Tempo curves estimation, generation 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 or heard. Since tempo may not be constant at the scale of a piece, even locally, we need some kind of reference to compare to in order to define said speed. Indeed, a note-wise speed would not match the intuitive notion of tempo, based on a regular *pulse*. Such a reference can be found in Western symbolic notations of music, called either *music score* or *sheet music*, that allows for a definition of tempo as symbolic speed. We present here some results regarding the generation and analysis of local tempo curves of musical performances, involving methods that need to be given some symbolic information, and methods that generate them on the fly.**

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

I. Introduction

The Music Information Retrieval (MIR) community focuses on three representations of musical information, presented here from the least to the most formatted. The first one is raw audio, either recorded or generated, encoded using WAVE 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 hear) and duration, 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 abstract unit, called a beat, and indicates as well the pitch and articulation of those events. Those symbolic indications are then to be interpretated by a performer. Hence, a 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 in this report as timings, or shifts).

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 symbolic notation expressed in MTU into real time events expressed in RTU. We will present a formal definition for both tempo and performance in Section II.A.

However, tempo itself is insufficient to translate a sheet music into an musical performance, i.e., a sequence of real time events. Indeed, S. D. Peter, C. E. Cancino-Chacón, E. Karystinaios, and G. Widmer [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 litterature. Besides, although velocity don't help to meaningfully estimate tempo, the latter allows to marginally improve velocity-related predictions. Actually, velocity can be predicted relatively well when trained across recordings of the same piece, but fail dismally when trained across a specific musician's recordings of other pieces, implying that score features may trump individual style when modeling loudness choices [2]. Here, we shall focus mainly 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 (music) score.

Tempo and related works actually hold a prominent place in litterature. Tempo inference was first computed based on probabilistic models [3]–[5], and physical / neurological models [6], [7] as methods for real time score following from a performance ; and later the community tried neural network models [2] and hybrids approaches [8]. As most of the previous examples, we shall focus here on mathematically and/or musically explainable methods.

Since tempo needs a symbolic representation to be meaningful, one can consider transcription as a tempo-related work. We will save this discussion for Section III.D.

However, note-alignement, that is matching between each note of a performance with those indicated by a given score is a very useful preprocessing technique, especially for direct tempo estimation and further analysis, such as [9]–[11]. Two main methods are to be found in litterature : a dynamic programming algorithm, equivalent to finding a shortest path [12], that can works on raw audio ; and a Hidden Markov Model that needs more formatted data, such as MIDI files [13].

We then present the following contributions :

- a formal definition of tempo based on [3], [9] and [11] (III.A) ; and some immediate consequences (III.B)
- a score based revision of [6] and [7] for tempo inference (III.C)
- a general theoretical framework for scoreless tempo estimation with application to [14], [15] (IV.B)
- a new technique of tempo inference, without score, based on [15], and related new theoretical results (IV.D, IV.E and Appendix C)
- A method for data augmentation, and related theoretical results, based on [16] and [17] (Section IV.A)

This document, and associated algorithm implementations, can be found on the dedicated github repository [18].

II. Score-based approaches

A. Formal considerations

Since we chose to focus on MIDI files, we will represent a performance as a strictly increasing sequence of timepoints, or events, $(t_n)_{n\in\mathbb{N}}$, 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, calculated by [5] represents the limits of human ability to tell two rhythmic events apart, and is widely used within the field [9], [11], [14]–[17], [8], [10].

Likewise, a sheet music will be represented as a strictly increasing sequence of symbolic events $(b_n)_{n\in\mathbb{N}}$. In both of those definition, the terms of the sequence do not indicate the nature of the event (chord, single note, rest...). Moreover, in terms of units, (t_n) corresponds to RTU, whereas (b_n) corresponds to MTU.

With those definitions, let us formally define tempo :

Definition

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

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

Appendix A proves the following proposition :

Proposition

A tempo curve $T \in (\mathbb{R}_+^*)^{\mathbb{R}}$ is a formal tempo according to (t_n) and (b_n) iff $\forall n \in \mathbb{N}, \int_{t_n}^{t_{n+1}} T(t)\,\mathrm{d}t = b_{n+1} - b_n$

A tempo curve $T \in (\mathbb{R}_+^*)^{\mathbb{R}}$ is a formal tempo according to (t_n) and (b_n) iff $\forall n \in \mathbb{N}, \int_{t_n}^{t_{n+1}} T(t)\,\mathrm{d}t = b_{n+1} - b_n$

However, tempo is only tangible (or observable) between two events *a priori*. We then introduce the following :

Definition

Given (t_n) and (b_n), respectively a performance and a score, the canonical tempo is defined as a stepwise constant function $T \in (\mathbb{R}_+^*)^{\mathbb{R}}$ so that :

$$\forall x \in \mathbb{R}^+, \forall n \in \mathbb{N}, x \in [t_n, t_{n+1}[\Rightarrow T^*(x) = \frac{b_{n+1}-b_n}{t_{n+1}-t_n}$$

Given (t_n) and (b_n), respectively a performance and a score, the canonical tempo is defined as a stepwise constant function $T \in (\mathbb{R}_+^*)^{\mathbb{R}}$ so 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

Definition

Definition

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

$T \in (\mathbb{R}_+^*)^{\mathbb{R}}$ is said to be a formal tempo (curve) accordi
according to to (t_n) and (b_n) when, for all $n \in \mathbb{N}$, $\int_{t_0}^{t_n} T(t)\,\mathrm{d}t = b_n - b$

From now on, we will consider the convention : $t_0 = 0$ 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 litterature : K. Shibata, E. Nakamura, and K. Yoshii [8] and E. Nakamura, N. Ono, S. Sagayama, and K. Watanabe [4] take $\frac{1}{T^*}$ as definition ; C. Raphael [3], [9] et P. Hu and G. Widmer [11] choose similar definitions than the one given here (approximated at the scale of a measure or a section for instance).

T^* has the advantage to coincide with the tempo actually indicated on traditional sheet music (and therefore on MusicXML format), hence allowing a simpler and more direct interpretation of results.

B. Naive use of formalism

As said in introduction, the more formatted data, the less accessible it is ; and the field contains only a few datasets containing both sheet music and corresponding audio, more or less anotated with various labels [9]–[11], [16], [17].

In our study, we chose to rely on the (n)-ASAP dataset [17] that presents a vast amount of performances, with over 1000 different pieces of classical music, all note-aligned with the corresponding score. From there, we can easily compute our definition of tempo. Figure 1 presents the results for a specific piece of the (n)-ASAP dataset with a logarithmic y-scale, that contains a few brutal tempo change, whilst maintaining a rather stable tempo value in-between.

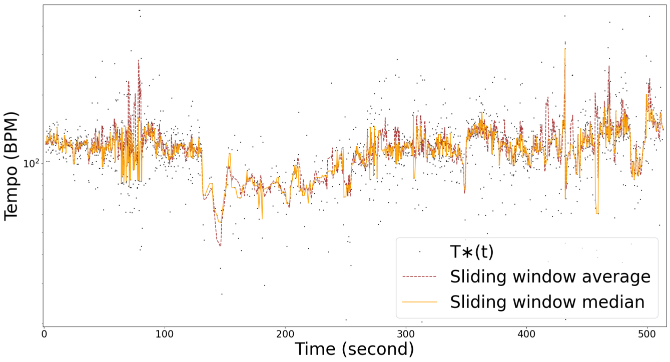

Figure 1: Tempo curve for a performance of Islamey, Op.18, M. Balakirev, with naive algorithm

In this graph, one can notice how T^* (plotted as little dots) appears 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 “feeled” tempo is quite constant for the listener, although the median line is a bit more stable than the average line, 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 as 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 notion of timing and tempo are mixed together in this computation, hence giving results that do not match the listener feeling of a stable tempo. Actually, timing can be seen as little modifications to the “official” score, and using the resulting score would allow for curves that fit better the listener feeling, though needing an actual transcription of the performance first.

C. *Physical models*

Among the tasks needing tempo estimation, the problem of real time estimation to allow a dedicated machine to play an accompagnement by following at least one real musician has been tackled by various approaches in litterature. C. Raphael [3] 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 [6]. In fact, their method has recently been developped to a commercial form[1], based on an a previous adaption of the latters by A. Cont [19].

The approach developped by E. W. Large and M. R. Jones [6] consider a simplified neurological model, where listening is a fundamentally active process, implying a synchronization between external events (those of the performance) and an internal oscillator, whose complexity depends of hypothesis on the shape of the first ones. The model consists of two equations for the internal parameters:

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

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

Here, (Φ_n) corresponds to the phase, or rather the phase shift between the oscillator and the external events, and (p_n) embodies its period. Finally, η_p and η_Φ are both constant parameters. This initial model is then modified to consider a notion of attending *via* the κ parameter, whose value change over time according to other equations. The new model contains the same formulas, with the following definition for F $F : \Phi, \kappa \to \frac{\exp(\kappa\cos(2\pi\Phi))}{\exp(\kappa)} \frac{\sin(2\pi\Phi)}{2\pi}$.

Even though this model shows pretty good results, has been validated through some experiments in [6], and is still used in the previously presented version (E. W. Large *et al.* [20]), a theoretical study of the system behavior remains quite complex, even in simplified theoretical cases [7], notably because of the function F expression.

In order to simplify the previous model, H.-H. Schulze, A. Cordes, and D. Vorberg [7] present *TimeKeeper*, that can be seen as a linearization of the previous approach, valid in the theoretical framwork 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, as shown by J. D. Loehr, E. W. Large, and C. Palmer [21]. Here, we used the derandomised version presented in [21], where $M_i = 0$ and $T_i = \tau_i$ for all $i \in \mathbb{N}$.

None of those models have an inherent comprehension of musical score information, since they both rely on a rather stable metronome. In the version displayed hereafter, they were modified to consider score information, in order to to create a more stable and precise value of tempo than the naive approach previously presented. Those modifications are detailled in the following paragraph (OR IN APPENDIX ?), and were made in order to keep consistency with the original models in their initial theoretical framework of validity. Let amin : $a, b \mapsto \begin{cases} a \text{ if } |a| < |b| \\ b \text{ otherwise} \end{cases}$

We first modify the E. W. Large and M. R. Jones [6] equations accordingly :

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

[1]https://metronautapp.com/

$$p_{n+1} = p_n \left(1 + \eta_p F(\Psi_n, \kappa_n)\right) \tag{4}$$

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

Using the analogy presented in [21], we then obtain the following equations for *TimeKeeper* :

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

$$\tau_{i+1} = \tau_i - \beta * \left(K_i \operatorname{mod}_{[-0.5,\, 0.5[}\, 1\right) \tag{6}$$

$$K_i = -\operatorname{amin}(k\tau + b_i - A_i, (k+1)\tau + b_i - A_i)$$
$$k = \left\lfloor \frac{A_i - b_i}{\tau_i} \right\rfloor \quad .$$

In both of those extensions, 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, i.e., $a_1 = \max_{n\in\mathbb{N}\,:\, b_n \leq B_i} \lfloor b_n \rfloor$ where B_i represents the internal value at time i acting as a beat unit (Φ_i for E. W. Large and M. R. Jones [6] and A_i for H.-H. Schulze, A. Cordes, and D. Vorberg [7]). The second argument embodies the correction according to the following beat, $a_2 = a_1 + 1$. One can notice that the phase is actually always used modulo 1 in E. W. Large and M. R. Jones [6], 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 here is equivalent to the original approach, i.e., (1), (2) $\Leftrightarrow$ (3), (4), hence justifying the designation “extension”.

Figure 2 displays the results of those two models, in regards with the canonical, or immediate tempo. One can notice that E. W. Large and M. R. Jones [6] model is less stable than *TimeKeeper*, although faster to converge.

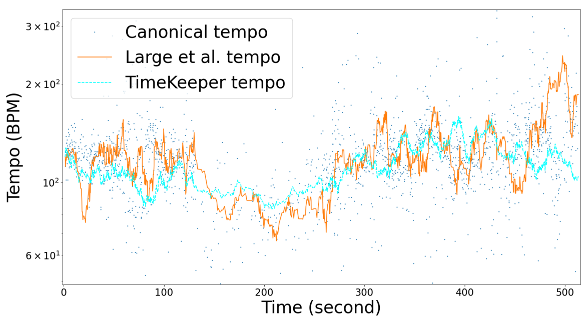

Figure 2: Tempo curve for a performance of Islamey, Op.18, M. Balakirev, according to various models

Figure 3 (below) exposes the visible difference in tempo initialization of the two models, starting both here with the initial tempo of 70 BPM (♩ = 70, i.e., the *beat* unit here is a quarter note). *TimeKeeper* does not manage to converge to any significant tempo. Such a behavior was to be expected, considering the theoretical framework for *TimeKeeper*, that is small tempo variation, and correct initialization. However, Large et al model manages to converge to a meaningful result. In fact, in the range 9 to 70 seconds, the estimated tempo according to Large is exactly half of the actual tempo hinted by the blue dots (canonical tempo).

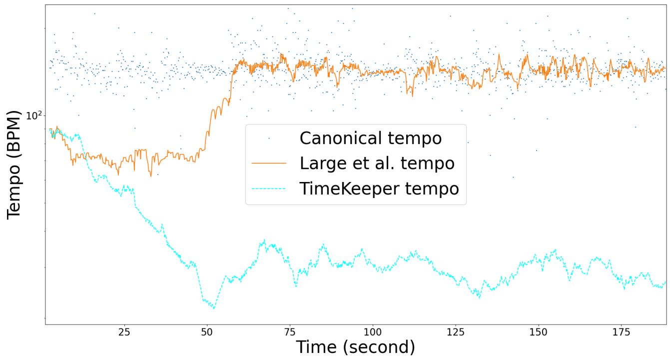

Figure 3: Tempo curve for a performance of Piano Sonata No. 11 in A Major (K. 331: III), W.A Mozart, according to the previous models with irrelevant initialization

III. Scoreless approaches

A. Motivations

There are three main issues with the previous models, apart from the prior knowledge of the sheet music, that are : salient sensibility to tempo initialization (cf. Figure 3), unstability that requires some time to (possibly) converge (cf. Figure 2 and Figure 3), and difficulty to accurately estimate relevant values of the constant internal parameters. According to our implementation, E. W. Large and M. R. Jones [6] is a particularly chaotic model regarding the latter.

We will present here two models focusing on tackling mainly the first two issues previously presented. Those rely on a specific musical property of division : in symbolic notations of music, every single event can be comprehended as a mutliple of a certain unit called a tatum, usually expressed in beat unit. Therefore, the real events of a performance, or rather their duration, can be interpreted as multiple of this tatum. However, considering a non-constant tempo, the real value (i.e., real duration in seconds) of this tatum may evolve through time, whereas the symbolic value remains constant anyway. Actually, detecting the tatum is equivalent to transcript the performance to sheet music, which a rather more complicated task than tempo estimation. For instance, there are several ways to write down sheet musics that are undistinguishable when performed. We call this ambiguity *tempo octaves* (cf Appendix A).

We considered here two quantization methods extracted from litterature : D. Murphy [14] and G. Romero-García, C. Guichaoua, and E. Chew [15]. Both of these papers present a theory of approximate division, that is a way to find a rational number, interpreted as the ratio of two symbolic events expressed in arbitrary unit (for instance in tatum) from the real events durations in seconds. Such a link is equivalent to defining a tempo with the formalism presented in III. Although [14] provides an algorithm to find candidate rationals, they do not include a way to compare those candidates, thus leaving no choice but an exhaustive approach (top-down in the paper). With another formalism, G. Romero-García, C. Guichaoua, and E. Chew [15] define a graph, restrained only to consistent values of tatum. We choose to adapt the latter, although their presentation is clearly user-oriented rather than automatic, since the introduced graph allows to mathematically (and therefore automatically) define a "good" and a "best" choice among all possible found tatum values.

B. *Introduction of an estimator based approach*

Given a sequence $(u_n)_{n\in\mathbb{N}}$, we now introduce the notation $(\Delta u_n)_{n\in\mathbb{N}} := (u_{n+1} - u_n)_{n\in\mathbb{N}}$ for the next sections.

The reason why the previous models have to converge is because they both try to find an exact value of tempo, and therefore sudden and huge tempo changes will lead to a uncertain period for the resulting tempo estimation, that is in this way the exact same problem as tempo initialization. When doing tempo estimation, we are in fact much more interested in a local tempo variation, relative to the previous estimation, rather than an absolute value, especially on a local time scale (where we can often assume tempo to be almost constant). Using the formalism presented in III, we first present the following result since $T^*_n > 0$:
$T^*_{n+1} = T^*_n \frac{T^*_{n+1}}{T^*_n} = T^*_n \frac{\Delta t_n}{\Delta t_{n+1}} \frac{\Delta b_{n+1}}{\Delta b_n}$

Let T_n be an estimation of T^*_n by a certain given model 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 $\frac{b_{n+2}-b_{n+1}}{b_{n+1}-b_n}$, where both the numerator and denumerator are translation permissive (i.e., we can afford a locally constant shift in both of our estimation), hence the resulting value is invariant by translation, or constant multiplication of our estimation. Furthermore, the value to be estimated only deals with symbolic units, meaning that we can use musical properties to find a consistent result. As a result, 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 of convergence, and that is robust to tempo initialization, while using only local methods (in other words, the resulting model is online). As noticed in the beginning of this section, symbolic value have some bindings that actually help to determine their values.

The point of this approach is to keep a constant factor between (T_n) and (T^*_n) in order to prevent the need for any convergence time. We will now define $T_{n+1} = \alpha_n T^*_{n+1}$.
We then find : $\frac{\Delta b_{n+1}}{\Delta b_n} = \frac{T^*_{n+1}\Delta t_{n+1}}{T^*_n \Delta t_n} = \frac{\frac{1}{\alpha_n}T_{n+1}}{\frac{1}{\alpha_n}T_n} \times \frac{\Delta t_{n+1}}{\Delta t_n}$,
hence $\frac{\Delta b_{n+1}}{\Delta b_n} = \frac{T_{n+1}}{T_n} \times \frac{\Delta t_{n+1}}{\Delta t_n}$.
Let us then write the actual formula of the model :

$$\frac{T_{n+1}}{T_n} = \frac{T^*_{n+1}}{T^*_n} = \frac{\Delta t_n}{\Delta t_{n+1}} \underbrace{E\left(\frac{T_{n+1}}{T_n} \times \frac{\Delta t_{n+1}}{\Delta t_n}\right)}_{\frac{\Delta b_{n+1}}{\Delta b_n}} \tag{7}$$

where E is a function-like object (closer to a *object* in computer science than an actual mathematical function), designated by *estimator*. This function 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 (7), therefore supposed to rarely match the theoretical values.

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 represents $\frac{T_{n+1}}{T_n}$ in (7) :

$$T_{n+1} = T_n \underset{x\in\left[\frac{\sqrt{2}}{2}T_n, \sqrt{2}T_n\right]}{\operatorname{argmin}} d\left(x, \frac{\Delta t_n}{\Delta t_{n+1}} E\left(x \frac{\Delta t_{n+1}}{\Delta t_n}\right)\right) \tag{8}$$

where $d : a, b \mapsto k_* |\log(\frac{a}{b})|, k_* \in \mathbb{R}^*_+$, is a logarithmic distance, choosen since an absolute distance would have favor small values by triangle inequality in the following process.

In the implementation presented here, the estimator role is more to quantize the ratio in order to output a musically relevant value. In our test, we limited these quantizations to accept only regular division (i.e., powers of 2). Furthermore, the numerical resolution for the previous equation was done by a logarithmically evenly spaced search and favor 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 going from the initial tempo. Appendix B gives further explanation about (8).

C. *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 $\left(\alpha_n := \frac{T_n}{T^*_n}\right)_{n\in[\![1,N]\!]}$

and $\left(\tilde{\alpha}_n := \exp\left(\ln\left(\frac{T_n}{T_n^*}\right) - \left\lfloor \log_2\left(\frac{T_n}{T_n^*}\right)\right\rfloor \ln(2)\right)\right)_{n \in [\![1,N]\!]}$. $(\tilde{\alpha}_n)_{n\in[\![1,N]\!]}$ is then called the *normalized* sequence of ratio, where each value is uniquely determined within the range $[1, 2[$. Such a choice allows for merging together the tempo octaves, as explained in Appendix A. In this representation, we actually have $\tilde{2} = \tilde{1}$, and adding $\tilde{1}$ is equivalent to multiply the initial value by 2. We will now define a *spectrum* $S = (\tilde{\alpha}_n)_{n\in[\![1,N]\!]}$. We then call $|S|$ the value $N \in \mathbb{N}$ and $\mathcal{C}$ the range $[1, 2[$ seen as a circle according to the following application : $\begin{array}{l} c : [1,2[\to \mathcal{C}(0,1) \\ x \quad\quad \mapsto (\cos(2\pi x), \sin(2\pi x)) \end{array}$, so that $c(1) = c(2)$

We then define the *measure* of a spectrum S, that embodies a standard deviation on $\mathcal{C}$, so that :

$$m(S, \Delta) = \max_{\tilde{d}\in\mathcal{C}} \frac{\left|\left\{n \in [\![1, |S|]\!] : d\left(\tilde{\alpha}_n, \tilde{d}\right) \le \Delta\right\}\right|}{|S|} \quad (9)$$

Where $\Delta \in \left[0, \frac{1}{2}\right]$ embodies the measure accuracy, and d is still a logarithmic distance, slightly modified on $\mathcal{C}$ to be consistent with $d\left(\tilde{1}, \tilde{2}\right) = 0$. Actually, it can be shown that on $\mathcal{C}$, $d : \tilde{a}, \tilde{b} \mapsto \min\left(\left|\tilde{a} - \tilde{b}\right|, 1 - \left|\tilde{a} - \tilde{b}\right|\right)$.

The reader can verify that this measure is invariant with respect to spectrum rotation by any $\lambda \in R_+^*$ (i.e., $m(S, \Delta) = m\left(\left(\widetilde{\lambda\alpha_n}\right)_{n\in[\![1,N]\!]}, \Delta\right)$), and does not depend on the normalisation interval (here $[1, 2[$, but actually $[\lambda, 2\lambda[$ would work just as well, with a different expression of d on $\mathcal{C}$). Finally, $0 \le m(S, \Delta) \le 1$, and $m(S, \Delta) = 0 \Leftrightarrow |S| = 0$, $m(S, \Delta) = 1$ iff S only contains values within a 2Δ range.

This measure allows to quantify the quality of this model, without considering tempo octaves, or equivalently to quantify the quality of the estimator. The C++ implementation of the measure used to obtain the following figures if available on [18].

résultats sur (n-)Asap selon les périodes, les compositeurs, etc..., éventuellement en annexe ?

D. *Towards a quantized approach*

The previous model supposes the existence of an oracle, more or less correct in its predictions, that is actually a (partial) transcriber. In this section, we will focus on this transcribing part, by extending G. Romero-García, C. Guichaoua, and E. Chew [15] model with the previous formalism. In fact, we extend the previous approach by considering the estimator as our central model and then extracting tempo values rather than the opposite.

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

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

This function is called the *transcription error*, and can be interpretated as maximum error (in real time unit) between all real events $d \in D$ and theoretical real duration of 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 local maxima of ε_D, except those that also are minima, 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

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

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

With this property, we can then choose to consider only local minima of ε_D as in [15], since there is exactly one local minima in-between two such successive 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 maxima (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 ≠ ∅, start, end) :
M ← { d/(k+1/2), d ∈ D, k ∈ ⟦d/end, d/start⟧ }
P ← { (d1+d2)/k, (d1,d2) ∈ D², k ∈ ⟦(d1+d2)/end, (d1+d2)/start⟧ }
localMinima ← ∅
for (m1, m2) ∈ M² two successive maxima :
    in_range ← {p ∈ P : m1 ≤ p < m2}
    localMinima ← localMinima ∪ {min(in_range)}
return localMinima
```

Algorithm 1: Returns all local minima within [start, end]

G. Romero-García, C. Guichaoua, and E. Chew [15] then defined $G = (V, E)$ a graph whose vertices are the local minima of ε_D with D a sliding window, or *frame*, on a given performance, and whose edges are so that they can guarantee a *consistency property*, explained hereafter.

The *consistency property* for two tatums a_1, a_2 specifies that, if $F_\cap$ is the set of all values in common between two successive frame, for all $d \in F_\cap$, d is quantized the same way according to the tatum 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 the same given tatum at some tempo (respectively $\frac{1}{a_1}$ and $\frac{1}{a_2}$ as shown in Section III.E). 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, but since an exact tempo curve would be (T_n^*), those two problems are actually 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 minima a_1 of ε_{F_1}. According to (11), we can divide the set of all local maxima in two, with those "caused" by $F_\cap$, $M_{F_\cap}$, and the others. Let then $m_1, m_2 \in M_{F_\cap}$. Those 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 minima within the range $]m_1, m_2[$ for both of these functions. However, thanks

Proposition

Proposition

Let m_1, m_2 be two successive local maxima of ε_D, $a_1 \in]m_1, m_2[$, $a_2 \in [m_1, m_2]$, $d \in D$ and $m \in \mathbb{Z}$.

Then $m \in \underset{k\in\mathbb{Z}}{\operatorname{argmin}}|d - ka_1| \Rightarrow m \in \underset{k\in\mathbb{Z}}{\operatorname{argmin}}|d - ka_2|$.

Let m_1, m_2 be two successive local maxima of ε_D, $a_1 \in]m_1, m_2[$, $a_2 \in [m_1, m_2]$, $d \in D$ and $m \in \mathbb{Z}$.

Then $m \in \underset{k\in\mathbb{Z}}{\operatorname{argmin}}|d - ka_1| \Rightarrow m \in \underset{k\in\mathbb{Z}}{\operatorname{argmin}}|d - ka_2|$.

to we know that both these local minima will quantize the elements of D the same way. Hence, by defining :

- $m_1 = \max\{m \in M_{F_\cap} : m < a_1\}$
- $m_2 = \min\{m \in M_{F_\cap} : m > a_1\}$
- a_2 a local minima of ε_{F_2} in the range $]m_1, m_2[$, which exists since m_1 and m_2 are local maxima (that are not local minima).

We obtain : (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, even if it means considering a bigger interval, for any given performance, and any given frame length for defining G.

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, even if it means considering a bigger interval, for any given performance, and any given frame length for defining G.

However, this restriction on G appears to have some interest. Indeed, let p a path in G locally inconsistent, i.e., such that $a_1, a_2 \in p$ 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 a two partial transcriptions of d being either : m_1 at tempo $\frac{1}{a_1}$ and m_2 at tempo $\frac{1}{a_2}$, m_1, m_2 expressed in tatum unit, with $m_1 \neq m_2$. WHAT IS THE POINT ?

E. *Quantization revised*

Let us define from now our tatum $\varepsilon = \frac{1}{60}$♩, which correspond to an sixteenth note wrapped within a triplet within a quintuplet, and has the property that $1\ \varepsilon/s = 1$ ♩$/m$.

With our tatum defined, we can now choose to express all our symbolic units as multiple of this tatum, hence the unit for symbolic values is now ε. 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*, since it embodies the error between a transcription of $d \in D$ as m expressed in tatum, hence 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 the definition of canonical tempo.

- LR
- bidi (2 passes: LR + RL) : justification (en annexe) : retour à la définition formelle de Tempo : valide dans les deux sens, d'où la possibilité de le faire en bidirectionnel + parler rapidement d'une application à Large
- RT : avec valeur initiale de tempo

,

F. *résultats évaluation (comparaison avec 3)*

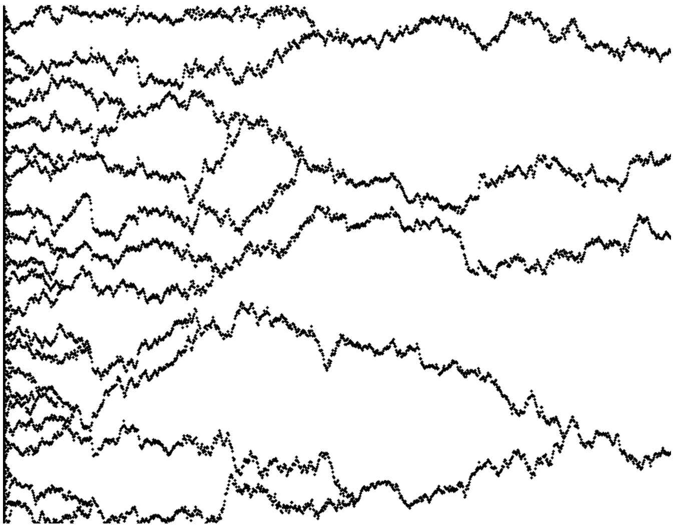

Figure 4: All potentials tempo curves found by a quantized approach for a performance of Piano Sonata No. 11 in A Major (K. 331: III), W.A Mozart. The tempo scale is linear between ♩ = 40 (bottom) and ♩ = 240 (top)

IV. Applications

A. Data extension

Let $(t_n)_{n\in\mathbb{N}}, (T^*_n)_{n(\in\mathbb{N})}, (T_n)_{n\in\mathbb{N}}$ be respectively a performance, a canonical tempo for this performance, an estimated tempo curve (supposed to be flattened with respect to the canonical tempo) and $T_c \in R^*_+$ a given tempo value. We define $(s_n := \Delta t_n (T_n - T^*_n))_{n\in\mathbb{N}}$ the symbolic shift of the performance according to (T^*_n) and (T_n). The *normalized* performance to tempo T_c of (t_n) is then defined as :

$$\hat{t}_0 = t_0, \forall n \in \mathbb{N}, \hat{t}_{n+1} = \hat{t}_n + \underbrace{\Delta t_n \frac{T^*_n}{T_c}}_{\alpha_n} + \underbrace{\frac{s_n}{T_c}}_{\beta_n} \qquad (12)$$

Where 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

$\left(\hat{t}_n\right)_{n\in\mathbb{N}}$ **is a performance as defined in Section II.A, and** $\left(\hat{T}^*_n\right)_{n\in\mathbb{N}} = \left(\frac{1}{1+\frac{s_n}{\Delta b_n}} T_c\right)_{n\in\mathbb{N}}$

$\left(\hat{t}_n\right)_{n\in\mathbb{N}}$ is a performance as defined in Section II.A, and $\left(\hat{T}^*_n\right)_{n\in\mathbb{N}} = \left(\frac{1}{1+\frac{s_n}{\Delta b_n}} T_c\right)_{n\in\mathbb{N}}$

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, $\hat{T}^*_n = \frac{\Delta b_n}{\Delta \hat{t}_n} = \frac{\Delta b_n}{\Delta \frac{b_n}{T_c} + \frac{s_n}{T_c}} = \frac{1}{1+\frac{s_n}{\Delta b_n}} T_c$ $\square$

Depending on the use of this data, one can choose to adapt the definition os (s_n), for instance by normalizing its values or cutting all shifts that represent over a certain portion of their duration. Proposition

Proposition

$\left(\hat{t}_n\right)_{n\in\mathbb{N}}$ **is a performance as defined in Section II.A, and** $\left(\hat{T}^*_n\right)_{n\in\mathbb{N}} = \left(\frac{1}{1+\frac{s_n}{\Delta b_n}} T_c\right)_{n\in\mathbb{N}}$

$\left(\hat{t}_n\right)_{n\in\mathbb{N}}$ is a performance as defined in Section II.A, and $\left(\hat{T}^*_n\right)_{n\in\mathbb{N}} = \left(\frac{1}{1+\frac{s_n}{\Delta b_n}} T_c\right)_{n\in\mathbb{N}}$

exposes how such definitions of (s_n) allow for guarantee on tempo change and deviation of the canonical tempo for the *normalized* performance with respect to T_c.

- génération de données "performance" : pour data augmentation ou test robustesse (fuzz testing) aplanissement de tempo démo MIDI?
- transcription MIDI par parsing : pre-processing d'évaluation tempo (approche partie 4)
- analyse "musicologique" quantitative de performances humaines de réf. (à la Mazurka BL) données quantitives de tempo et time-shifts

V. Conclusion & perspectives

- intégration pour couplage avec transcription par parsing (+ plus court chemin multi-critère)
- Section III.E presented a model which appear to share some similarities with E. W. Large and M. R. Jones [6] as a score-based approach. Therefore, studying the formalism for a quantizer might allow to obtain some theoretical results for the previous model, regarding for instance convergence guarantee and meaningfulness of the result.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[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, "ANTESCOFO: Anticipatory Synchronization and Control of Interactive Parameters in Computer Music.," in *International Computer Music Conference (ICMC)*, 2008, pp. 33–40.

[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.

Appendix A : Some formal considerations

A. Equivalence of tempo formal definitions

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.

B. The tempo octave problem

When estimating 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 $t = \left(\frac{b_n}{2}\right)_{n\in\mathbb{N}}$.
Then, the canonical tempo according to t, called (T_1^*), and the one according to (b_n), called (T_2^*) verify :
$\forall n \in \mathbb{N}, T_{1,n}^* = \frac{\frac{b_{n+1}}{2} - \frac{b_n}{2}}{t_{n+1} - t_n} = \frac{1}{2} T_{2,n}^*$. Actually, the t transcription corrections 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 those two transcription is better than the other. This problem is here known as the tempo octave problem, and should be keeped in mind when transcribing or estimating tempo. We present in Section III.B a model resistant to these tempo octaves, as well as other kind of octaves not discussed here (for instance multiplying the tempo by 3 by using triplet).

C. Tempo convservation when reversing time

First, we want to insist on the fact that none of the sequence (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, called $|(b_n)|$ and $|(t_n)|$ respectively, with $|(b_n)| = |(t_n)|$. Let us then define the reversed sequence of $(u_n)_{n\in\mathbb{N}}$ as $r\left((u_n)_{n\in\mathbb{N}}\right) := \left(\overline{u_n} = u_{|(u_n)|} - u_{|(u_n)|-n}\right)_{n\in[\![0,\,|(u_n)|]\!]}$. Both $\overline{b} = r\left((b_n)_{n\in\mathbb{N}}\right)$ and $\overline{t} = r\left((t_n)_{n\in\mathbb{N}}\right)$ are correct representations of a sheet music and performance respectively, as defined in Section I-I.A.

Let $t^* = t_{|(t_n)|}$ and $q = |(t_n)|$, T a formal tempo according to (t_n) and (b_n), i.e., $\forall n \in \mathbb{N}, \int_{t_n}^{t_{n+1}} T(t)\,\mathrm{d}t = b_{n+1} - b_n$, $n \in [\![0, q-1]\!]$, 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 according to $(\overline{t_n})$ and $\left(\overline{b_n}\right)$.

D. Musical explication 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.

Appendix B : Estimator Model

A. Formal explanations and proofs

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

Then, (8) 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 $k_* \in R_+^*$ 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 $x \in R_+^*$. 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 :

$$d(T_{n+1}, T_n) = k_* |\log(y)| = d(1, y) \tag{13}$$

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 $d(1, x) = d(1, 2x)$.

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, the latter being incorrect with regards to usual transcription convention, and their corresponding tempo curves.

One can notice that these are quite similar, and in fact, a human being could not tell them apart, as shown by Figure 8.

Figure 5: Transcription A

Figure 6: Transcription B

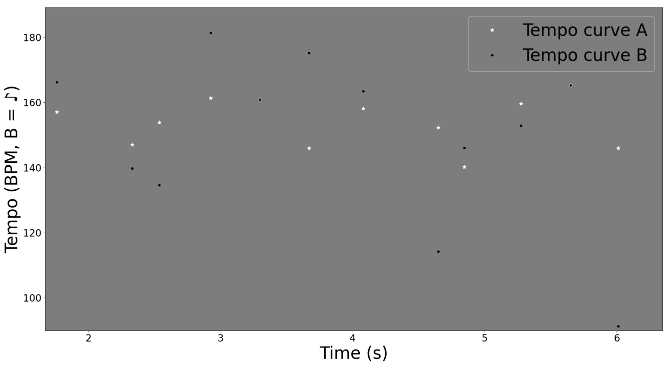

Figure 7: Tempo curves A & B

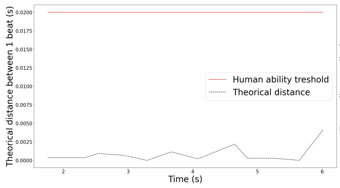

Figure 8: Tempo distance (s)

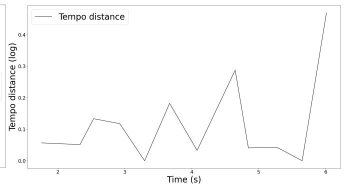

Figure 9: Tempo distance (log)

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 suppose impossible for a humain being according to the value of ε defined (and displayed as the top line in Figure 8) in Section II.A.

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 :

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

hence we have proved ε_T to be continous on R_+^*.

Furthermore, for $n \in \mathbb{N}^*, T \subset \left(\mathbb{R}_+^*\right)^n, a \in R_+^*$,
$\varepsilon_T(a) = a \max_{t \in T} g\left(\frac{t}{a}\right) = a \times 1 \max_{t \in T} g\left(\frac{t}{a} 1^{-1}\right) = a \varepsilon_{T/a}(1)$.
Hence the intuitive following result : the smaller the tatum, the smaller the bound of the error.

A. *Characterization of local maxima that happen not be a local minima*

1) *First implication:*
Let a be a local maxima (and not a local minima) 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 $t \in \underset{t' \in T}{\operatorname{argmax}}\, g\left(\frac{t'}{a}\right)$, hence $\varepsilon_T(a) = a g\left(\frac{t}{a}\right)$.
For all $\delta \in]-\varepsilon, \varepsilon[, \varepsilon_T(a) = a g\left(\frac{t}{a}\right) \geq \varepsilon_T(a + \delta)$,
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 $a g\left(\frac{t}{a}\right) \geq (a + \delta) g\left(\frac{t}{a + \delta}\right) \geq a g\left(\frac{t}{a + \delta}\right)$
Hence, $g\left(\frac{t}{a}\right) \geq g\left(\frac{t}{a + \delta}\right)$ since $a > 0$, for all $\delta \in [0, \varepsilon[$.
Therefore, g <u>increases monotonically</u> in the range $]\frac{t}{a + \delta}, \frac{t}{a}[$, since g has a unique local maxima (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 $g\left(\frac{t}{a}\right) = \frac{t}{a} - \left\lfloor \frac{t}{a} \right\rfloor, \varepsilon_T(a) = t - a\left\lfloor \frac{t}{a} \right\rfloor$.

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

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

Hence $A \subset \left\{\frac{t_1 \mp t_2}{n}, (t_1, t_2) \in T^2, n \in \mathbb{Z}^*\right\}$, because $T \subset (R_+^*)^{|T|}$. Therefore, there is a countable set of closed convex intervals, whose union is R_+^* so that on each of these intervals, ε_T is equal to $f_t : a \mapsto ag\left(\frac{t}{a}\right)$ 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, $t \in \underset{t' \in T}{\operatorname{argmax}}\, g\left(\frac{t'}{a}\right)$. 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 $N\left(\frac{t}{a}\right)^+$ in the following, the previous expression of g remains valid, i.e., $\forall x \in N\left(\frac{t}{a}\right)^+, g(x) = x - \lfloor x \rfloor$. Moreover, $x \mapsto \lfloor x \rfloor$ is right-continuous, hence by restricting $N\left(\frac{t}{a}\right)^+$, we can assure for all $x \in N\left(\frac{t}{a}\right)^+, \lfloor x \rfloor = \left\lfloor \frac{t}{a} \right\rfloor$. Let then $y = a - \frac{t}{x}$ so that $x = \frac{t}{a-y}$, we then have $\varepsilon_T(a-y) = t - (a-y)\left\lfloor \frac{t}{a} \right\rfloor \leq \varepsilon_T(a) = t - a\left\lfloor \frac{t}{a} \right\rfloor$ because a is a local maxima of ε_T and $a - y$ is within a (left) neighbourhood of a, even if it means restricting δ' or $N\left(\frac{t}{a}\right)^+$. Hence, $t - \left\lfloor \frac{t}{a} \right\rfloor a \geq t - (a-y)\left\lfloor \frac{t}{a} \right\rfloor$ i.e., $a\left\lfloor \frac{t}{a} \right\rfloor \leq (a-y)\left\lfloor \frac{t}{a} \right\rfloor \Leftrightarrow 0 \leq -y\left\lfloor \frac{t}{a} \right\rfloor$ i.e., $\left\lfloor \frac{t}{a} \right\rfloor \leq 0$ i.e., $\left\lfloor \frac{t}{a} \right\rfloor = 0$, since y, t and a are all positive values. Then, $a > t$ and therefore $\varepsilon_T(a-y) = t = \varepsilon_T(a)$ for $\left\lfloor \frac{t}{a} \right\rfloor = 0$. 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 $x \mapsto \max\left(\underset{t' \in T \setminus \{t\}}{\max} xg\left(\frac{t'}{x}\right), \varepsilon_T(a)\right)$ on $[a, +\infty[$. Hence a is a local minima on the right, and since ε_T is constant on a left neighbourhood of a, a is also a local minima on the left. Finally, a is a local minima, which is absurd by definition.
- else, g is decreasing on $N\left(\frac{t}{a}\right)^+$, $\frac{t}{a}$ is by definition a local maxima of g. However, g only has a unique local maxima 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 $\underline{a = \frac{t}{k + \frac{1}{2}}, k \in \mathbb{N}, \text{ since } a > 0}$.

2) *Second implication:*

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

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

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

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., $\left|\frac{t_1}{a} - \frac{t_2}{a}\right| = k_a \in \mathbb{N}$.

Then, $\frac{t_1}{x^*} = \pm \frac{t_2}{x^*} \bmod 1$ i.e., $\left|\frac{t_1}{x^*} \mp \frac{t_2}{x^*}\right| = k_* \in \mathbb{N}$, 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 <u>a finite set</u>, 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 $a_2 = \underset{(t_1, t_2) \in T^{*2}}{\min} \underbrace{x_{t_1, t_2}}_{> a}$ and $a_1 \in]a, a_2[$,

let $t^* = \underset{t' \in T^*}{\operatorname{argmax}}\, g\left(\frac{t'}{a_1}\right)$ 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 maxima 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, $\underline{\varepsilon_T(a) > \varepsilon_T(x)}$.
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 local maxima of ε_T and not a local minima, and then the set of all such local maxima of ε_T is $M_T = \left\{ \frac{t}{k+\frac{1}{2}}, t \in T, k \in \mathbb{N} \right\}$ and therefore with the notations introduced in Section III.D, we proved (11)

■

B. Necessary condition to be a local minima

Let a be a local minima of ε_T, i.e.,

ε_T is constant on a neighbourhood of $+\infty$, donc on va d'abord trouver un maximum local qui n'est pas un minimum local, puis le reste de la preuve fonctionne

Par continuité de ε_T, on est assuré de l'existence d'exactement un unique minimum local entre deux maximums locaux, qui est alors global sur cet intervalle.
Par la condition nécessaire précédente, il suffit donc, pour déterminer ce minimum local, de déterminer le plus petit élément parmi les points obtenus, contenus dans l'intervalle. On en déduit ainsi un algorithme en $\mathcal{O}\left(|T^2| \frac{t^*}{\tau} \log\left(|T| \frac{t^*}{\tau}\right)\right)$ permettant de déterminer tous les minimums locaux accordés par le seuil τ fixé, sur l'intervalle $]2\tau, t_* + \tau[$

On en déduit la correction de Algorithm 1.

TODO : send report to Gonzalo, (Rigaux, Lemaitre)

Appendix D : Glossary

A. Acronyms

***MIR* – Music Information Retrieval**: Interdisciplinary science aiming at retrieving information from music, in several ways. Amoungst 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

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

***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. 12

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, 14

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

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. 5

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.. 4, 6

Definition

Definition

$T \in \left(\mathbb{R}_+^*\right)^{\mathbb{R}}$ **is said to be a formal te**
according to (t_n) **and** (b_n) **when,**
$\int_{t_0}^{t_n} T(t)\,\mathrm{d}t = b_n - b_0$

tempo: Formally defined in $T \in \left(\mathbb{R}_+^*\right)^{\mathbb{R}}$ is said to be a formal tempo (to (t_n) and (b_n) when, for all $n \in \mathbb{N}$, $\int_{t_0}^{t_n}$ 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. 13

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

transcription: Process of converting an audio recording into symbolic notation, such as sheet music or MIDI file. This process involves several audio analysis tasks, which may include multi-pitch detection, duration or tempo estimation, instrument identification… 1, 13

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

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