Tempo curves generation, estimation and analysis
L3 intership (CNAM / INRIA)
27/05/24 - 02/08/24

Sylvain Meunier (intern)
sylvain.meunier@ens-rennes.fr

Florent Jacquemard (supervisor)
florent.jacquemard@inria.fr

***Abstract*— Tempo estimation consists in detecting the speed at which a musician plays, or more broadly at which a piece of music is played. Such speed is usually expressed with respect to symbolic representation of the piece, in order to match the intuitive notion of a regular *pulse*. We present here some results regarding the generation and analysis of local tempo curves from musical performances involving, first, methods that need to be given some symbolic information, and then methods that don't. More precisely, we focus here on tempo estimation for a given performance recorded as a MIDI file, on both a local and global level, and with or without prior knowledge of a reference *sheet music* (score). In order to do so, we introduce mathematical formalisms based on underlying notions in the literature.**

***Index terms*—Music Information Retrieval, tempo estimation, quantization, musical formalism, musical data representation**

I. Introduction

The Music Information Retrieval (MIR) community focuses on three representations of musical information, presented here from lowest to highest level of formatting. The first one is raw audio, either recorded or generated, encoded using WAV or MP3 formats. The computation is based on a physical understanding of signals, using audio frames and spectrum, and represents the most common and accessible kind of data. The second is a more musically-informed format, representing notes with both pitch (i.e., the note that the listener hears) and duration in Real Time Unit (RTU) (sec.), encoded within a MIDI file. The last way to encode musical information is a MusicXML file, mainly used for display and analysis purposes. The latter relies on a symbolic and abstract notation for time, that only describes the length of events in relation to a specific Musical Time Unit (MTU), called a beat, and indicates as well the pitch and articulation of those events. These symbolic indications are then to be interpretated by a performer. An expressive musical performance is not only about theoretical compliance with rhythmic musical theory (a task that computers excel at), but rather, and actually mostly, about sprinkling micro-errors (refered to here as or shifts, or (micro) timings).

Moreover, to actually play a sheet music, one needs a given tempo, usually indicated as an amount of beat per minute (BPM). Therefore, the notion of tempo allows to translate MTU symbolic notation into RTU events. We will present a formal definition for both tempo and performance in Section II.A.

However, tempo itself is insufficient to translate a music score into musical performance, i.e., a sequence of real time events. Indeed, S. D. Peter et al. [1] present four parameters, among which tempo and articulation appear the most salient as opposed to velocity and timing. Even though the MIR community studies the four parameters, the hierarchy exposed by [1] embodies quite well their relative priority within literature.

Tempo and associated works actually hold a prominent place in literature. Tempo inference was first computed based on probabilistic models [2]–[4], and physical or neurological models [5], [6] as methods for real time (musical) score synchronization with a performance ; and later the community tried neural network models [7] and hybrids approaches [8].

A very useful preprocessing task for tempo inference and further analysis, such as [9]–[11], is note-alignement, that is a matching between each note of a MIDI performance and those indicated by a given score. Two main methods are to be found in literature : a dynamic programming algorithm, equivalent to finding a shortest path, that can work on raw audio [12] and a Hidden Markov Model that needs more formatted data, such as MIDI files [13]. As most of the previous examples, we shall focus here on mathematically or musically explainable methods.

We shall present below our following contributions :

- Formal definition of tempo, based on [2], [9] and [11] (II.A) ; and some immediate consequences (II.B)
- Revision of [5] and [6] for score based tempo inference (II.C)
- Original techniques of tempo inference, without score, based on [14] (III.A), and [15], with related new theoretical results (III.B, III.C and Appendix C)
- Method for data augmentation, and related results, based on [16] and [17] (IV.A)

This document, along with some algorithm implementations and detailled results can be found on the dedicated github repository [18]. Most proofs are to be found in Appendices.

II. Score-based approaches

A. Preliminary works

Definition Let $u = (u_n)_{n\in\mathbb{N}}$ be a sequence, we introduce the notation : (u_n) for u, where n is a dummy variable, and its introduction $n \in \mathbb{N}$ is implicit.

Since we chose to focus on MIDI files, we will represent a (monophonic) performance as a strictly increasing sequence of timepoints, or events, $(t_n) \in \mathbb{R}^{\mathbb{N}}$, each element of whose indicates the onset of a corresponding performance event. Such a definition is very close to an actual MIDI representation.
For practical considerations, we will stack together all events whose distance in time is smaller than $\varepsilon = 20$ ms. This order of magnitude represents the limits of human ability to tell two rhythmic events apart [4], and is widely used within the field [8]–[11], [14]–[17]. Likewise, a music score will be represented as a strictly increasing sequence of symbolic events $(b_n) \in \mathbb{R}^{\mathbb{N}}$. An extension of this formalism for polyphonic pieces is discussed in Appendix A.
Please note that, in both definitions, the terms of the sequence do not indicate the nature of the corresponding event (chord, single note, rest...). Moreover, in terms of time units, (t_n) is expressed in RTU, whereas (b_n) is expressed in MTU.
With these definitions, let us formally define tempo :

Definition II.1 $T \in (\mathbb{R}_+^*)^{\mathbb{R}}$ is said to be a formal tempo (curve) with respect to (t_n) and (b_n) when,
for all $n \in \mathbb{N}$, $\int_{t_0}^{t_n} T(t)\,\mathrm{d}t = b_n - b_0$

Proposition II.2 Let $T \in (\mathbb{R}_+^*)^{\mathbb{R}}$.
T is a formal tempo with respect to (t_n) and (b_n) iff
$\forall n \in \mathbb{N}, \int_{t_n}^{t_{n+1}} T(t)\,\mathrm{d}t = b_{n+1} - b_n$

Since tempo is only observable between two events *a priori*, we introduce a definition for a canonical tempo T^*, also called immediate tempo.

Definition II.3 Given (t_n) and (b_n), respectively a performance and a score, the canonical tempo is defined as a stepwise constant function $T^* \in (\mathbb{R}_+^*)^{\mathbb{R}}$ such that :
$\forall x \in \mathbb{R}^+, \forall n \in \mathbb{N}, x \in [t_n, t_{n+1}[\Rightarrow T^*(x) = \frac{b_{n+1}-b_n}{t_{n+1}-t_n}$

The reader can verify that this function is a formal tempo as defined in Definition II.1. From now on, we will assume by convention that $t_0 = 0$ RTU et $b_0 = 0$ MTU.

Even though there is a general consensus in the field as for the interest and informal definition of tempo, several formal definitions coexist within literature : [2], [9] and [11] choose definitions very similar to T^*, approximated at the scale of a measure or a section for instance, whereas [3] and [8] use $\frac{1}{T^*}$.

When a performance perfectly fits theoretical expectations, T^* has the advantage to coincide with the tempo indicated on a traditional sheet music (and therefore on a corresponding MusicXML) when expressed in BPM, hence allowing for a simpler and more direct interpretation of results.

B. Computation of the canonical tempo

There exists a few datasets containing note-alignment matching between both music score and corresponding audio, more or less anotated with various labels [9]–[11], [16], [17]. For this study, we chose to rely on the (n)-ASAP dataset[1] that presents a vast amount of piano performances on MIDI format, with over 1000 different pieces of classical music, all note-aligned with their corresponding score. From there, we can easily visualize our definition of canonical tempo.

Figure 1 presents the results for a specific piece of the (n)-ASAP dataset with a logarithmic y-scale, that shows two abrupt tempo changes, whilst maintaining a rather stable tempo value in-between.

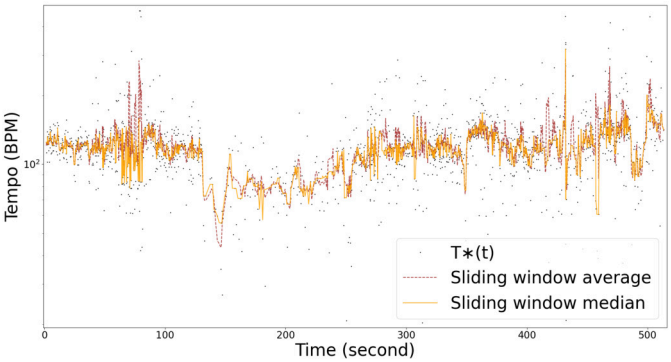

Figure 1: Graph of T^* for a performance of Islamey, Op.18, M. Balakirev, extracted from the (n)-ASAP dataset.

In this graph, one can notice how T^* (plotted as little dots) appears to be noisy over time; even though allowing to distinguish a tempo change at $t_1 = 130$ s and $t_2 = 270$ s. Both the sliding window average (dotted line) and median (full line) of T^* seem unstable, presenting undesirable peaks, whereas the perceived tempo is quite constant for the listener. The median curve is a bit more stable than the average curve, as expected. There are two explanation for those results. First, fast events are harder to play exactly on time, and the very definition being a ratio with a small theoretical value at the denominator explains the deviation and absurd immediate tempo plotted. In fact, we can read that about 10 points are plotted over 400 BPM (keep in mind that usual tempo are in the range 40 - 250 BPM). Second, the notions of timing and tempo are merged in this computation, hence giving results that do not match the listener feeling of a stable tempo. Actually, timing can be seen as expressive modifications to the "official" score. Using the score containing said modifications would allow for curves that fit better the listener feeling, though needing an actual transcription of the performance first. For instance, if the performer plays "off-the-beat", with added syncopations, which is notably common in swing interpretations, one would need the actual transcription in order to define a meaningful tempo.

[1] https://github.com/CPJKU/asap-dataset

C. Two physical models for tempo generation

Among the tasks requiring tempo generation, score following, that is the real time inference of tempo to allow a dedicated machine to play an accompagnement by following at least one actual musician, has been tackled by various approaches in literature. C. Raphael [2] started with a probabilistic model, but those methods have found themselves replaced by a more physical understanding of tempo *via* the notion of internal pulse, as explained by E. W. Large and M. R. Jones [5]. In fact, a combination of these methods has recently been developped as a commercial product[2], based on an a previous work by A. Cont [19].

The approach [5] considers a simplified neurological model, where listening is a fundamentally active process, implying a synchronization between *observations*, i.e., external events (those of the performance) and *expectations*, here being an internal oscillator whose complexity depends of hypotheses on the shape of *observations*. The model consists of two equations for the internal parameters presented hereafter for all $n \in \mathbb{N}$:

$$\Phi_{n+1} = \left[\Phi_n + \frac{t_{n+1} - t_n}{p_n} - \eta_\Phi F(\Phi_n, \kappa)\right] \underset{[-0.5,\ 0.5[}{\mathrm{mod}}\ 1 \quad (1)$$

$$p_{n+1} = p_n\big(1 + \eta_p F(\Phi_n, \kappa)\big) \quad (2)$$

where Φ_n corresponds to the phase, or rather the phase shift at each event t_n between the oscillator and the external events, p_n embodies its period, $\eta_p \in \mathbb{R}^+$ and $\eta_\Phi \in \mathbb{R}^+$ are both constant damping parameters, and F is the correction at t_n to match *expectations* and *observations*.

This initial model is then modified to consider a notion of attending *via* the κ parameter, whose value changes over time according to other equations not showed here.
We finally have : $F : \Phi, \kappa \mapsto \frac{\exp(\kappa\cos(2\pi\Phi))}{\exp(\kappa)} \frac{\sin(2\pi\Phi)}{2\pi}$

This model being fit for beat tracking, we modified it to consider score information in order to generate a more stable and precise value of tempo than the naive approach previously presented. The modifications presented hereafter were made in order to keep consistency with respect to the original model theoretical framework of validity :

$$\Phi_{n+1} = \Phi_n + \frac{t_{n+1} - t_n}{p_n} - \eta_\Phi F(\Psi_n, \kappa) \quad (3)$$

$$p_{n+1} = p_n\big(1 + \eta_p F(\Psi_n, \kappa)\big) \quad (4)$$

where $\mathrm{amin} : a, b \mapsto \begin{cases} a \text{ if } |a| < |b| \\ b \text{ otherwise} \end{cases}$

$$\Psi_n = -\,\mathrm{amin}(k + b_n - \Phi_n, k + 1 + b_n - \Phi_n)$$
$$k = \lfloor \Phi_n - b_n \rfloor$$

Here, the amin function is used in order to represent a choice between two corrections. The first argument can be interpreted as a correction with respect to the most recent passed beat time occuring exactly on a actual beat. Said beat is formally defined as $a_1 = \max_{n \in \mathbb{N}\ :\ b_n \leq \Phi_i} \lfloor b_n \rfloor$ where Φ_i is the internal value at time i acting as a beat unit. The second argument embodies the correction according to $a_2 = a_1 + 1$, the following beat. One can notice that the phase is actually always considered modulo 1 in [5], since it appears only multiplied by 2π in either cos or sin functions. Using this remark, one can verify that, in the initial presentation of the model with a metronome, i.e., $\forall n \in \mathbb{N}, b_n = 0 \bmod 1$, the extension proposed in $(3, 4)$ is equivalent to the original approach $(1, 2)$, hence justifying the designation "extension". We will from now on refer to this model as *Large et al.* since the modifications presented here were inspired by various works in literature, including [19]. The original model will not be considered in this report.

Even though the latter has been validated experimentally in [5], and is still used in the presented version [20], a theoretical study of the system behavior remains quite complex, even in simplified theoretical cases, notably because of the function F expression. [6] thus presents the *TimeKeeper* model, that can be seen as a linearization of the previous approach, valid in the theoretical framework of a metronome presenting small tempo variations. In fact, there is a strong analogy between the two models, that are almost equivalent under specific circumstances [21]. We used the derandomised version considered by J. D. Loher et al. [21]. Using their analogy, we then obtain the following equations for *TimeKeeper* :

$$A_{i+1} = K_i(1 - \alpha) + \tau_i - (t_{i+1} - t_i) \quad (5)$$

$$\tau_{i+1} = \tau_i - \beta \times \left[K_i \underset{[-0.5,\ 0.5[}{\mathrm{mod}}\ 1\right] \quad (6)$$

where $K_i = -\,\mathrm{amin}(k\tau + b_i - A_i, (k+1)\tau + b_i - A_i)$

$$k = \left\lfloor \tfrac{A_i - b_i}{\tau_i} \right\rfloor$$

.

Here, A_i is the absolute asynchrony at time t_i, with a similar role than the phase shift in (1), α and β are both constant damping parameters, and τ_i is the time value that represents the current tempo, similarly to the period in (2).
Figure 2 displays the results of those two models, compared to the canonical tempo. One can notice that the *Large et al.* model is less stable than *TimeKeeper*, although faster to converge.

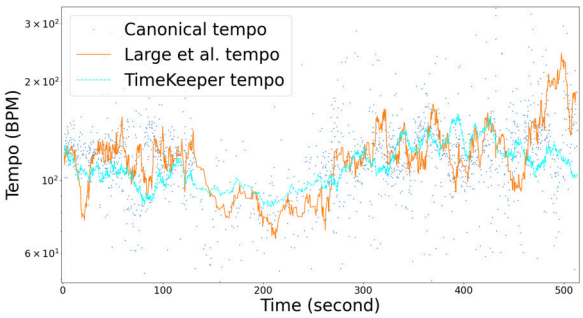

Figure 2: Tempo curve for the same performance of Islamey, Op.18, M. Balakirev, according to the models presented here

[2] https://metronautapp.com/

Figure 3 illustrates the differences in managing an irrelevant tempo initialization value of the two models, starting here both with the initial tempo value of 70 BPM (♩ = 70, i.e., the *beat* unit here is a quarter note). As expected, *TimeKeeper* does not manage to converge to any significant tempo : its theoretical framework supposes small tempo variations (and preferably relevant initialization). However, *Large et al.* model manages to converge to a meaningful result. In fact, in the range 9 to 70 seconds, its estimated tempo is exactly half of the actual tempo hinted by the blue dots (canonical tempo). This is an example of a *tempo octave*, defined and discussed in Appendix A.

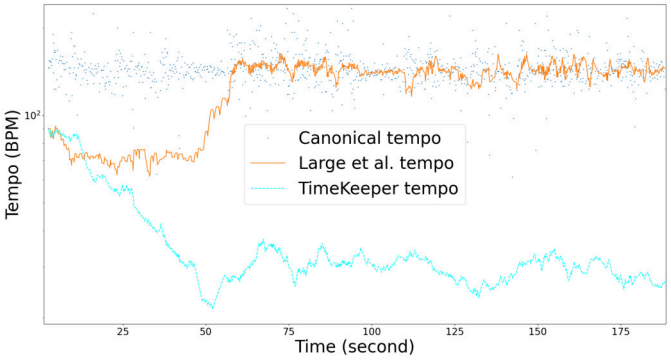

Figure 3: Tempo curve for a performance of Piano Sonata No. 11 in A Major (K. 331: III), W.A Mozart, according to the two previously modified models, with irrelevant initialization

A solution to adress this latter problem could be to begin the computation from the end of the piece, and the going backwards to the start, hence hopefully obtaining a meaningful initialization value. Such a method is theoretically valid as explained in Appendix A.

III. Scoreless approaches

The first issue with the two previous approaches is the requirement of both a reference score and a note-alignment between the given performance and the latter, which is something the field lacks at large scale. Therefore, we will now focus on methods for tempo estimation that **do not** require the prior knowledge of a reference score. However, we will suppose such a score actually exists, and use the notation (b_n) to designate it for formal proofs. One may notice that in this framework, estimating T^* is equivalent to transcribing the actual performance, which we can consider to be the most exact tempo curve one can compute. Since such a tempo cannot be uniquely determined (see Appendix A for details on the *tempo octave* problem), we will here try to relax the problem by finding a "flattened" tempo curve that intuitively gives the general tempo hinted by T^*. To a lesser extent, we will try to find methods that do not present salient sensitivity to tempo initialization, unstability nor require to accurately estimate relevant values of some constant internal parameters.y According to our implementation, *Large et al.* model is a particularly chaotic model regarding the latter.

This section presents two models, respectively based on [14] and [15] that rely on the notion of quantization, i.e., the process of converting real values into simple enough rational numbers, according to restrictions.

A. Introduction of an estimator based approach

Definition Given a sequence (u_n), let from now on (Δu_n) be $(u_{n+1} - u_n)_{n\in\mathbb{N}}$

(Δu_n) embodies the durations of the different events of (u_n). The previous expression is actually valid in monophonic pieces, under the hypothesis that the end of each note is exactly the begining of the next one. Since our framework is firstly monophonic, we will consider the given expression for durations, but this sequence could be defined otherwise in order to consider polyphonic pieces. Most of the following results do not depend on the actual expression of durations. See Appendix A and Appendix B for polyphonic considerations.

This first method aims at tracking tempo variation rather than actual values. Hence, we suppress the need for a convergence time. In fact, we search to estimate αT^*, where $\alpha \in \mathbb{R}_+^*$ is an unknown multiplicative factor that we try to make constant over time. Using the formalism presented in III, we first present the following result since $T_n^* > 0$:

$$T_{n+1}^* = T_n^* \frac{T_{n+1}^*}{T_n^*} = T_n^* \frac{\Delta t_n}{\Delta t_{n+1}} \frac{\Delta b_{n+1}}{\Delta b_n} \tag{7}$$

Let T_n be an estimation of T_n^* by a given model at a given time t_n and $\alpha_n = \frac{T_n}{T_n^*}$. We obtain $\alpha_n T_{n+1}^* = \underbrace{\alpha_n T_n^*}_{T_n} \frac{\Delta t_n}{\Delta t_{n+1}} \times \frac{\Delta b_{n+1}}{\Delta b_n}$

In the above formula, the only value to actually estimate is therefore $\frac{b_{n+2}-b_{n+1}}{b_{n+1}-b_n}$, which allows for a locally constant shift in both our estimations of the numerator and denominator. Hence the resulting value is invariant by translation, or constant multiplication of our estimation of (b_n). Furthermore, this value only deals with symbolic units, meaning that we can apply musical properties to find a consistent result.

The point of this approach is to keep a constant factor between (T_n) and (T_n^*). We thus define $T_{n+1} = \alpha_n T_{n+1}^*$, to find :

$\frac{\Delta b_{n+1}}{\Delta b_n} = \frac{T_{n+1}^* \Delta t_{n+1}}{T_n^* \Delta t_n} = \frac{T_{n+1}/\alpha_n}{T_n/\alpha_n} \times \frac{\Delta t_{n+1}}{\Delta t_n}$,

Hence $\frac{\Delta b_{n+1}}{\Delta b_n} = \frac{T_{n+1}}{T_n} \times \frac{\Delta t_{n+1}}{\Delta t_n}$.

If we manage to correctly estimate T_{n+1}, we can obtain a tempo estimation with the same multiplicative shift as the previous estimation T_n, thus by using the formula recursively, we obtain a model that can track tempo variations over time without any need for convergence, hence being robust to irrelevant tempo initialization, while using only local methods (i.e., the resulting model is online). We then obtain :

$$\frac{T_{n+1}}{T_n} = \frac{T_{n+1}^*}{T_n^*} = \frac{\Delta t_n}{\Delta t_{n+1}} \underbrace{E\left(\frac{T_{n+1}}{T_n} \times \frac{\Delta t_{n+1}}{\Delta t_n}\right)}_{(\Delta b_{n+1})/(\Delta b_n)} \tag{8}$$

where E, designated by *estimator*, is supposed to act on a theoretical ground as an oracle that returns the correct value of the symbolic $\frac{\Delta b_{n+1}}{\Delta b_n}$ from the given real values indicated in (8). In practice, E is actually a rhythmic quantizer.

Given an estimator E, the tempo value defined as T_{n+1}, computed from both T_n and local data, is obtained *via* the following equation, where x embodies $\frac{T_{n+1}}{T_n}$ in (8) :

$$T_{n+1} = T_n \underset{x\in\left[\frac{\sqrt{2}}{2}T_n,\sqrt{2}T_n\right]}{\operatorname{argmin}} d\left(x, \frac{\Delta t_n}{\Delta t_{n+1}} E\left(x\frac{\Delta t_{n+1}}{\Delta t_n}\right)\right) \quad (9)$$

with $d : a, b \mapsto k_*|\log(\frac{a}{b})|, k_* \in \mathbb{R}_+^*$ a logarithmic distance, choosen since an absolute distance would have favored small values by triangle inequality in the following process.
Further explanations about (9) can be found in Appendix B.

In the implementation presented here, the estimator role is to output a musically relevant value, given that the real durations contain (micro)-timing. In our tests, we limited these outputs to be either regular divisions (i.e., powers of 2) or triplets. Furthermore, the numerical resolution for the previous equation was done by a logarithmically evenly spaced search and favored x values closer to 1 (i.e., T_{n+1} closer to T_n) in case of distance equality.

Such a research allows for a musically explainable result : the current estimation is the nearest most probable tempo, and both halving and doubling the previous tempo is considered as improbable, and as further away from the initial tempo.

Figure 4 compares the canonical tempo and the tempo curve obtained by our naive estimator (here almost correctly initialized to simplify the interpretation of the result). One can notice that the constant distance between the two curves indicates in our logarithmic scale a constant multiplicative factor, except for the dotted line. There, the actual ratio of $\frac{T^*_{n+1}}{T^*_n}$ exceeds our bound of $[\frac{1}{\sqrt{2}}, \sqrt{2}[$. Hence, the computation finds the representative within the range, here being $\frac{1}{2}\frac{T^*_{n+1}}{T^*_n} > \frac{\sqrt{2}}{2}$. This example illustrates that the values obtained through this method may be off by a power of 2. Appendix B presents a formal study of this model adressing this problem.

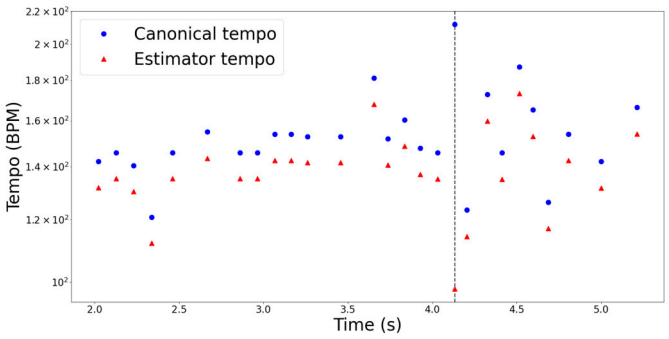

Figure 4: Normalized canonical tempo and estimation according to the model presented here for the first measures of Piano Sonata No. 11 in A Major (K. 331: III), W.A Mozart.

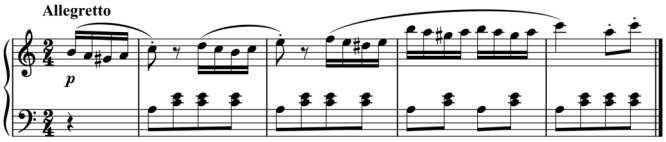

Figure 5: The first measures of Piano Sonata No. 11 in A Major (K. 331: III), W.A Mozart.

B. *Towards a quantized approach*

Definition Let $f \in \mathbb{R}^{\mathbb{R}}$ be a continuous function, $a \in \mathrm{dom}(f)$ is said to be a semi-strict local minimum (resp. maximum) of f when a is a local minimum (resp. maximum) of f, and a is not a local maximum (resp. minimum) of f, or in other words, f is not constant on a neighbourhood of a.

In this section, we extend the previous approach by considering the estimator as our central model and only then extracting tempo values rather than the opposite. We based our work on G. Romero-García et al. [15] with the previous formalism.

Let $n \in \mathbb{N}^*$ and $D \subset (\mathbb{R}^+)^n$ be a set of some durations of real time events. The function ε_D is defined by [15] as :

$$\varepsilon_D : a \mapsto \max_{d\in D} \min_{m\in\mathbb{Z}} |d - ma| \quad (10)$$

This continuous function is called the *transcription error*, and can be interpretated as maximum error (in RTU) between all real events $d \in D$ and theoretical real duration ma, where m is a symbolic notation expressed in arbitrary symbolic unit, and a a real time value corresponding to a tatum at a given tempo. We proove in Appendix C that the set of all semi-strict local maxima of ε_D is :

$$\begin{aligned} M_D &= \left\{\frac{d}{k+\frac{1}{2}}, d \in D, k \in \mathbb{N}\right\} \\ &= \bigcup_{d\in D}\left\{\frac{d}{k+\frac{1}{2}}, k \in \mathbb{N}\right\} \end{aligned} \quad (11)$$

In fact, each of these local maxima corresponds to a change of the m giving the minimum in the expression of ε_D, hence the following result : in-between two such successive local maxima, the quantization remains the same, i.e. Proposition III.1 :

Proposition III.1 Let m_1, m_2 be two successive local maxima of ε_D, $a_1 \in]m_1, m_2[, a_2 \in [m_1, m_2], d \in D$ Then,
$\forall m \in \mathbb{Z}, m \in \underset{k\in\mathbb{Z}}{\operatorname{argmin}}|d - ka_1| \Rightarrow m \in \underset{k\in\mathbb{Z}}{\operatorname{argmin}}|d - ka_2|$.

Corollary III.2 Let $d \in D, a \in \mathbb{R}_+^*, A = \underset{k\in\mathbb{Z}}{\operatorname{argmin}}|d - ka|$.
Then : $0 < |A| \leq 2$ and $|A| = 2 \Leftrightarrow a \in M_D$.

Proof $A \subset \left\{\left\lfloor\frac{d}{a}\right\rfloor, \left\lfloor\frac{d}{a}\right\rfloor + 1\right\}$, $\lim\limits_{|k|\to+\infty} |d - ka| = +\infty$, hence $A \neq \emptyset$. Finally, let $k = \left\lfloor\frac{d}{a}\right\rfloor$,

$$\begin{aligned} |A| = 2 &\Leftrightarrow A = \left\{\left\lfloor\tfrac{d}{a}\right\rfloor, \left\lfloor\tfrac{d}{a}\right\rfloor + 1\right\} \\ &\Leftrightarrow |d - ka| = |d - (k+1)a| = \tfrac{a}{2} \\ &\Leftrightarrow a = \tfrac{d}{k+\frac{1}{2}} \Leftrightarrow a \in M_D \end{aligned}$$

□

With this property, we can then choose to consider only semi-strict local minima of ε_D as in [15], since there is exactly one semi-strict local minimum in-between two semi-strict local maxima, and choosing any other value in this range would result in the exact same transcription, with a higher error by definition of a local minimum (that is global on the considered interval). The correctness of the following algorithm to find all local minima within a given interval is proven in Appendix C.

FindLocalMinima($D \neq \emptyset$, start, end) :

$M \leftarrow \left\{ \frac{d}{k+\frac{1}{2}}, d \in D, k \in \left[\!\left[\frac{d}{\text{end}}, \frac{d}{\text{start}} \right]\!\right] \right\}$
$P \leftarrow \left\{ \frac{d_1+d_2}{k}, (d_1, d_2) \in D^2, k \in \left[\!\left[\frac{d_1+d_2}{\text{end}}, \frac{d_1+d_2}{\text{start}} \right]\!\right] \cap \mathbb{N}^* \right\}$
localMinima $\leftarrow \emptyset$
for $(m_1, m_2) \in M^2$ two successive local maxima :
    in_range $\leftarrow \{p \in P : m_1 \leq p < m_2\}$
    localMinima $\leftarrow$ localMinima $\cup \{\min(\text{in\_range})\}$
return localMinima

Algorithm 1: Finds all semi-strict local minima of [start, end]

[15] then defined $G = (V, E)$ a graph whose vertices are the semi-strict local minima of ε_D, where D is a sliding window, or *frame*, on a given performance, and whose edges are such that they can guarantee a *consistency property*, explained hereafter.

The *consistency property* for two tatums a_1, a_2 specifies that, for all $d \in F_\cap$ where $F_\cap$ is the set of all values in common between two successive frames, d is equally quantized according to a_1 and a_2, i.e., the symbolic value of d is the same when considering either a_1 of a_2 as the duration of a given tatum at some tempo (respectively $\frac{1}{a_1}$ and $\frac{1}{a_2}$, as shown in Section III.C). From these definitions, we can now define a *tempo curve* as a path in G. In fact, [15] call such a path a *transcription* rather than a tempo curve, yet, since an exact tempo curve would be (T_n^*), these two problems are equivalent.

Actually, the consistency property is not that restrictive when considering tempo curves. Let F_1, F_2 be two successive frames, $F_\cap = F_1 \cap F_2$, $d \in F_\cap$, and p a path in G containing a local minimum a_1 of ε_{F_1}. According to (11), we can divide the set of all semi-strict local maxima in two, with those "caused" by $F_\cap$, $M_{F_\cap}$, and the others. Let then $m_1, m_2 \in M_{F_\cap}$. These are local maxima for both ε_{F_1} and ε_{F_2} by (11) since $F_\cap \subset F_2$, and therefore there is at least one local minimum within the range $]m_1, m_2[$ for both of these functions. However, thanks to Proposition III.1, we know that both these local minima will equally quantize the elements of D. Hence, by defining :

- $m_1 = \max\left\{m \in M_{F_\cap} : m < a_1\right\}$
- $m_2 = \min\left\{m \in M_{F_\cap} : m > a_1\right\}$
- a_2 a local minimum of ε_{F_2} in the range $]m_1, m_2[$, which exists since m_1 and m_2 are semi-strict local maxima,

(a_1, a_2) is *consistent* according to the consistency property.

Corollary The consistency property only implies restrictions relative to the interval of research. In other words, any given strictly partial path p in G can be extended to a *consistent* path, even if it means considering a larger interval, for any given performance, and any given frame length used to define G.

Remark III.3 This approach supposes that we can define a tatum within all frames i.e., that there is a real value of the same tatum that can maintain a single value within the frame. In other words, let a be the symbolic value of a given tatum, $n \in \mathbb{N}$, and $F = \{f_1, ..., f_n\}$ a given frame. If we suppose all the events of F to embody a musical meaning, we can define an immediate tempo for each of them, and therefore we can express a_i the value of a in RTU at the tempo t_i corresponding to f_i, $i \in [\![1, n]\!]$. In order for this approach to be meaningful, we then need the existence of $\hat{a} \in \mathbb{R}_+^*$ such that each f_i is equally quantized according to a_i and $\hat{a}$. Since, by definition, f_i is quantized by $f_i \times t_i$, we obtain the following definition.

Definition III.4 With the notations introduced in Remark III.3, a tatum $a \in \mathbb{R}_+^*$ is said to have an actual meaning with respect to a frame $F = \left\{f_1, ..., f_{|F|}\right\}$ when, for all $i \in [\![1, |F|]\!]$, $f_i \times t_i \in \underset{k \in \mathbb{Z}}{\operatorname{argmin}}(f_i - k\hat{a})$.

Proposition III.5 When $t_i = t$ for all $i \in [\![1, |F|]\!]$, $\frac{1}{t}$ has an actual meaning with respect to F.

C. *Quantization revised*

We choose to define our tatum ε as $\frac{1}{60}$♩, which corresponds to a sixteenth note (i.e., 𝅘𝅥𝅯) wrapped within a triplet within a quintuplet, and has the property that $1\ \varepsilon/\text{sec} = 1$ ♩/min. This definition implies that we restrain to symbolic durations that are integer multiples of ε in our framework. Hence, we can take ε as our MTU. We then have $T := \frac{\Delta b}{\Delta t} = \frac{1\varepsilon}{a}$, where a is the theoretical duration of ε at tempo T. From there, we can define $\sigma_D : a \mapsto \frac{1}{a}\varepsilon_D(a)$ the *normalized error*, or *symbolic error*, which embodies the error between a transcription m of $d \in D$ expressed in tatum (thus a quantized and valid transcription), and $d \times \frac{1}{a} = d \times T$, which is the expression of the symbolic duration of d at tempo T according to Definition II.3.

Definition III.6 Let $A \subset \mathbb{R}$ a countable set and $f : A \to \mathbb{R}$. $a \in A$ is said to be a local minimum of f when :
$\exists \eta \in \mathbb{R}_+^* : \forall a' \in \mathrm{H} := \{a' \in A : |a - a'| \leq \eta\}$,
$f(a') \geq f(a) \wedge \sup(\mathrm{H}) > a > \inf(\mathrm{H})$.

In order to reduce the amount of local minima considered during computation, we propose the following conditions on a local minimum a of ε_D, where $m = \{a' \in m_D : a' \geq a\}$ with m_D the set of all semi-strict local minima of ε_D :

1. $\forall a' \in m, \varepsilon_D(a) \leq \varepsilon_D(a')$
2. $\forall a' \in m, \sigma_D(a) \leq \sigma_D(a')$
3. a is a local minimum of $\lambda_1 : x \in m_D \mapsto \varepsilon_D(x)$
4. a is a local minimum of $\lambda_2 : x \in m_D \mapsto \sigma_D(x)$

Since $2 \Rightarrow 1$ and $4 \Rightarrow 3$, one can only consider conditions 2, 4.

We present here a modified version of [15] introduced in the previous section. We similarly define a graph $G = (V, E)$, with the relaxation that the edges no longer guarantee the consistency property. In order to define our tempo curves, we then sequentially define paths in G according to the following process, where d is a logarithmic distance, P is the set of all possible paths, and A_n the set of all local minima of ε_{F_n} ; F_n being the n-th frame.

Definition III.7 A path $p = (p_1, ..., p_f)$ in G is said to be a fixpoint when there exists a path $p' = (p'_1, ..., p_f)$ in G such that $\forall i \in [\![1, |p'| - 1]\!]$, $p'_{i+1} \in \underset{s \in A_{i+1}}{\operatorname{argmin}}\, d(s, p'_i)$, and $p'_1 \in \underset{s \in A_1}{\operatorname{argmin}}\, d(s, p_f)$.

FindPotentialPaths(performance, frame_length, start, end) :

$F_1 \leftarrow \{p_i \in \text{performance}, i \in [\![1, \text{frame\_length}]\!]\}$
$P \leftarrow \text{FindLocalMinima}(F_1, \text{start}, \text{end})$
for n corresponding to a frame :
    $F_n \leftarrow \{p_{n+i} \in \text{performance}, i \in [\![1, \text{frame\_length}]\!]\}$
    $A_n \leftarrow \text{FindLocalMinima}(F_n, \text{start}, \text{end})$
    for all path $p = (p_1, ..., p_n)$ in P :
        $p_{n+1} \leftarrow \underset{a \in A_n}{\operatorname{argmin}}\, d(p_n, a)$
        $p \leftarrow (p_1, ..., p_n, p_{n+1})$
return the fixpoints of P

Algorithm 2: Returns potential tempo curves of a performance

We prove in Appendix C that this algorithm is correct under some hypotheses, notably that the final tempo of the section is the same as the one at the begining, which is true in particular if the tempo is quasi-constant during the considered section.

D. *Raw results and comparison with Section II*

All the tempo curves presented in this section have been obtained on performances without abrupt tempo changes. In order to output a single curve, we choose here to select the path whose first tatum value corresponds to the nearest tempo to a given initialization value, according to d.

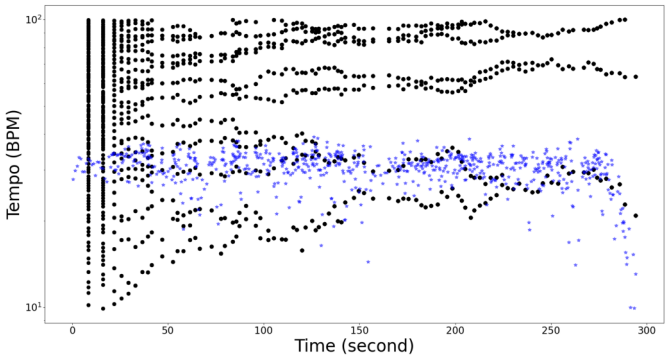

Figure 6: Potential tempo curves and canonical tempo for a performance of J-S. Bach, Italian Concerto, BWV 971

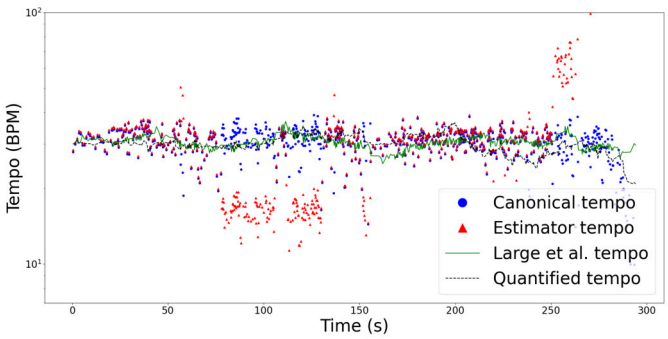

Figure 7: Comparison of different models for the previous performance of J-S. Bach, Italian Concerto, BWV 971

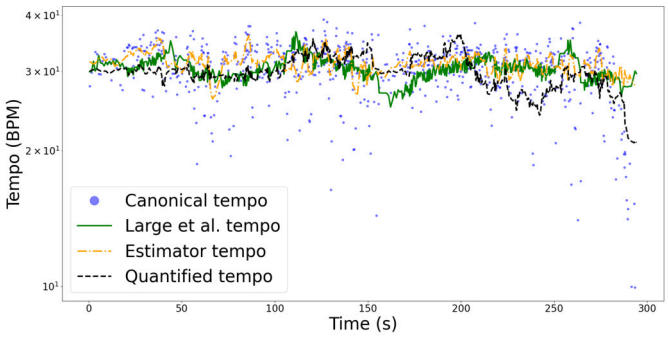

Figure 8: Comparison of different models for the same performance of J-S. Bach, Italian Concerto, BWV 971

Figure 6 presents the potential tempo curves (black dots) and actual canonical tempo (blue stars). Figure 7 illustrates that the estimator approach is sensitive to tempo octaves, hence we only plotted values within the range [10, 100]. Even though, the associated transcription appears satisfying, there would need for an actual formatting of said transcription in order to obtain satisfying tempo values. The latter is done on Figure 8, where one can see the tempo curve hinted by the estimator approach is actually really similar to the mean of the canonical tempo notably presented in Figure 1, though less unstable. However, such an approach needs for post-processing, and a way to determine the constant factor (obtained *via Large et al.* model here). The quantized approach - which do not require such heavy post-processing - appears more stable than the methods presented in Section II, as shown by Figure 9.

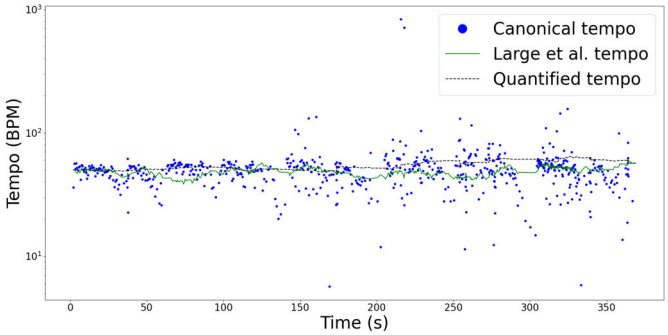

Figure 9: Comparison of different models for a performance of M. Ravel, *Pavane pour une infante défunte*

IV. Applications

A. A method for (monophonic) data augmentation

Let $(\Delta t_n), (T_n^*), (T_n)$ be respectively a performance, the canonical tempo for this performance, an estimated tempo curve (supposed to be flattened with respect to the canonical tempo) and $T_c \in \mathbb{R}_+^*$ a given tempo value.

Definition IV.1 We define the symbolic shift of the performance according to (T_n^*) and (T_n) as $(s_n := \Delta t_n(T_n - T_n^*))$.

Definition IV.2 The *normalized* performance at tempo T_c is defined as : $\hat{t}_0 = t_0, \hat{t}_{n+1} = \hat{t}_n + \underbrace{\Delta t_n \frac{T_n^*}{T_c}}_{\alpha_n} + \underbrace{\frac{s_n}{T_c}}_{\beta_n} \forall n \in \mathbb{N}$.

Here a_n represents the new duration of Δt_n at tempo T_c, since $\alpha_n = \frac{\Delta b_n}{T_c}$, and β_n embodies the actual time shift at tempo T_c.

Proposition IV.3 $\left(\hat{t}_n\right)$ is indeed a performance, as defined in Section II.A. Furthermore, $\left(\hat{T}_n^*\right) = \left(\frac{\Delta b_n}{\Delta \hat{t}_n}\right) = \left(\frac{1}{1+\frac{s_n}{\Delta b_n}} T_c\right)$.

Proof For all $n \in \mathbb{N}^*$:
$\Delta \hat{t}_n = \alpha_n + \beta_n = \Delta \frac{t_n}{T_c}(T_n^* + T_n - T_n^*) = \Delta t_n \frac{T_n}{T_c} > 0$
Furthermore, $\hat{T}_n^* = \frac{\Delta b_n}{\Delta \hat{t}_n} = \frac{\Delta b_n}{\Delta \frac{b_n}{T_c} + \frac{s_n}{T_c}} = \frac{1}{1+\frac{s_n}{\Delta b_n}} T_c$ □

Remark In order to adapt this formalism for polyphonic pieces, one should also consider a shift in terms of onsets in addition to the shifts in terms of duration considered here.

Depending on the use of this data, one can choose to adapt the definition of (s_n), for instance by normalizing its values or cutting all shifts that represent over a certain portion of their duration. Proposition IV.3 illustrates how such definitions of (s_n) allow for guarantees on deviations of the canonical tempo for the *normalized* performance with respect to T_c.

B. Monophonic performance generation

By considering the set of all shifts obtained from various performances, that one can see as a rhythmic language, and trying to generate a word of this language (i.e., a rhythmic template for a performance), we can generate rough performances from a given score. A few examples of such generated pieces are to be found on [18]. A more subtle method could be to train some algorithms to reproduce rhythmic patterns within a database depending on the context of a particular piece (rhythmic (un)stability, begining of a phrase, cadence...), as score features may sometimes eclipse individual styles [7].

C. Quantitative musicological analysis

A formal definition for time-shifts allows for a quantitative, and statistical analysis of human performances regarding the latter. The analysis of quantitative variations of the canonical tempo at the end of a phrase, or depending on the kind of cadence already exist within literature [9]–[11], and could be complemented with such a study.

D. MIDI transcription

The methods presented in Section III could be used as preprocessing for transcription problems, allowing to add complementary data with respect to the transcription (such as tempo stability, consistency, and potentially tempo changes).

V. Conclusion & perspectives

We presented here some methods for monophonic tempo analysis, with applications to data augmentation (Section IV.A), rough performance generation (Section IV.B), and transcription. The latter could allow for further development in addition to a generative grammar for global consistency, as a multi-criteria optimisation [22] including tempo stability, specifically when coupled by a method similar to Section III.C. In particular, a dynamic programming algorithm could be used to detect potential abrupt tempo changes, or quasi-constant tempo sections. Even though we focused here on monophonic works, or rather on monophic views of performances, the very definition of canonical tempo presented here is more fit for monophonic pieces than polyphonic ones. The reader shall find some extensions of this formalism suited for polyphonic works in Appendix A.

Section III.C also presented a model which appears to share some similarities with the *Large et al.* model (Section II.C), modified as a score-based method. Therefore, studying the formalism for a quantizer might allow to obtain some theoretical results for the previous model, regarding for instance time of convergence guarantee and meaningfulness of the result.

Finally, the most immediate case of use of Section IV.A is fuzz testing for algorithm in tasks such as beat-tracking, transcription or tempo inference. In fact, this section presented a way to "midify" an actual performance, that we can see as a function from the set of musical human performances to the set of plain and flat midi files as generated by standard softwares. Being able to reverse this function could then allow for generating convincing performances. We presented in Section IV.B a rough way to extrapolate such data augmentation for performance generation, that could be improved with formal langage algorithms or machine learning techniques.

References

[1] S. D. Peter, C. E. Cancino-Chacón, E. Karystinaios, and G. Widmer, "Sounding Out Reconstruction Error-Based Evaluation of Generative Models of Expressive Performance," in *Proceedings of the 10th International Conference on Digital Libraries for Musicology*, 2023, pp. 58–66.

[2] C. Raphael, "A Probabilistic Expert System for Automatic Musical Accompaniment," *Journal of Computational and Graphical Statistics*, vol. 10, no. 3, pp. 487–512, Sep. 2001, doi: 10.1198/106186001317115081.

[3] E. Nakamura, N. Ono, S. Sagayama, and K. Watanabe, "A Stochastic Temporal Model of Polyphonic MIDI Performance with Ornaments," *Journal of New Music Research*, vol. 44, no. 4, pp. 287–304, Oct. 2015, doi: 10.1080/09298215.2015.1078819.

[4] E. Nakamura, T. Nakamura, Y. Saito, N. Ono, and S. Sagayama, "Outer-Product Hidden Markov Model and Polyphonic MIDI Score Following," *Journal of New Music Research*, vol. 43, no. 2, pp. 183–201, Apr. 2014, doi: 10.1080/09298215.2014.884145.

[5] E. W. Large and M. R. Jones, "The dynamics of attending: How people track time-varying events," *Psychological Review*, vol. 106, no. 1, pp. 119–159, 1999, doi: 10.1037/0033-295X.106.1.119.

[6] H.-H. Schulze, A. Cordes, and D. Vorberg, "Keeping Synchrony While Tempo Changes: Accelerando and Ritardando," *Music Perception: An Interdisciplinary Journal*, vol. 22, no. 3, pp. 461–477, 2005, doi: 10.1525/mp.2005.22.3.461.

[7] O. F. B. Katerina Kosta Rafael Ramírez and E. Chew, "Mapping between dynamic markings and performed loudness: a machine learning approach," *Journal of Mathematics and Music*, vol. 10, no. 2, pp. 149–172, 2016, doi: 10.1080/17459737.2016.1193237.

[8] K. Shibata, E. Nakamura, and K. Yoshii, "Non-local musical statistics as guides for audio-to-score piano transcription," *Information Sciences*, vol. 566, pp. 262–280, Aug. 2021, doi: 10.1016/j.ins.2021.03.014.

[9] "MazurkaBL: Score-aligned Loudness, Beat, and Expressive Markings Data for 2000 Chopin Mazurka Recordings." Accessed: Jun. 18, 2024. [Online]. Available: https://zenodo.org/records/1290763

[10] J. Hentschel, M. Neuwirth, and M. Rohrmeier, "The Annotated Mozart Sonatas: Score, Harmony, and Cadence," vol. 4, no. 1, pp. 67–80, May 2021, doi: 10.5334/tismir.63.

[11] P. Hu and G. Widmer, "The Batik-plays-Mozart Corpus: Linking Performance to Score to Musicological Annotations." Accessed: Jun. 18, 2024. [Online]. Available: http://arxiv.org/abs/2309.02399

[12] M. Müller, "Memory-restricted Multiscale Dynamic Time Warping," [Online]. Available: https://www.academia.edu/25724042/MEMORY_RESTRICTED_MULTISCALE_DYNAMIC_TIME_WARPING

[13] E. Nakamura, K. Yoshii, and H. Katayose, "Performance Error Detection and Post-Processing for Fast and Accurate Symbolic Music Alignment," 2017. Accessed: Jun. 18, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Performance-Error-Detection-and-Post-Processing-for-Nakamura-Yoshii/37e9f5e23cada918c2b8982d71a18972140d9d5a

[14] D. Murphy, "Quantization revisited: a mathematical and computational model," *Journal of Mathematics and Music*, vol. 5, no. 1, pp. 21–34, Mar. 2011, doi: 10.1080/17459737.2011.573674.

[15] G. Romero-García, C. Guichaoua, and E. Chew, "A Model of Rhythm Transcription as Path Selection through Approximate Common Divisor Graphs," May 2022. Accessed: Jun. 19, 2024. [Online]. Available: https://hal.science/hal-03714207

[16] F. Foscarin, A. Mcleod, P. Rigaux, F. Jacquemard, and M. Sakai, "ASAP: a dataset of aligned scores and performances for piano transcription," Oct. 2020. Accessed: Jun. 18, 2024. [Online]. Available: https://cnam.hal.science/hal-02929324

[17] S. D. Peter *et al.*, "Automatic Note-Level Score-to-Performance Alignments in the ASAP Dataset," vol. 6, no. 1, pp. 27–42, Jun. 2023, doi: 10.5334/tismir.149.

[18] S. Meunier, "Report github," Jun. 2024, [Online]. Available: https://github.com/sylvain-meunier/stageL3

[19] A. Cont, "A coupled duration-focused architecture for real-time music-to-score alignment," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 6, pp. 974–987, 2009.

[20] E. W. Large *et al.*, "Dynamic models for musical rhythm perception and coordination," *Frontiers in Computational Neuroscience*, vol. 17, May 2023, doi: 10.3389/fncom.2023.1151895.

[21] J. D. Loehr, E. W. Large, and C. Palmer, "Temporal coordination and adaptation to rate change in music performance," *Journal of Experimental Psychology. Human Perception and Performance*, vol. 37, no. 4, pp. 1292–1309, Aug. 2011, doi: 10.1037/a0023102.

[22] F. Foscarin, F. Jacquemard, P. Rigaux, and M. Sakai, "A parse-based framework for coupled rhythm quantization and score structuring," in *Mathematics and Computation in Music: 7th International Conference, MCM 2019, Madrid, Spain, June 18–21, 2019, Proceedings 7*, 2019, pp. 248–260.

Appendix A : Some formal considerations

A. Proof of Proposition II.2

Let $n \in \mathbb{N}$, $\int_{t_0}^{t_n} T(t)\,\mathrm{d}t = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} T(t)\,\mathrm{d}t$

Furthermore, $\int_{t_n}^{t_{n+1}} T(t)\,\mathrm{d}t = \int_{t_0}^{t_{n+1}} T(t)\,\mathrm{d}t + \int_{t_n}^{t_0} T(t)\,\mathrm{d}t$

$= \int_{t_0}^{t_{n+1}} T(t)\,\mathrm{d}t - \int_{t_0}^{t_n} T(t)\,\mathrm{d}t$

Let T be a formal tempo according to the first definition.
For all $n \in \mathbb{N}$, we then have :

$$\begin{aligned}\int_{t_n}^{t_{n+1}} T(t)\,\mathrm{d}t &= \int_{t_0}^{t_{n+1}} T(t)\,\mathrm{d}t - \int_{t_0}^{t_n} T(t)\,\mathrm{d}t \\ &= b_{n+1} - b_0 - (b_n - b_0) \\ &= b_{n+1} - b_n\end{aligned}$$

Let T be a formal tempo according to the second definition.
For all $n \in \mathbb{N}$:

$$\begin{aligned}\int_{t_0}^{t_n} T(t)\,\mathrm{d}t &= \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} T(t)\,\mathrm{d}t \\ &= \sum_{i=0}^{n-1} b_{i+1} - b_i \\ &= b_n - b_0\end{aligned}$$

We thus obtain the two implications, hence the equivalence stated in Proposition II.2.

B. The ***tempo octave*** problem

When defining tempo, or transcribing a performance, there always exist several equivalent possibilities. For instance, given a "correct" transcription (b_n) of a performance (t_n), one can choose to define its own transcription as $t = \left(\frac{b_n}{2}\right)$.
Then, the canonical tempo with respect to t, called (T_1^*), and the one with respect to (b_n), called (T_2^*) verify :
$\forall n \in \mathbb{N}, T_{1,n}^* = \frac{\frac{1}{2}b_{n+1} - \frac{1}{2}b_n}{t_{n+1} - t_n} = \frac{1}{2}T_{2,n}^*$. Actually, the transcription t corresponds to (b_n) where all durations are indicated doubled, but played twice faster, hence giving the exact same theoretical performance. Unfortunately, there is no absolute way to decide which of these two transcription is better than the other. This problem is known as the tempo octave problem, and should be kept in mind when transcribing, or estimating tempo. We present in Section III.A a model robust to these tempo octaves, and other kind of octaves not discussed here (for instance multiplying the tempo by 3 using triplet).

C. Tempo conservation when reversing time

First, we want to insist on the fact that none of the sequences (b_n) and (t_n) are infinite, but in order to simplify the notation, we chose to indicate them as usual infinite sequences, or rather only consider them on a finite number of indexes, refered to as $|(b_n)|$ and $|(t_n)|$ respectively, with $|(b_n)| = |(t_n)|$. Let us then define the reversed sequence of (u_n) as $r((u_n)) := \left(\overline{u_n} = u_{|(u_n)|} - u_{|(u_n)|-n}\right)_{n \in [\![0,\ |(u_n)|]\!]}$

Both $\overline{b} = r((b_n))$ and $\overline{t} = r((t_n))$ are correct representations of a music score and performance respectively, as defined in Section II.A.
Let $q = |(t_n)|$, $t^* = t_q$, T a formal tempo with respect to (t_n) and (b_n), $n \in [\![0, q-1]\!]$ and $T_r : t \mapsto T(t^* - t)$.

$$\begin{aligned}\int_{\overline{t_n}}^{\overline{t_{n+1}}} T_r(t)\,\mathrm{d}t &= \int_{t^*-t_{q-n}}^{t^*-t_{q-n-1}} T(t^* - t)\,\mathrm{d}t = \int_{t_{q-n}}^{t_{q-n-1}} -T(x)\,\mathrm{d}x \\ &= \int_{t_{q-n-1}}^{t_{q-n}} T(t)\,\mathrm{d}t \\ &= b_{q-n} - b_{q-n-1} \\ &= \left(b_{q-n} - b_q\right) - \left(b_{q-n-1} - b_q\right) \\ &= -\overline{b_n} + \overline{b_{n+1}} \\ &= \overline{b_{n+1}} - \overline{b_n}\end{aligned}$$

Hence T_r is a formal tempo with respect to $(\overline{t_n})$ and $\left(\overline{b_n}\right)$.

D. Musical explanation of the choice of a tempo distance

In terms of tempo, halving and doubling are considered as far as each other from the initial value. Therefore a usual absolute distance does not fit this notion, and we will rather use a logarithmic distance when comparing tempi.

E. Monophony and polyphony

Definition II.3 is valid when considering monophonic pieces. In order to adapt the formalism for polyphony, one should consider the sequence (b_n) to be increasing (but not necessarily strictly), and instead of using $b_{n+1} - b_n$, the sequence (Δb_n) embodying the different durations of the events should be defined. The same modifications applied to (t_n) allow for defining a polyphonic performance of a polyphonic piece. However, in such a piece, the tempo may vary between the different voices, hence different formalisms for tempo may be defined, especially for a formal tempo curve.

Nonetheless, according to Section III.C and Section IV.A, defining a single global tempo for all voices, and considering deviations as timings may be the easiest way to extend our formalism, although more studies should be done in order to verify this assumption.

Therefore, the canonical tempo for a polyphonic performance (t_n) of a polyphonic piece (b_n) is defined as $T_n^* = \frac{\Delta b_n}{\Delta t_n}$, for all $n \in \mathbb{N}$ in all our polyphonic works.

Appendix B : Estimator Model

A. Formal explanations and proofs

First d is indeed a mathematical distance : let $a, b \in (\mathbb{R}_+^*)^2$, $d(a,b) = d(b,a)$ and $d(a,b) = 0 \Leftrightarrow \left|\log\left(\frac{a}{b}\right)\right| = 0 \Leftrightarrow a = b$. Finally, let $c \in \mathbb{R}_+^*$, $d(a,c) = k_* \left|\log\left(\frac{a}{c}\right)\right| = k_* \left|\log\left(\frac{a}{b} \times \frac{b}{c}\right)\right| = k_* \left|\log\left(\frac{a}{b}\right) + \log\left(\frac{b}{c}\right)\right| \leq d(a,b) + d(b,c)$.

Then, formula of the model is valid on a monophic context, where all the grace notes are explicit (in other words, (b_n) is strictly increasing).

The estimator E is not exactly a function in practice. Its actual expression is only supposed to remains the same between two computations of T_n, in order for the argmin to make sens, as explained hereafter. (9) presents an argmin, that makes sense when E is a increasing right-continuous function-like object, even though its actual expression may change after each computed value of T_{n+1}. In fact, E can only output a countable set of values, hence E is piecewise constant under those hypothesis.

Finally, one can notice that the value of $k_* \in \mathbb{R}_+^*$ does not affect the result of the process.

B. *About the range* $\left[\frac{\sqrt{2}}{2}T_n, \sqrt{2}T_n\right]$

In order to resist to the tempo octave problem discussed in Appendix A, we choose here to consider a unique candidate within a range $[x, 2x] \subset \left[\frac{1}{2}, 2\right]$, for a given $x \in \mathbb{R}_+^*$. Then, we want this range to be centered around 1, since its values corresponds to tempo variation, and our system should not favor increasing nor decreasing the tempo *a priori*. For this musical reason, we then take x as solution of : $\|x - 1\| = \|2x - 1\|$ that implies $1 - x = 2x - 1$, i.e., $x = \frac{2}{3}$ with the absolute value distance.
With a logarithmic distance, the same reasoning would give : $\log\left(\frac{1}{x}\right) = \log(2x) \Leftrightarrow -\log(x) = \log(2) + \log(x) \Leftrightarrow \log(x^2) = -\log(2) \Leftrightarrow x^2 = \frac{1}{2} \Leftrightarrow x = \frac{\sqrt{2}}{2}$ since $x > 0$.

Then, when considering the tempo distance between T_{n+1} and T_n, we find :

$$d(T_{n+1}, T_n) = k_*|\log(y)| = d(1, y) \tag{12}$$

where $y = \underset{x' \in [x,2x]}{\operatorname{argmin}}\, d\left(x', \frac{\Delta t_n}{\Delta t_{n+1}} E\left(x' \frac{\Delta t_{n+1}}{\Delta t_n}\right)\right)$.
Therefore, since we want the extreme possible values of our range to imply an equal distance between T_n and T_{n+1}, we choose the logarithmic distance, and hence $x = \frac{\sqrt{2}}{2}$, so that $d(1, x) = d(1, 2x)$.

In fact, those two distances give the same results according to the measure. We favor the logarithmic since it embodies more musical meaning.

C. *About the estimator* E

One can notice that $E = \mathrm{id}$ implies, by the hypothesis that E acts as an oracle, that the theoretical and actual values are the same, or that the performance is a perfect interpretation of the piece. Since real players do not make such performance, we can expect a relevant estimator to act rather differently than the identity function.

Moreover, E is not a function : its expression only has to be fixed when computing the numerical resolution for the argmin. Hence, an given output can depends on several previous outputs. In an extreme case, E can even be a transcripting system. However, in our problem of tempo estimation, we do not have as much constraints as in transcription.
Indeed, the following figures displays two transcription A and B and their corresponding tempo curves. The latter transcription is actually incorrect with regards to usual transcription convention, with inconsistent duration for a measure that do not match the indicated time signature, and increased reading complexity with respect to transcription A.

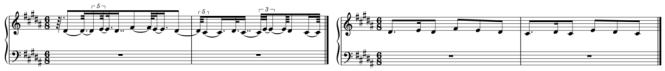

Figure 10: Transcription A

Figure 11: Transcription B

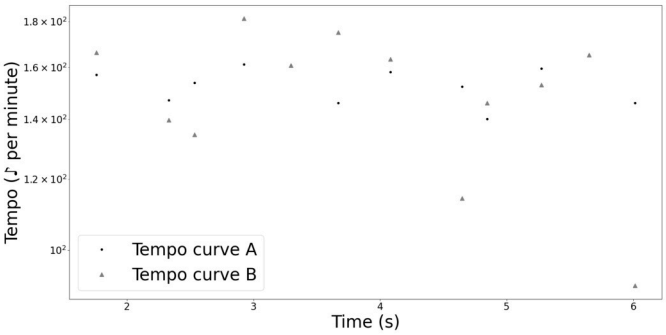

Figure 12: Tempo curves A and B plotted together

One can notice that these tempo curves are quite similar, and in fact, a human being could not tell them apart, as shown by Figure 13.

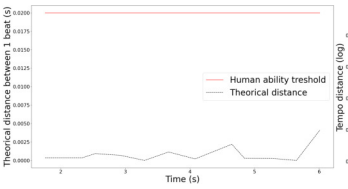

Figure 13: Tempo distance (s)

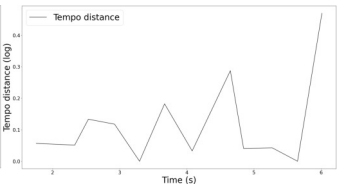

Figure 14: Tempo distance (log)

Tempo distance between the two previous curves. Being able to differentiate them would imply to tell apart two rhythmic events within 4 ms, which is supposed impossible for a human being according to the value of ε defined in Section II.A (and displayed as the top line in Figure 13).

D. *Formal study of the model*

Since this approach fundamentally search to estimate tempo variation rather than actual values, it is not easy to visualize the relevance of the result by naive means. We choose here to define the sequence of ratios $\left(\alpha_n := \frac{T_n}{T_n^*}\right)_{n \in [\![1,N]\!]}$, and then$(\tilde{\alpha}_n := \exp(\ln(\alpha_n) - \lfloor \log_2(\alpha_n) \rfloor \ln(2)))_{n \in [\![1,N]\!]}$. The latter is called the *normalized* sequence of ratio, where each

value is uniquely determined within the range $[\tilde{1}, \tilde{2}[$. Such a choice allows for merging together the tempo octaves, as explained in Appendix A. One can notice that adding $\tilde{1}$ to a normalized value is equivalent to multiplying the initial value by 2. We now define a *spectrum* $S = (\tilde{\alpha}_n)_{n\in[\![1,N]\!]}$, and call $|S|$ the value $N \in \mathbb{N}$. Finally, we define $\mathcal{C}$ the range $[\tilde{1}, \tilde{2}[$ seen as a circle according to the following application : $c : [\tilde{1}, \tilde{2}[\to \mathcal{C}(0,1)$

$\tilde{x} \mapsto (\cos(2\pi\tilde{x}), \sin(2\pi\tilde{x}))$, so that $c(\tilde{1}) = c(\tilde{2})$

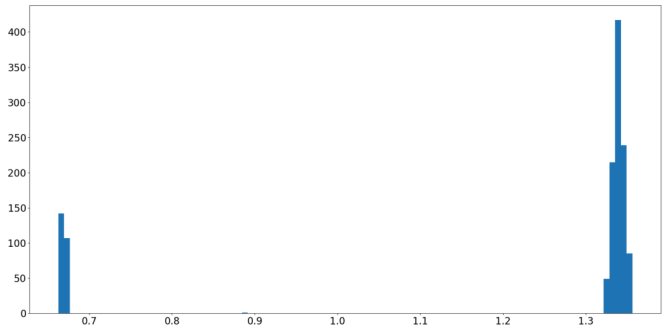

Figure 15: Distribution of (α_n) for a monophic piece

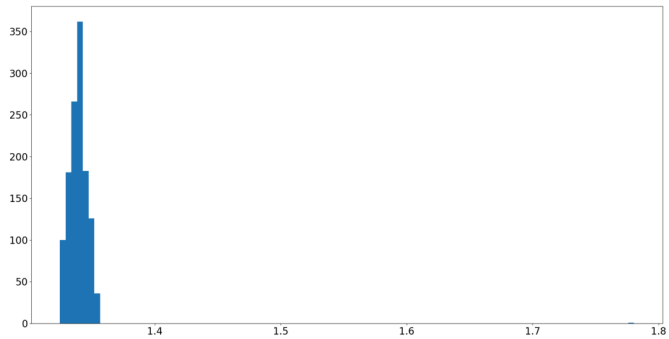

Figure 16: Distribution of the spectrum $(\tilde{\alpha}_n)$ for the same piece

Definition N.1 Let S be a spectrum, and $\Delta \in \left[0, \frac{1}{2}\right]$.
We define as follow the *measure* of S with imprecision Δ, that embodies a standard deviation on $\mathcal{C}$:

$$m(S,\Delta) = \max_{\tilde{d}\in\mathcal{C}} \frac{\left|\left\{n \in [\![1,|S|]\!] : d(\tilde{\alpha}_n, \tilde{d}) \leq \Delta\right\}\right|}{\min(1,|S|)} \quad (13)$$

Here d is still a logarithmic distance with $k_* = \frac{1}{\ln(2)}$, slightly modified on $\mathcal{C}$ to be consistent with $d(\tilde{1},\tilde{2}) = 0$.
Actually, $d : \tilde{a}, \tilde{b} \mapsto \min\left(\left|\log_2\left(\frac{\tilde{a}}{\tilde{b}}\right)\right|, 1 - \left|\log_2\left(\frac{\tilde{a}}{\tilde{b}}\right)\right|\right)$ on $\mathcal{C}$.

Proposition N.2 Let $a, b \in (\mathbb{R}_+^*)^2$, $\widetilde{a^{-1}} \times \tilde{a} = \tilde{1}$, and $\widetilde{ab} = \tilde{a}\tilde{b}$

Definition N.3 Let S be a spectrum and $\lambda \in \mathbb{R}_+^*$, we define the rotation of S by λ as : $\lambda S = \left(\widetilde{\lambda S_n}\right)_{n\in[\![1,\,|S|]\!]}$

Proposition N.4 Let $\Delta \in \left[0, \frac{1}{2}\right]$, $\lambda \in \mathbb{R}_+^*$ and S a spectrum. Let S' be the spectrum of the same initial values as S, but normalized within the interval $[\tilde{\lambda}, \widetilde{2\lambda}[$ instead of $[\tilde{1}, \tilde{2}[$, and $S^{-1} := \left(\widetilde{\frac{1}{\alpha_n}}\right)_{n\in[\![1,\,|S|]\!]}$

Then :

- $m(S',\Delta) = m(S,\Delta) = m(S^{-1},\Delta) = m(\lambda S,\Delta)$
- $0 \leq m(S,\Delta) \leq 1$
- $m(S,\Delta) = 0 \Leftrightarrow |S| = 0$
- $m(S,\Delta) = 1 \Leftrightarrow \forall(\tilde{a},\tilde{b}) \in S^2, d(\tilde{a},\tilde{b}) \leq 2\Delta$

This *measure* allows to quantify the quality of this model, without considering tempo octaves, or equivalently to quantify the quality of the estimator. A C++ implementation of this measure is available on [18], as well as the detailled performance of our test model over the (n)-ASAP dataset. Figure 17 presents those results in a global display. The blue values corresponds to the pieces written by W.A Mozart, that usually contain mainly regular division, and thus are expected to produce a *measure* closer to 1 than average. On the other hand, the red values corresponds to M. Ravel's pieces, much more rhythmically expressive, hence we expect a *measure* closer to 0. To understand the global results, we present in Figure 23 the same results for a random estimator.

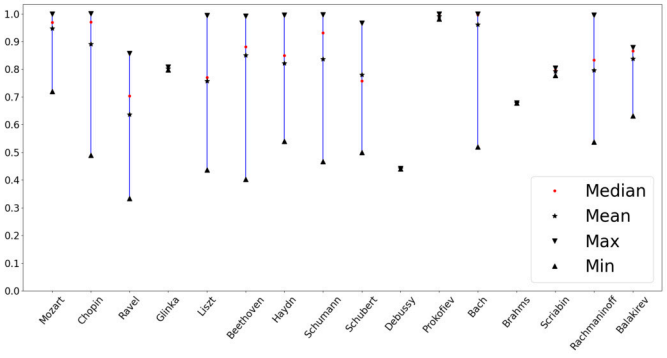

Figure 17: Measures of the resulting spectrum over the whole (n)-ASAP dataset with $\Delta = 0.075$, with naive estimator.

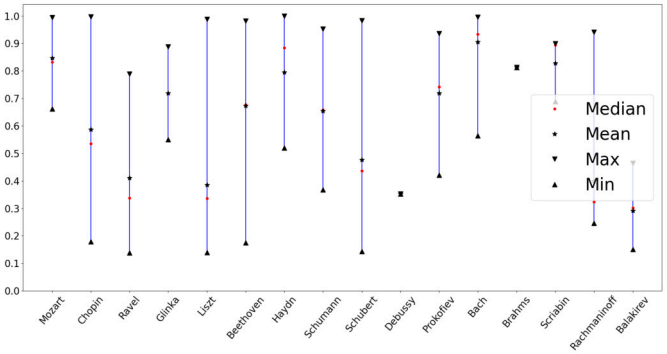

Figure 18: Measures of the resulting spectrum over the whole (n)-ASAP dataset with $\Delta = 0.075$, with naive estimator.

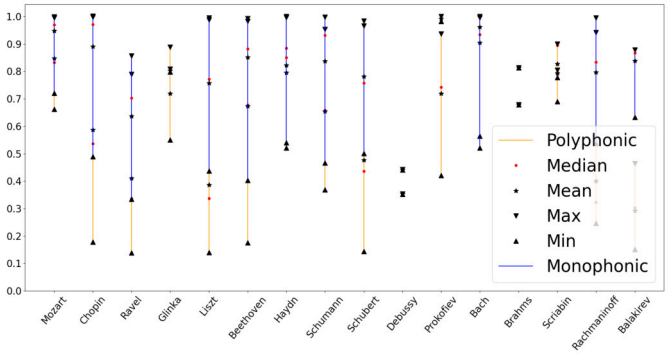

Figure 19: Comparison between the two latters.

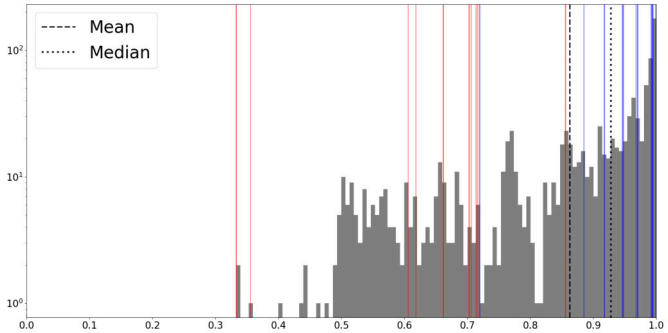

Figure 20: Measures of the resulting spectrums over the whole (n)-ASAP dataset with $\Delta = 0.025$, with 63% over mean.

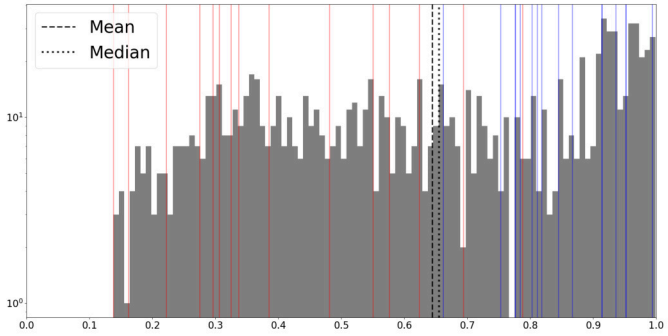

Figure 21: Measures of the resulting spectrums over the whole (n)-ASAP dataset with $\Delta = 0.025$, with 51% over mean

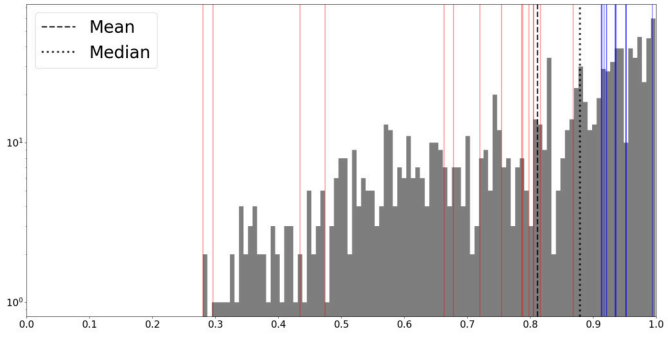

Figure 22: Measures of the resulting spectrums over the whole (n)-ASAP dataset with $\Delta = 0.075$, with 62% over mean

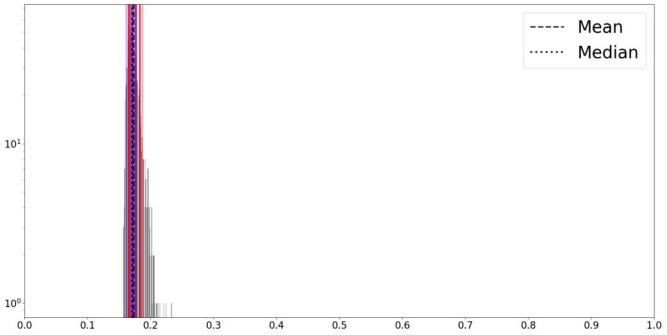

Figure 23: Measures of the resulting spectrums over the whole (n)-ASAP dataset with $\Delta = 0.075$, with an estimator outputing random quantized values.

Figure 17 shows results indicating that our naive estimator performs well on pieces with regular division and small tempo changes, which is typically the style of the classical era. The random estimator could actually output regular division, and to a lesser extent triplets (with a lower probability). Hence, it performed better than the naive one for a few pieces, especially those containing such irregular divisions. Finally, our naive estimator has 34.2% of its value strictly over its average, whereas this value is 27.6% for the random estimator.

Absolute distance and log distance presents the same results regarding the measure value.

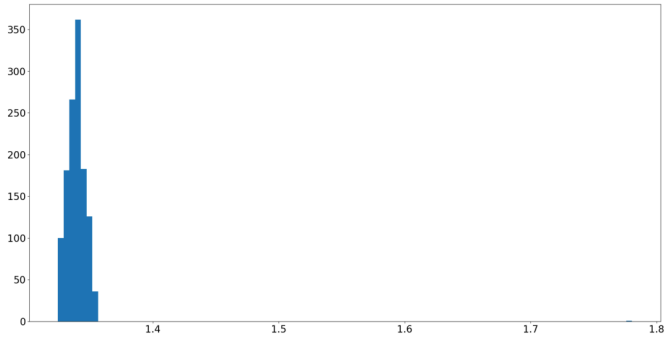

Figure 24: Example of a spectrum for a performance of a Mozart piece with our naive estimator in a monophic context

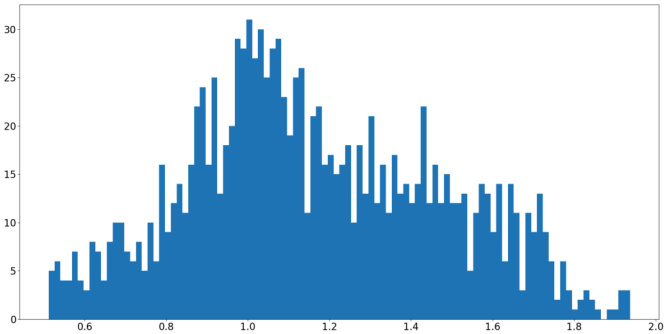

Figure 25: Example of a spectrum with random estimator. Such a spectrum actually reflects the distribution of ratios throughout the piece's actual score

Appendix C : Quantized Model

In this section we use the notation introduced by G. Romero-García, C. Guichaoua, and E. Chew [15], that essentially replace D by T.

Let us first define : $g : x \mapsto \min(x - \lfloor x \rfloor, 1 + \lfloor x \rfloor - x)$

One can make sure that $g : x \mapsto \begin{cases} x - \lfloor x \rfloor \text{ si } x - \lfloor x \rfloor \leq \frac{1}{2} \\ 1 - (x - \lfloor x \rfloor) \text{ sinon} \end{cases}$ and that g is 1-periodic, continuous on $\mathbb{R}$.

Then, by definition :

$$\begin{aligned} \varepsilon_T(a) &= \max_{t \in T} \Big(\min_{m \in \mathbb{Z}} |t - ma| \Big) \\ &= \max_{t \in T} \min\big(t - \lfloor \tfrac{t}{a} \rfloor a, (\lfloor \tfrac{t}{a} \rfloor + 1)a - t\big) \\ &= a \max_{t \in T} \underbrace{\min\left(\frac{t}{a} - \left\lfloor \frac{t}{a} \right\rfloor, \left\lfloor \frac{t}{a} \right\rfloor + 1 - \frac{t}{a} \right)}_{g(\frac{t}{a})} \\ &= a \max_{t \in T} g\left(\frac{t}{a} \right) \end{aligned}$$

hence we have proved ε_T to be continous on $\mathbb{R}_+^*$.

Furthermore, for $n \in \mathbb{N}^*, T \subset (\mathbb{R}_+^*)^n, a \in \mathbb{R}_+^*$, $\varepsilon_T(a) = a \max_{t \in T} g(\frac{t}{a}) = a \times 1 \max_{t \in T} g(\frac{t}{a} 1^{-1}) = a \varepsilon_{T/a}(1)$.

Hence the intuitive following result : the smaller the tatum, the smaller the bound of the error.

A. Characterization of semi-strict local maxima

1) *First implication:*

Let a be a semi-strict local maximum of ε_T, $a > 0$. By definition, there is a $a > \varepsilon > 0$ so that :

$\forall \delta \in]-\varepsilon, \varepsilon[, \varepsilon_T(a) \geq \varepsilon_T(a + \delta)$.

Let $t \in \operatorname{argmax}_{t' \in T} g\left(\frac{t'}{a}\right)$, hence $\varepsilon_T(a) = a g(\frac{t}{a})$.

For all $\delta \in]-\varepsilon, \varepsilon[, \varepsilon_T(a) = a g(\frac{t}{a}) \geq \varepsilon_T(a + \delta)$,

and $\varepsilon_T(a + \delta) = (a + \delta) \max_{t' \in T} g\left(\frac{t'}{a+\delta}\right) \geq (a + \delta) g\left(\frac{t}{a+\delta}\right)$.

For $\delta \geq 0, a + \delta \geq a$, so $a g(\frac{t}{a}) \geq (a + \delta) g\left(\frac{t}{a+\delta}\right) \geq a g\left(\frac{t}{a+\delta}\right)$

Hence, $g(\frac{t}{a}) \geq g\left(\frac{t}{a+\delta}\right)$ since $a > 0$, for all $\delta \in [0, \varepsilon[$.

Therefore, g increases monotonically in the range $]\frac{t}{a+\delta}, \frac{t}{a}[$, since g has a unique local maximum (modulo 1), considering the previous range as a neighbourhood of $\frac{t}{a}$.

Hence, $g = x \mapsto x - \lfloor x \rfloor$ within the considered range, and $g(\frac{t}{a}) = \frac{t}{a} - \lfloor \frac{t}{a} \rfloor, \varepsilon_T(a) = t - a \lfloor \frac{t}{a} \rfloor$.

The function ε_T is the maximum of a finite set of continuous functions, with a countable set of A of intersection, i.e., $A = \left\{ x \in \mathbb{R}_+^* : \exists (t_1, t_2) \in T^2 : t_1 \neq t_2 \wedge x g\left(\frac{t_1}{x}\right) = x g\left(\frac{t_2}{x}\right) \right\}$. Indeed,

$$\begin{aligned} x \in A &\Leftrightarrow \exists (t_1, t_2) \in T^2 : t_1 \neq t_2 \wedge g\left(\tfrac{t_1}{x}\right) = g\left(\tfrac{t_2}{x}\right) \\ &\Leftrightarrow \exists (t_1, t_2) \in T^2 : t_1 \neq t_2 \wedge \tfrac{t_1}{x} = \pm \tfrac{t_2}{x} \bmod 1 \\ &\Leftrightarrow \exists (t_1, t_2) \in T^2 : t_1 \neq t_2 \wedge x = \tfrac{t_1 \mp t_2}{n}, n \in \mathbb{Z}^* \end{aligned}$$

Hence $A \subset \left\{ \frac{t_1 \mp t_2}{n}, (t_1, t_2) \in T^2, n \in \mathbb{Z}^* \right\}$, because $T \subset (\mathbb{R}_+^*)^{|T|}$. Therefore, there is a countable set of closed convex intervals, whose union is $\mathbb{R}_+^*$ so that on each of these intervals, ε_T is equal to $f_t : a \mapsto a g(\frac{t}{a})$ for a $t \in T$. Let then t be so that for all $x \in]a - \delta', a[, \varepsilon_T(x) = f_{t(x)}$, where $]a - \delta', a[$ is included in one the previous intervals. Since f_t and ε_T are both continuous on $[a - \delta', a]$, $f_t(a) = \varepsilon_T(a)$ and therefore, $t \in \operatorname{argmax}_{t' \in T} g\left(\frac{t'}{a}\right)$. The previous paragraph showed that g is increasing on a left neighbourhood of $\frac{t}{a}$. Therefore, on a right neighbourhood of $\frac{t}{a}$, g is either increasing or decreasing by its definition.

- if g is increasing on this neighbourhood, called $N(\frac{t}{a})^+$ in the following, the previous expression of g remains valid, i.e., $\forall x \in N(\frac{t}{a})^+, g(x) = x - \lfloor x \rfloor$. Moreover, $x \mapsto \lfloor x \rfloor$ is right-continuous, hence by restricting $N(\frac{t}{a})^+$, we can assure for all $x \in N(\frac{t}{a})^+, \lfloor x \rfloor = \lfloor \frac{t}{a} \rfloor$. Let then $y = a - \frac{t}{x}$ so that $x = \frac{t}{a-y}$, we then have $\varepsilon_T(a - y) = t - (a - y)\lfloor \frac{t}{a} \rfloor \leq \varepsilon_T(a) = t - a \lfloor \frac{t}{a} \rfloor$ because a is a local maximum of ε_T and $a - y$ is within a (left) neighbourhood of a, even if it means restricting δ' or $N(\frac{t}{a})^+$. Hence, $t - \lfloor \frac{t}{a} \rfloor a \geq t - (a - y)\lfloor \frac{t}{a} \rfloor$ i.e., $a \lfloor \frac{t}{a} \rfloor \leq (a - y)\lfloor \frac{t}{a} \rfloor \Leftrightarrow 0 \leq -y \lfloor \frac{t}{a} \rfloor$ i.e., $\lfloor \frac{t}{a} \rfloor \leq 0$ i.e., $\lfloor \frac{t}{a} \rfloor = 0$, since y, t and a are all positive values. Then, $a > t$ and therefore $\varepsilon_T(a - y) = t = \varepsilon_T(a)$ for $\lfloor \frac{t}{a} \rfloor = 0$. This interval where ε_T is constant is then either going on infinitely on the right of a, or else ε_T will reach a value greater than $\varepsilon_T(a) = t$, since ε_T can then be rewritten as $x \mapsto \max\left(\max_{t' \in T \setminus \{t\}} x g\left(\frac{t'}{x}\right), \varepsilon_T(a) \right)$ on $[a, +\infty[$. Hence a is a local minimum on the right, and since ε_T is constant on a left neighbourhood of a, a is also a local minimum on the left. Finally, a is a local minimum, which is absurd by definition.
- else, g is decreasing on $N(\frac{t}{a})^+$, $\frac{t}{a}$ is by definition a local maximum of g. However, g only has a unique local maximum modulo 1, that is $\frac{1}{2}$. Hence, $\frac{t}{a} = \frac{1}{2} \bmod 1$, i.e., $\frac{t}{a} = \frac{1}{2} + k, k \in \mathbb{Z}$, or $a = \frac{t}{k + \frac{1}{2}}, k \in \mathbb{N}$, since $a > 0$.

2) *Second implication:*

Let $(t, k) \in T \times \mathbb{N}, a = \frac{t}{k + \frac{1}{2}}$.

By definition : $g(\frac{t}{a}) = g(\frac{1}{2} + k) = g(\frac{1}{2}) = \frac{1}{2} = \max_{\mathbb{R}} g$.

Therefore, $\varepsilon_T(a) = a \max_{t' \in T} g\left(\frac{t'}{a}\right) = a g(\frac{t}{a}) = \frac{a}{2}$.

For all $x \in]0, a[, \varepsilon_T(x) = x \max_{t' \in T} g\left(\frac{t'}{x}\right) \leq \frac{x}{2} < \frac{a}{2} = \varepsilon_T(a)$, i.e., $\varepsilon_T(a) > \varepsilon_T(x)$.

Let $T^* = \left\{ t' \in T : g\left(\frac{t'}{a}\right) = \frac{1}{2} \right\}$. Since $t \in T^*, |T^*| > 1$.

Let $t^* \in T^*$. For all $t' \in T \setminus T^*, g\left(\frac{t'}{a}\right) < g\left(\frac{t^*}{a}\right)$.

Since $h_{t'} : x \mapsto g\left(\frac{t'}{x}\right) - g\left(\frac{t^*}{x}\right)$ is continuous in a neighbourhood of $a > 0$, we have the existence of $\varepsilon_{t'} > 0$ so that $h_{t'}$ is strictly positive within $[a, a + \varepsilon_{t'}[$.

Let $\varepsilon_{t^*} = \min_{t' \in T} \varepsilon_{t'}$ and finally $\varepsilon_1 = \min_{t^* \in T^*} \varepsilon_{t^*}$.

Let $(t_1, t_2) \in (T^*)^2$.
In the following, $N(a)^+$ is a right neighbourhood of a such that $a \notin N(a)^+$.

Let $\text{tmp} : x \mapsto g\left(\frac{t_1}{x}\right) - g\left(\frac{t_2}{x}\right)$ be a continous function on $N(a)^+$ and A be the set of all $x^* \in N(a)^+$ so that $\text{tmp}(x^*) = 0 \Leftrightarrow g\left(\frac{t_1}{x^*}\right) = g\left(\frac{t_2}{x^*}\right)$.

We have for all $x^* \in A$, $g\left(\frac{t_1}{x^*}\right) = g\left(\frac{t_2}{x^*}\right)$ by definition. Considering the expression of g, we then find : $\frac{t_1}{x^*} = \pm\frac{t_2}{x^*} \bmod 1$. Moreover, since g only reach $g\left(\frac{t_1}{a}\right) = \frac{1}{2}$ once per period, we have $\frac{t_1}{a} = \frac{t_2}{a} \bmod 1$, i.e., $|\frac{t_1}{a} - \frac{t_2}{a}| = k_a \in \mathbb{N}$.

Then, $\frac{t_1}{x^*} = \pm\frac{t_2}{x^*} \bmod 1$ i.e., $|\frac{t_1}{x^*} \mp \frac{t_2}{x^*}| = k_* \in \mathbb{N}$, and therefore $|t_1 \mp t_2| = ak_a = x^*k_*$, and $x^* > a$ implies $k_a > k_* \geq 0$. However, $x^* = \frac{|t_1 \mp t_2|}{k_*}$, hence A is finite if $A \neq \emptyset$, and $\emptyset$ is a finite set. Finally, A is a finite set, i.e., $|A| \in \mathbb{N}$.

Let then $x_{t_1,t_2} = \begin{cases} \min A \text{ if } A \neq \emptyset \\ x \in N(a)^+ \setminus \{a\} \text{ otherwise} \end{cases}$ WLOG, $g\left(\frac{t_1}{x}\right) \geq g\left(\frac{t_2}{x}\right) \forall x \in \left[a, x_{t_1,t_2}\right]$

Let $a_2 = \min_{(t_1,t_2)\in T^{*2}} \underbrace{x_{t_1,t_2}}_{>a}$ and $a_1 \in]a, a_2[$,

let $t^* = \underset{t' \in T^*}{\text{argmax}}\, g\left(\frac{t'}{a_1}\right)$ We finally have $\forall x \in]a, a_2[, g\left(\frac{t^*}{x}\right) \geq g\left(\frac{t'}{x}\right), \forall t' \in T^*$.

Let then $\tilde{a} = \min(a + \varepsilon_1, a_2)$ so that for all $x \in]a, \tilde{a}[, t' \in T, g\left(\frac{t^*}{x}\right) \geq g\left(\frac{t'}{x}\right)$, hence $\varepsilon_T(x) = xg\left(\frac{t^*}{x}\right)$.
Let $f : x \mapsto g\left(\frac{t^*}{x}\right), f(a) = g\left(\frac{t^*}{a}\right) = \frac{1}{2}$ because $t^* \in T^*$ hence f is increasing on a right neighbourhood of a, $N(a)^+$, since $\frac{1}{2}$ is a global maximum of g, therefore g is increasing on $N\left(\frac{t^*}{a}\right)^-$ a left neighbourhood of $\frac{t^*}{a}$, i.e., f is increasing on $N(a)^+$. Therefore, we know that $g\left(\frac{t^*}{x}\right) = \frac{t^*}{x} - \left\lfloor \frac{t^*}{x} \right\rfloor$, since the only other possible expression for g would imply a decreasing function on $N\left(\frac{t^*}{a}\right)^-$.

Hence, $\varepsilon_T(x) = xg\left(\frac{t^*}{x}\right) = x\left(\frac{t^*}{x} - \left\lfloor \frac{t^*}{x} \right\rfloor\right) = t^* - x\left\lfloor \frac{t^*}{x} \right\rfloor$ and $\varepsilon_T(a) = t^* - a\left\lfloor \frac{t^*}{a} \right\rfloor$ since ε_T is continuous on $\mathbb{R}_+^*$.
By definition : $\left\lfloor \frac{t^*}{a} \right\rfloor \leq \frac{t^*}{a} < \left\lfloor \frac{t^*}{a} \right\rfloor + 1$.
However, $f(a) = \frac{1}{2} = \frac{t^*}{a} - \left\lfloor \frac{t^*}{a} \right\rfloor$, therefore $\left\lfloor \frac{t^*}{a} \right\rfloor < \frac{t^*}{a}$.
Then, there is $\alpha \in \mathbb{R}_+^*$ so that $\left\lfloor \frac{t^*}{a} \right\rfloor < \alpha < \frac{t^*}{a}$, let $y = \frac{t^*}{\alpha}$, i.e., $\alpha = \frac{t^*}{y}$, with $y > a$.
Let $a' = \min(y, \tilde{a})$ and $k = \left\lfloor \frac{t^*}{a} \right\rfloor$. For all $x \in]a, a'[$,

- $x \leq y = \frac{t^*}{\alpha}$ hence $\left\lfloor \frac{t^*}{a} \right\rfloor \leq \alpha \leq \frac{t^*}{x}$
- $x \geq a$ hence $\frac{t^*}{x} \leq \frac{t^*}{a} < \left\lfloor \frac{t^*}{a} \right\rfloor + 1$

In the end, $\left\lfloor \frac{t^*}{x} \right\rfloor = \left\lfloor \frac{t^*}{a} \right\rfloor = k$ by definition.
Then, $\varepsilon_T(x) = t^* - xk$ and $\varepsilon_T(a) = t^* - xa$, with $a < x$.
Finally, $\varepsilon_T(a) > \varepsilon_T(x)$.
To conclude, for all $x \in]0, a'[$,

- if $x \leq a$, $\varepsilon_T(a) \geq \varepsilon_T(x)$
- if $x \geq a, x \in [a, a'[$ and $\varepsilon_T(a) \geq \varepsilon_T(x)$

Hence a is a semi-strict local maximum of ε_T, and then the set of all semi-strict local maxima is $M_T = \left\{ \frac{t}{k+\frac{1}{2}}, t \in T, k \in \mathbb{N} \right\}$.

Finally, with the notations introduced in Section III.B, we proved (11)

□

B. Necessary condition to be a semi-strict local minimum

Let a be a semi-strict local minimum of ε_T.
There exists $\varepsilon > 0 : \forall \delta \in]-\varepsilon, \varepsilon[, \varepsilon_{T(a+\delta)} \geq \varepsilon_{T(a)}$.
Thanks to the previous results, we know $\varepsilon_{T(a)} < \frac{a}{2}$ since otherwise, $a \in M_T$, hence a is a strict local maximum.
We now consider two neighbourhoods of a : $N(a)^+$ and $N(a)^-$. Similarly to the previous proof, we can define $(t^+, t^-) \in T^2$ so that :

- for all $x \in N(a)^+, \varepsilon_T(x) = xg\left(\frac{t^+}{x}\right)$
- for all $x \in N(a)^-, \varepsilon_T(x) = xg\left(\frac{t^-}{x}\right)$

We can then consider $\varepsilon' > 0$ so that, for all $\delta \in]0, \varepsilon'[$:

- $\varepsilon_T(a + \delta) = (a + \delta)g\left(\frac{t^+}{a+\delta}\right) \geq \varepsilon_T(a) = ag\left(\frac{t^+}{a}\right)$
- $\varepsilon_T(a - \delta) = (a - \delta)g\left(\frac{t^-}{a-\delta}\right) \geq \varepsilon_T(a) = ag\left(\frac{t^-}{a}\right)$

However $a - \delta < a$, hence $g\left(\frac{t^-}{a-\delta}\right) > g\left(\frac{t^-}{a}\right)$, or in other words, g is increasing on $]\frac{t^-}{a}, \frac{t^-}{a-\delta}[$ since $\frac{t^-}{a-\delta} > \frac{t^-}{a}$, and g is stepwise monotonic. Note that this implies ε_T is decreasing on $N(a)^-$, which is conveniently consistent with the hypothesis of a being a local minimum of the latter.
Then, we have : $g\left(\frac{t^-}{a-\delta}\right) = \frac{t^-}{a-\delta} - \left\lfloor \frac{t^-}{a-\delta} \right\rfloor$
Moreover, since the functions considered here are all continuous on $]a - \varepsilon', a + \varepsilon'[$, we have $g\left(\frac{t^-}{a}\right) = g\left(\frac{t^+}{a}\right) = \frac{1}{a}\varepsilon_T(a)$.
If g were increasing on $]\frac{t^+}{a+\delta}, \frac{t^+}{a}[$, then a is also a local maximum of ε_T according to the first implication of the previous proof. Hence, a is not a semi-strict local minimum, which is absurd.
Therefore, g is decreasing on $]\frac{t^+}{a+\delta}, \frac{t^+}{a}[$, i.e., $g\left(\frac{t^+}{a+\delta}\right) = 1 + \left\lfloor \frac{t^+}{a+\delta} \right\rfloor - \frac{t^+}{a+\delta}$.

Finally, since $g\left(\frac{t^+}{a}\right) = g\left(\frac{t^-}{a}\right)$, we obtain :
$1 + \left\lfloor \frac{t^+}{a} \right\rfloor - \frac{t^+}{a} = \frac{t^-}{a} - \left\lfloor \frac{t^-}{a} \right\rfloor$,
hence : $t^- + t^+ = a\left(1 + \left\lfloor \frac{t^-}{a} \right\rfloor + \left\lfloor \frac{t^+}{a} \right\rfloor\right)$. We finally obtain the following necessary condition : $t^- + t^+ = ak, k \in \mathbb{N}^*$.
Therefore, by defining $m_T = \left\{ \frac{t_1+t_2}{k}, (t_1, t_2) \in T^2, k \in \mathbb{N}^* \right\}$, we have $a \in m_T$.

C. Conclusion about the correctness of Algorithm 1

Since ε_T is constant on a neighbourhood of $+\infty$, we know that the first semi-strict local minimum will be strictly contained in-between two semi-strict local maxima. Thanks to the previous necessary condition, to find the semi-strict local minimum (which exists since ε_T is a continuous function on $\mathbb{R}_+^*$) in-between two given successive semi-strict local maxima m_1 and m_2 of M_T, we only have to determine the value of $\underset{m \in m_T \cap]m_1, m_2[}{\text{argmin}} \varepsilon_T(m)$.
Finally Algorithm 1 is correct, and can actually run in $\mathcal{O}\left(|D|^2\left(1 + \frac{d^*}{\text{start}} - \frac{d_*}{\text{end}}\right)\right)$ with $d_* = \min T$ and $d^* = \max T$.

Proposition N.1 Let (b_n) be a correct transcription for a performance (t_n). This informally means that (b_n) is for instance the original score, or an equivalent transcription (with a different tempo or time signature). If the canonical tempo is within range $\left[\frac{1}{\text{end}}, \frac{1}{\text{start}}\right]$, and for all $n \in \mathbb{N}, \exists k \in \mathbb{N} : b_n = k\varepsilon$ then the complete graph whose verticies are all the local maxima contains a path which is equivalent to the canonical tempo, at least in terms of transcription.

D. Remarks about the fixpoints

Figure 26 presents the potential tempo curves, where the y-axis represents tempo, and is linear between ♩ = 40 (bottom) and ♩ = 240 (top). Since we only extend previously existing paths, without creating new, we see some path convergence, i.e., the merging of two paths into one. In this case, we end up with only 6 potential paths, whereas we started with over a thousand (exactly 1010). The x-axis corresponds to the index of each event in the performance, hence it does not contain any information regarding actual time. Such a representation allows for displaying results over a whole performance, instead of extracts. Furthermore, among the 6 paths present at the end, only 3 are fixpoints, hence can actually be a tempo curve.

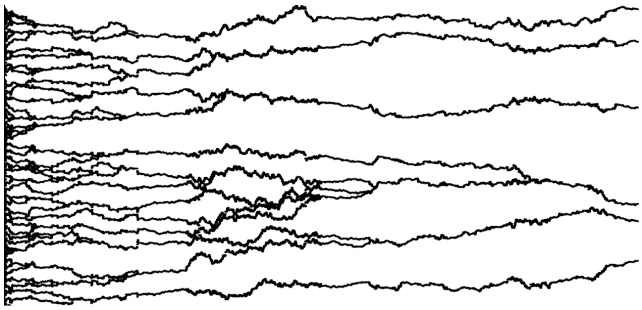

Figure 26: All potentials tempo curves found by our quantized approach for a performance of K. 331: III, W.A Mozart.

Definition N.2 Usually, a correct transcription does not indicates much tempo variations. Hence, if the considered section of a performance is played at a quasi-constant tempo, then the correct tempo curve is defined as the tempo curve $T = (t_1, ...t_N)$ in the graph that minimizes $\sum_{i=1}^{N-1} d(t_i, t_{i+1})$, and that is the nearest to the canonical tempo among all such tempo curves according to the tempo distance (or logarithmic distance) d. We can then extend this definition by merging different correct tempo curves corresponding to different sections of the performance under the hypothesis that an "actual" tempo curve is either stepwise quasi-constant, or slowly varying (and in this latter case, we will consider the tempo to be quasi-constant at the scale of two successive frames only).

Moreover, under the hypothesis that there exists a tatum t that has an actual meaning with respect to $F = F_1 \cup F_N, N \in \mathbb{N}$, we can extend the performance by duplicating it, as shown in Figure 28. In this situation, the fixpoints of the graph are all the lines (that are actually all possible paths) which end is also a starting point when the performance duplicates, hence the top one, but not the bottom one.

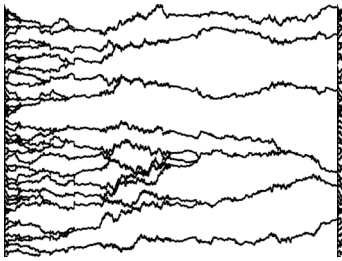

Figure 27: First duplication

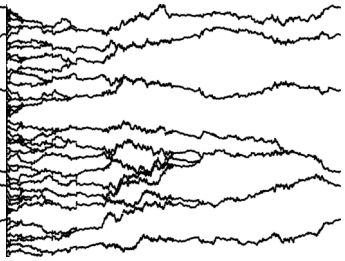

Figure 28: Second duplication

As the previous figures suggest, one can verify that Figure 26 only contains 3 fixpoints by following each line and checking that the end point of the first duplication is the same as for the second duplication.

The following proposition is actually a conjecture at the time being, for we did not have time to formally prove it.

Proposition N.3 If the tempo at times t_1 and t_N is the same, $N \in \mathbb{N}$, then the correct tempo curve is a fixpoint for the section $(t_1, ..., t_N)$ of a performance (t_n).

Corollary N.4 The correct tempo curve for a duplicated performance is the duplicated correct tempo curve for the initial performance.

Appendix A explains that reversing the performance allows for verifying a tempo curve, since any formal tempo is correct when reversed in time. One can verify that the notion of fixpoints is actually a more subtle way to discriminate tempo curves, and that all fixpoints are actually correct when reversed in time, since the computation minimises a distance, that is thus symetrical.

In the case of Figure 6 and Figure 7, the canonical tempo clearly indicates a rallentando at the end of the piece, hence the previous remarks about fixpoints cannot be applied to discriminate tempo curves.

Proposition N.5 For all partial performance $(t_n)_{n \in [\![0,N]\!]}, N \in \mathbb{N}$, there is a fixpoint among the potential paths given by Algorithm 2.

E. Remarks about the distances used in the quantized approach

In the definition of ε_T (10), we used an absolute distance, that allows for defining a distance to 0, in case of grace notes. However, when considering $t_a \in \mathbb{R}_+^*$ a tatum and T^* a canonical tempo for a given performance, we know the tempo corresponding to t_a is $\frac{1}{t_a}$ and $d\left(\frac{1}{t_a}, T^*\right) = d\left(\frac{1}{t_a}, \frac{\Delta b_n}{\Delta t_n}\right) = d(\Delta t_n, \Delta b_n t_a)$. Hence, if we consider $\underset{m \in \mathbb{Z}}{\text{argmin}}\, \text{dist}(\Delta t_n, m t_a)$ to be a possible transcription of Δt_n at tempo $\frac{1}{t_a}$ as we did in our approach, where dist is a distance, using d instead of dist would imply to minimize the distance between the canonical

tempo, and the actual tempo of the transcription, which may embody more musical meaning than the absolute distance. However, such a distance can consider a distance to 0, or in other words accepts grace notes, whereas a logarithmic one cannot, by definition. This implies that, if we were to use a logarithmic distance, then the grace notes would have to me marked down explicitly, at least at first processing.

Furthermore, we then minimize a norm over all elements of D, that was actually a max in (10), hence an infinity norm. Such a norm gives actual guarantees about the result, and may be simpler to understand and use, especially on theoretical proofs, but we could not find any evidence that point towards the use of this particular norm in our study.

F. A remark about the consistency property

Definition N.6 formal definition of the property

Definition N.7 locally consistent and inconsistent

Let p be a path in G locally inconsistent, i.e., such that there are $a_1, a_2 \in p$ so that $d \in D$ is quantized differently according to a_1 and a_2, with a_1 and a_2 local minima of successive frames. We therefore have two partial transcriptions of d being either : m_1 at tempo $\frac{1}{a_1}$ and m_2 at tempo $\frac{1}{a_2}$, both expressed in tatum unit, with $m_1 \neq m_2$. By hypothesis of our model, both a_1 and a_2 have an actual meaning within their respective frame F_1 and F_2. Let $A_i = \underset{k\in\mathbb{Z}}{\operatorname{argmin}}(d - ka_i), i \in \{1, 2\}$, $A = A_1 \cap A_2$ and t be the canonical tempo corresponding to d according to a correct transcription of the given performance. Thanks to Corollary III.2, we know that $|A_i| \geq 2$ iff a_i is a local maximum, which is absurd in our case, since both a_1 and a_2 are semi-strict local minima. Hence, $|A_i| = 1$. Moreover, by definition, $td \in A$, hence $A \neq \emptyset$, hence $A_1 = A_2$. Since $m_1 \in A_1, m_2 \in A_2$, both correspond to the same value, possibly express in different MTU. However, by definition, they both are expressed in tatum unit, since a_1 and a_2 embody a RTU value for the same tatum. Finally $m_1 = m_2$. this result is absurd and shall be investigated and updated !

G. Others formal proofs

Proof *(Proposition III.5)* Let $i \in [\![1, |F|]\!]$, since $t_i = t$, we have $f_i t \in \mathbb{N}$ MTU, when expressing symbolic values in tatum. Therefore, since $\min_{k\in\mathbb{Z}}\left(f_i - \frac{k}{t}\right) = \frac{1}{t}\min_{k\in\mathbb{Z}}\left(t\left(f_i - \frac{k}{t}\right)\right)$, we obtain : $\underset{k\in\mathbb{Z}}{\operatorname{argmin}}\left(f_i - \frac{k}{t}\right) = \underset{k\in\mathbb{Z}}{\operatorname{argmin}}\left(t\left(f_i - \frac{k}{t}\right)\right) = \underset{k\in\mathbb{Z}}{\operatorname{argmin}}(f_i t - k) = \{f_i t\}$ □

Appendix D : Glossary

A. Acronyms

***MIR* – Music Information Retrieval**: Interdisciplinary science aiming at retrieving information from music, in several ways. Amongst the various problems tackled by the community, one can notice transcription, automatic or semi-automatic musical analysis, and performance generation or classification... 1

***MTU* – Musical Time Unit**: Time unit for a symbolic, or musical notation, e.g., beat, quarter note (♩), eighth note (♪). 1, 2, 18

***RTU* – Real Time Unit**: Time unit to represent real events. Here, we usually use seconds as RTU. 1, 2, 18

***WLOG* – Without loss of generality**: The term is used to indicate the assumption that what follows is chosen arbitrarily, narrowing the premise to a particular case, but does not affect the validity of the proof in general. The other cases are sufficiently similar to the one presented that proving them follows by essentially the same logic. 15

B. Definitions

articulation: Describes how a specific note is played by the performer. For instance, *staccato* means the note shall not be maintained, and instead last only a few musical units, depending on the context. On the other hand, a fermata (*point d'orgue* in French) indicates that the note should stay longer than indicated, to the performer's discretion. 1

beat: Symbolic time unit of a score, its value is defined by a time signature. Although its value can change within a score, or through various transcription of a same piece, this notion is usually the most convenient way to describe a rhythmic sequence of events, since it is supposed to embody the *pulse* of the music felt by the listener. 1, 18

beat tracking: Common problem in the MIR community that consists in detecting the onsets of the theoretical beats from a performance, thus creating a partial note-alignment. 3

cadence: A cadence is can be defined as a progression of at least two chords which concludes a musical phrases. Actually, cadence is often used to refer to some parts of the punctuation within a musical section. 8, 18

chord: A chord is by definition the simultaneous production of at least three musical events with different pitches 2

measure: A measure is a symbolic time unit corresponding to a fixed amount (integer) of beats. This value is indicated by the time signature. 2

monophonic: Describes a piece played so that a single note can be heard at a time. A common hypothesis for monophonic pieces is to consider the end of a note as the beginning of the next one. The formalism presented in Section II.A is more fit for a monophic piece thant a polyphonic one. 2, 8, 10

online: In computer science, an online algorithm is one that can process its input piece-by-piece in a serial fashion, i.e., in the order that the input is fed to the algorithm, without having the entire input available from the start. 4

phrase: A musical phrase is defined similarly as a sentence in formal speech, usually depicting a single idea with clear punctuation. In this analogy, cadence act as a dots or comas within or in-between the phrases. 8

quantization: We consider here rhythm quantization, i.e., a way to find a rational number expressed in MTU from the real events durations in RTU, based on specific musical properties of time division. Indeed, in symbolic rhythmic notations of music, every single event can be expressed as a multiple of a certain unit called a tatum, usually expressed in beat. Then, the rhythm quantization consists in expressing each real event of a given performance, or rather its duration, as an integer. This integer is to be interpretated as the value of the duration, expressed in tatum converted to RTU. Hence, rythm quantization is equivalent to tempo inference, as explained in Section III.C 4

rallentando: Musical direction used to indicate a slackening in the pace. 16

rest: A symbolic notation for silence, following the same rules as actual note notations. 2

score – sheet music: Symbolic notation for music. The version considered here is supposed to fit a simplified version of the rhythmic Western notation system 1

tatum: Minimal resolution of a musical unit, expressed in beats. Although several values are possible, a tatum is usually indicates the following value for a given score (b_n) : $\sup\{r \mid \forall n \in \mathbb{N}, \exists k \in \mathbb{N} : b_n = kr, r \in \mathbb{R}_+^*\}$. For practical reasons, a tatum may be defined as smaller value than the one previously given, especially if this value is easier to express within the current time signature, or makes more sense musically.. 5, 18

tempo: Formally defined in Definition II.1 by $T_n^* = \frac{b_{n+1}-b_n}{t_{n+1}-t_n}$, tempo is a measure of the immediate speed of a performance, usually written on the score. It can be seen as a ratio between the symbolic speed indicated by the score, and the actual speed of a performance. Tempo is often expressed in beat per minute, or bpm 1

time signature: The time signature is a convention in Western music notation that specifies how many note values of a particular type are contained within each measure. It is composed of two integers : the amount of beat contained within a measure, and the value of these beats, indicated as division of a whole note, i.e., four quarter notes. 16, 17

timing – shifts: Delay between the theoretical real time onset according to the current tempo, and the actual onset heard in the performance. Even though such a delay is inevitable for neurological and biological reasons, those timings are usually overemphasized and understood as part of the musical expressivity of the performance 1, 5, 10

transcription: Process of converting an audio recording into symbolic notation, such as music score or MIDI file. This process involves several audio analysis tasks, which may include multi-pitch detection, duration or tempo estimation, instrument identification... 17

triplet: A triplet is a musical symbol indicating to play a third of the indicated duration. 5, 10

velocity: The velocity describes how loud a sound shall be played, or is actually played. 1