference tempo. It illustrates what kind of estimation error we can expect depending on a given true tempo. For tempi greater than 100 BPM, Böck and DeepTemp tend to suffer from negative octave errors.

Stability. Figure 4d shows that accuracy is higher when considering only segments with low c_{var}-values—our proxy for tempo variability. When only considering relatively stable segments with $c_{\mathrm{var}} < 0.025$ (Figure 4e), the accuracy scores for all five Mazurkas increase substantially. The comparison of $\mathrm{DT-Maz_M}$ and $\mathrm{DT-Maz_V}$ shows that $\mathrm{DT-Maz_M}$ performs much worse for some Mazurkas (Op. 17, 4, Op. 63, 3, and Op. 68, 3) than $\mathrm{DT-Maz_V}$. Apparently, differences in stability cannot fully explain differences in accuracy for the five works.

Musical Properties. We have seen in Figure 4e that even for stable segments, $\mathrm{DT-Maz_V}$ performs better than $\mathrm{DT-Maz_M}$. To find out why, we exploit beat annotations for each recording of the five Mazurkas. They allow us to compute stability and the absolute octave error $|\,\mathrm{OE}\,|$ for 11.9 s segments with a beat at their center, i.e., stability and error on a musical time axis. Using musical time, we can summarize errors and stability measures *cross-version* by averaging per beat over all recordings of a given Mazurka. Figure 6 shows the results for Op. 17, 4 and as expected, the $\overline{c_{\mathrm{var}}}$-curve roughly correlates with errors by both $\mathrm{DT-Maz_M}$ and $\mathrm{DT-Maz_V}$. For $\mathrm{DT-Maz_M}$ we see four additional peaks around beats 42, 89, 162, and 305 (highlighted in light-blue). These peaks loosely correlate with the occurrence of dense mixtures of ornamented beats (red) and weak bass beats (cyan), i.e., piece-dependent musical properties (classification from [13]), which are apparently the main reason for the difference in accuracy. Trained on the V-split, $\mathrm{DT-Maz_V}$ was able to learn piece-specific musical properties and generalize them across versions. This implies that expecting $\mathrm{DT-Maz_M}$'s accuracy levels is more realistic when using either model on unseen Mazurkas.

5. DISCUSSION AND CONCLUSIONS

With five of Chopin's Mazurkas as use case, we have shown that local tempo for expressive music can be modeled using median aggregated IBIs, and tempo stability can be measured using the coefficient of variation (c_{var}) of local tempo values. Using these tools, we have found that the five Chopin Mazurkas exhibit high intra-track tempo variability, but low intra-segment variability, i.e., the local tempo is relatively stable and thus musically meaningful. This has allowed us to conduct a local tempo estimation experiment. As was to be expected, a standard beat-tracker like Böck and a tempo estimation CNN like DeepTemp—trained on Pop, EDM, and Ballroom music—perform relatively poorly for Mazurkas. Even when ignoring tempo octave errors, the results are by far inferior to those achieved by the same kind of CNN as DeepTemp, but trained on recordings from the target genre. It is reasonable to assume that training the Böck system on Mazurkas would also improve performance substantially—at the price of a strong genre bias. More interestingly, we have been able to confirm a relationship between estimation accuracy and tempo stability measured in c_{var}. Arguably, segments with a very high c_{var} may not have a meaningful local tempo and should therefore be excluded from local tempo evaluation. Another valuable insight results from the comparison of local accuracy results for DT−Maz-models trained on either the piece-wise M- or the performance-wise V-split. It

[10] To keep the evaluation concise, the reported local accuracy in all following accuracy figures use 4% tolerance.

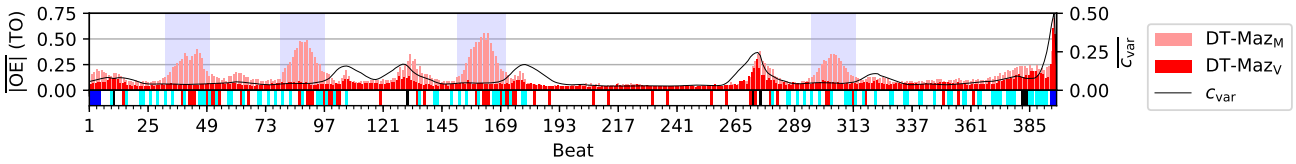


Figure 6: Averaged c_{var} and $|\,\text{OE}\,|$ for Op. 17, 4 around beats with classifications from [13]: non-event beats (black ■), boundary beats (blue ■), ornamented beats (red ■), and weak bass beats (cyan ■). High-error sections, unexplained by tempo instability, are highlighted in light-blue.

allows identification of piece-specific, musically difficult passages. When training and testing on the V-split, the network apparently has a chance to learn these piece-specific features not covered by data augmentation. One might also argue, DT−Maz$_V$ overfits to the pieces ("cover song effect" [25]). While usually seen as a negative effect, we exploit this to learn about our dataset by contrasting results with DT−Maz$_M$.

As with all deep learning systems, performance depends largely on the training data. For a production system, one is therefore well advised to use a larger and more diverse training set than we did in this case study.

6. FUTURE WORK

We consciously refrained from attempting to find ideal segment lengths and aggregation functions. We would therefore welcome studies on larger corpora of expressive music that search for optimal selection and aggregation functions as well as c_{var} ranges useful for meaningful evaluations.

Additional Material. Trained models are available at https://github.com/hendriks73/tempo-cnn

Acknowledgments. This work was supported by the German Research Foundation (DFG MU 2686/10-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS. The authors gratefully acknowledge the compute resources and support provided by the Erlangen Regional Computing Center (RRZE).

7. REFERENCES

[1] F. Gouyon, A. P. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1832–1844, 2006.

[2] B. H. Repp, "Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei"," *The Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2546–2568, 1992.

[3] ——, "On determining the basic tempo of an expressive music performance," *Psychology of Music*, vol. 22, no. 2, pp. 157–167, 1994.

[4] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *Journal of New Music Research*, vol. 30, pp. 39–58, 2001.

[5] P. Grosche and M. Müller, "Extracting predominant local pulse information from music recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1688–1701, 2011.

[6] E. Chew and C. Callender, "Conceptual and experiential representations of tempo: effects on expressive performance comparisons," in *Proceedings of the International Conference on Mathematics and Computation in Music (MCM)*. Springer, 2013, pp. 76–87.

[7] D. P. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[8] A. Gkiokas, V. Katsouros, G. Carayannis, and T. Stafylakis, "Music tempo estimation and beat tracking by applying source separation and metrical relations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 421–424.

[9] G. Peeters, "Time variable tempo detection and beat marking," in *Proceedings of the International Computer Music Conference (ICMC)*, Barcelona, Spain, 2005.

[10] ——, "Template-based estimation of time-varying tempo," *EURASIP Journal on Advances in Signal Processing*, 2007.

[11] J. L. Oliveira, F. Gouyon, L. G. Martins, and L. P. Reis, "IBT: A real-time tempo and beat tracking system," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 291–296.

[12] H. Schreiber and M. Müller, "A single-step approach to musical tempo estimation using a convolutional neural network," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 98–105.

[13] P. Grosche, M. Müller, and C. S. Sapp, "What makes beat tracking difficult? A case study on Chopin

Mazurkas," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 649–654.

[14] A. Lerch, C. Arthur, A. Pati, and S. Gururani, "Music performance analysis: A survey," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 33–43.

[15] K. Kosta, O. F. Bandtlow, and E. Chew, "Dynamics and relativity: practical implications of dynamic markings in the score," *Journal of New Music Research*, vol. 47, no. 5, pp. 438–461, 2018.

[16] J. Peperkamp, K. Hildebrandt, and C. C. S. Liem, "A formalization of relative local tempo variations in collections of performances," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 158–164.

[17] C. E. Cancino-Chacón, M. Grachten, W. Goebl, and G. Widmer, "Computational models of expressive music performance: A comprehensive and critical review," *Frontiers in Digital Humanities*, vol. 5, 2018.

[18] M. Fu, G. Xia, R. B. Dannenberg, and L. A. Wasserman, "A statistical view on the expressive timing of piano rolled chords," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, 2015, pp. 578–583.

[19] C. S. Sapp, "Hybrid numeric/rank similarity metrics," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Philadelphia, USA, 2008, pp. 501–506.

[20] F. Krebs, S. Böck, and G. Widmer, "Rhythmic pattern modeling for beat and downbeat tracking in musical audio," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Curitiba, Brazil, 2013, pp. 227–232.

[21] S. Böck, F. Krebs, and G. Widmer, "Joint beat and downbeat tracking with recurrent neural networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, New York City, USA, 2016, pp. 255–261.

[22] H. Schreiber and M. Müller, "Musical tempo and key estimation using convolutional neural networks with directional filters," in *Proceedings of the Sound and Music Computing Conference (SMC)*, Málaga, Spain, 2019, pp. 47–54.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference for Learning Representations (ICLR)*, San Diego, California, USA, 2015.

[24] H. Schreiber and M. Müller, "A post-processing procedure for improving music tempo estimates using supervised learning," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 235–242.

[25] H. Schreiber, C. Weiß, and M. Müller, "Local key estimation in classical music recordings: A cross-version study on Schubert's Winterreise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 501–505.