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. 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.**

***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 music score features may trump individual style when modeling loudness choices [2].

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 synchronization with a performance ; and later the community tried neural network models [2] and hybrids approaches [8]. A very useful preprocessing technique for tempo inference and further analysis, such as [9]–[11], is note-alignement, that is a matching between each note of a performance and those indicated by a given score. 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]. As most of the previous examples, we shall focus here on mathematically and/or musically explainable methods.

terminologie: choisir entre
- music score
- sheet music

shall present below
We thus introduce the following contributions :

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

This document, and associated algorithm implementations, & results can be found on the dedicated github repository[1] [18].

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 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 [5], and is widely used within the field [8]–[11], [14]–[17]. Likewise, a sheet music will be represented as a strictly increasing sequence of symbolic events $(b_n) \in \mathbb{R}^{\mathbb{N}}$.
Please note that, in both of those definitions, the terms of the sequence do not indicate the nature of the corresponding 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 II.1 $T \in (\mathbb{R}_+^*)^{\mathbb{R}}$ is said to be a formal tempo (curve) ~~according, or~~ 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 according to (t_n) and (b_n) if and only if $\forall n \in \mathbb{N}, \int_{t_n}^{t_{n+1}} T(t)\,\mathrm{d}t = b_{n+1} - b_n$

Proof See Appendix A. □

Since tempo is only ~~tangible (or~~ observable) between two events given *a priori*, we introduce the notion of canonical tempo, also called immediate tempo T^*.

[1] https://github.com/sylvain-meunier/stageL3

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 according to Definition II.1. From now on, we will consider the convention : $t_0 = 0$ RTU et $b_0 = 0$ MTU. assume by conv. that

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 : [3], [9] and [11] choose definitions very similar to T^*, approximated at the scale of a measure or a section for instance, whereas [4] and [8] use $\frac{1}{T^*}$. When a performance perfectly fit 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 a simpler and more direct interpretation of results.

B. Computation of the canonical tempo

There exists
~~The field contains only~~ a few datasets containing note-alignment matching between both sheet music and corresponding audio, more or less anotated with various labels [9]–[11], [16], [17]. For our study, we chose to rely on the (n)-ASAP dataset [17] 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 ~~contains~~ a few (2) abrupt tempo changes, whilst maintaining a rather stable tempo value in-between.

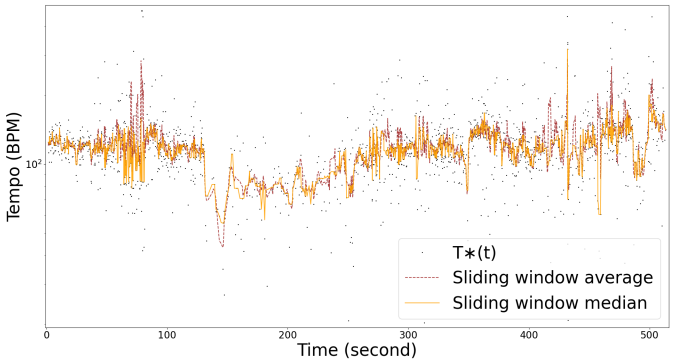


Figure 1: Graph of T^* for a performance of Islamey, Op.18, M. Balakirev, extracted from the (n)-ASAP dataset.

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 "felt" perceived tempo is quite constant for the listener, ~~although~~ 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 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 notions 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 expressive 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. Two physical models for tempo estimation

Among the tasks requiring tempo estimation, 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 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, a combination of these methods has recently been developped to a commercial form[2], based on an a previous work by A. Cont [19].

The approach ~~developped by~~ [6] 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 hypothesis on the shape of *observations*. The model consists of two equations for the internal parameters presented hereafter for all $n \in \mathbb{N}$:

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

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

Here, Φ_n corresponds to the phase, or rather the phase shift at each event t_n between the oscillator and the external events, and p_n embodies its period. Finally, $\eta_p \in \mathbb{R}^+$ and $\eta_\Phi \in \mathbb{R}^+$ are both constant damping 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 not showed here. Finally $F : \Phi, \kappa \mapsto \frac{\exp(\kappa\cos(2\pi\Phi))}{\exp(\kappa)} \frac{\sin(2\pi\Phi)}{2\pi}$ is the correction at t_n to match *expectations* and *observations*. This model being fit for ~~score following, or more precisely~~ 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.

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

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

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

where

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

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, i.e., $a_1 = \max_{n \in \mathbb{N} \,:\, b_n \le B_i} \lfloor b_n \rfloor$ where B_i represents the internal value at time i acting as a beat unit (Φ_i in our case). 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 [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". We will from now on refer to this model as *Large et al.* since the modifications presented here were inspired by various works within litterature, including [19].

Even though this model has been validated experimentally in [6], 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 [7]. H.-H. Schulze, A. Cordes, and D. Vorberg [7] thus present their *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]. Here, we used the derandomised version presented ~~and considered~~ by J. D. Loehr, E. W. Large, and C. Palmer [21]. Using their analogy, we then obtain the following equations for *TimeKeeper* :

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

$$\tau_{i+1} = \tau_i - \beta * \left(K_i \bmod_{[-0.5,\, 0.5[} 1\right) \quad (6)$$

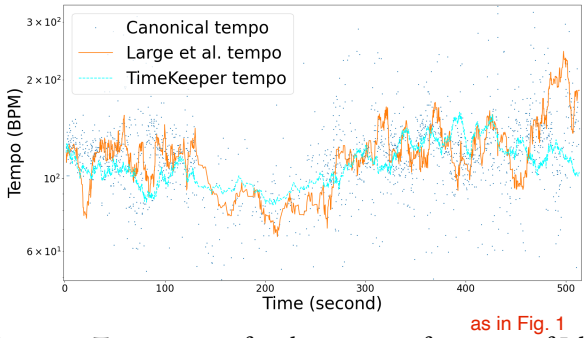
$$K_i = -\operatorname{amin}(k\tau + b_i - A_i, (k+1)\tau + b_i - A_i)$$
$$k = \left\lfloor \frac{A_i - b_i}{\tau_i} \right\rfloor$$

[2] https://metronautapp.com/

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). Finally, (b_i) and (t_i) are the formal representation of the score and performance respectively. Figure 2 displays the results of those two models, in regards with the canonical, or immediate tempo. One can notice that the modified E. W. Large and M. R. Jones [6] model is less stable than *TimeKeeper*, although faster to converge.

deja dit

compared to



as in Fig. 1

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

illustrates

Figure 3 exposes 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).

(*)

original or modified ?

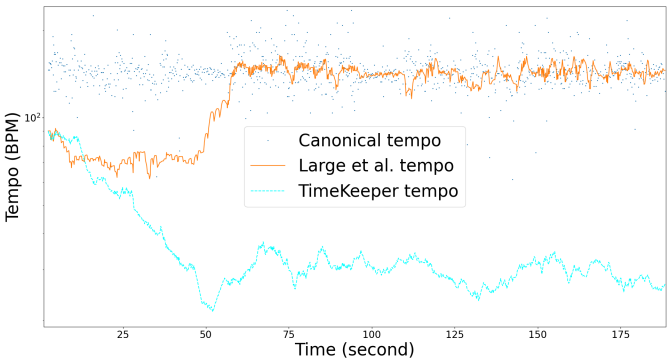


Figure 3: Tempo curve for a performance of Piano Sonata No. 11 in A Major (K. 331: III), W.A Mozart, according to the two previously modified models, with irrelevant initialization

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. We therefore 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 to actually exists, and use to notation (b_n) to designate it for formal proofs. One would 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 sensibility to tempo initialization, unstability nor require to accurately estimate relevant values of some constant internal parameters. According to our implementation, *Large et al.* model is a particularly chaotic model regarding the latter.

This section present 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 $(u_n)_{n\in\mathbb{N}}$, let from now on $(\Delta u_n)_{n\in\mathbb{N}}$ be $(u_{n+1} - u_n)_{n\in\mathbb{N}}$

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 α 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 allow for a locally constant shift in both of our estimation of the numerator and denumerator). 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 will then define $T_{n+1} = \alpha_n T^*_{n+1}$, to find : $\frac{\Delta b_{n+1}}{\Delta b_n} = \frac{T^*_{n+1}\Delta t_{n+1}}{T^*_n \Delta t_n} = \frac{\frac{1}{\alpha_n}T_{n+1}}{\frac{1}{\alpha_n}T_n} \times \frac{\Delta t_{n+1}}{\Delta t_n}$,

(*) ill faudraut oeut être donner des noms au modele de [6] et sa version modifiée, pour les différentier, par exemple Oscillator et Oscillator-mod

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 estimate correctly 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).
Let us then write the actual formula of the model :

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

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). Actually, in practice, E is 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) :

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

where $d : a, b \mapsto k_* \left|\log\left(\frac{a}{b}\right)\right|, k_* \in \mathbb{R}^*_+$, is a logarithmic distance, choosen since an absolute distance would have favor small values by triangle inequality in the following process.
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 contained micro-errors (that we call here timing). In our tests, we limited these outputs to be 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.

B. *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 $\left(\alpha_n := \frac{T_n}{T^*_n}\right)_{n \in [\![1,N]\!]}$ and $\left(\tilde{\alpha}_n := \exp(\ln(\alpha_n) - \lfloor \log_2(\alpha_n) \rfloor \ln(2))\right)_{n \in [\![1,N]\!]}$. The latter is then 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 : $c : [\tilde{1}, \tilde{2}[\to \mathcal{C}(0,1)$

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

Definition III.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 $\mathcal{C}$:

$$m(S, \Delta) = \max_{\tilde{d} \in \mathcal{C}} \frac{\left|\left\{n \in [\![1, |S|]\!] : d(\tilde{\alpha}_n, \tilde{d}) \leq \Delta\right\}\right|}{\min(1, |S|)} \tag{10}$$

Here d is still a logarithmic distance, slightly modified on $\mathcal{C}$ to be consistent with $d(\tilde{1}, \tilde{2}) = 0$. Actually, it can be shown that on $\mathcal{C}$, $d : \tilde{a}, \tilde{b} \mapsto \min(|\tilde{a} - \tilde{b}|, 1 - |\tilde{a} - \tilde{b}|)$.

Proposition III.2 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 $[\tilde{1}, \tilde{2}[$.
Then : $m(S', \Delta) = m(S, \Delta)$ and :

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

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 4 presents those results in a global display. The red values corresponds to the pieces written by W.A Mozart, that usually contain mainly regular division, and thus we expect to obtain a *measure* closer to 1. On the other hands, the blue values corresponds to M. Ravel's pieces, much more rhythmically expressive, where we expect a *measure* closer to 0. To understand the global results, we present in Figure 5 the same results for a random estimator.

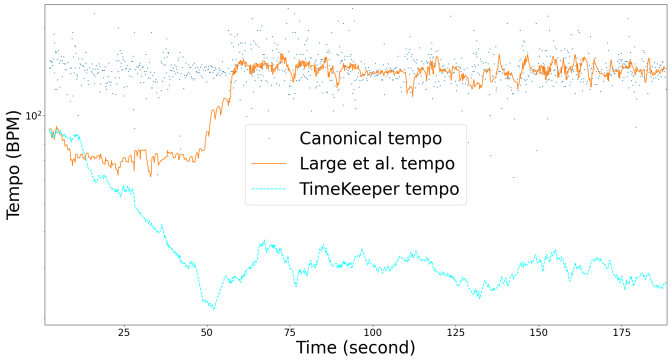


Figure 4: Measure of the resulting spectrum over the whole (n)-ASAP dataset with $\Delta = 0.075$

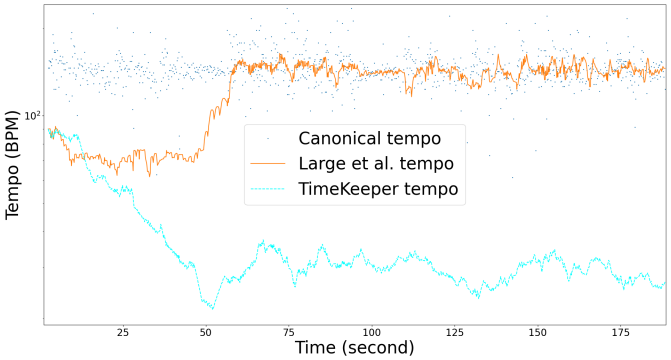


Figure 5: Measure of the resulting spectrums over the whole (n)-ASAP dataset with $\Delta = 0.075$, with an estimator outputing random quantized values. We therefore consider typical values of this plot to be representative of an unacceptable spectrum, corresponding thus to an inadequate tempo curve.

C. *Towards a quantized approach*

In this section, we extend the previous approach by considering the estimator as our central model and then extracting tempo values rather than the opposite, by extending G. Romero-García, C. Guichaoua, and E. Chew [15] model with the previous formalism.

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

$$\varepsilon_D : a \mapsto \max_{d \in D} \min_{m \in \mathbb{Z}} |d - ma| \tag{11}$$

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 local maxima of ε_D, except those that also are minima, is :

$$\begin{aligned} M_D &= \left\{ \frac{d}{k + \frac{1}{2}}, d \in D, k \in \mathbb{N} \right\} \\ &= \bigcup_{d \in D} \left\{ \frac{d}{k + \frac{1}{2}}, k \in \mathbb{N} \right\} \end{aligned} \tag{12}$$

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, $a_1 \in]m_1, m_2[$, $a_2 \in [m_1, m_2]$, $d \in D$ and $m \in \mathbb{Z}$. Then $m \in \underset{k \in \mathbb{Z}}{\operatorname{argmin}} |d - ka_1| \Rightarrow m \in \underset{k \in \mathbb{Z}}{\operatorname{argmin}} |d - ka_2|$.

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

$M \leftarrow \left\{ \frac{d}{k+\frac{1}{2}}, d \in D, k \in \left[\!\left[\frac{d}{\text{end}}, \frac{d}{\text{start}} \right]\!\right] \right\}$
$P \leftarrow \left\{ \frac{d_1+d_2}{k}, (d_1, d_2) \in D^2, k \in \left[\!\left[\frac{d_1+d_2}{\text{end}}, \frac{d_1+d_2}{\text{start}} \right]\!\right] \right\}$
localMinima $\leftarrow \emptyset$
for $(m_1, m_2) \in M^2$ two successive 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: 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.D). 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 (12), 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 (12) 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 to Proposition III.1, we know that both these local minima will quantize the elements of D the same way. 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 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 III.2 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 $m_1 \neq m_2$. WHY IS IT ABSURD ?

D. 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

E. résultats évaluation (comparaison avec 3)

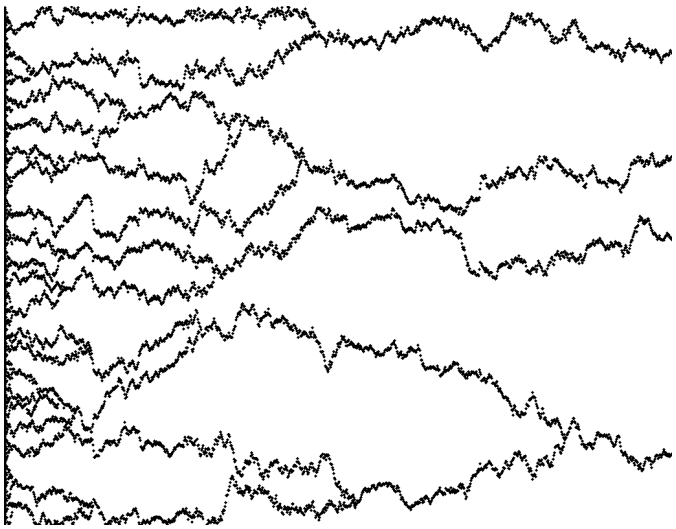


Figure 6: All potentials tempo curves found by a quantized approach for a performance of Piano Sonata No. 11 in A Major (K. 331: III), W.A Mozart. The tempo scale is linear between ♩ = 40 (bottom) and ♩ = 240 (top)

IV. Applications

A. A method for data augmentation

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 :

$$\hat{t}_0 = t_0, \forall n \in \mathbb{N}, \hat{t}_{n+1} = \hat{t}_n + \underbrace{\Delta t_n \frac{T^*_n}{T_c}}_{\alpha_n} + \underbrace{\frac{s_n}{T_c}}_{\beta_n} \quad (13)$$

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 IV.1 $\left(\hat{t}_n\right)_{n\in\mathbb{N}}$ 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, $\hat{T}^*_n = \frac{\Delta b_n}{\Delta \hat{t}_n} = \frac{\Delta b_n}{\Delta \frac{b_n}{T_c} + \frac{s_n}{T_c}} = \frac{1}{1+\frac{s_n}{\Delta b_n}} T_c$ □

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 IV.1 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?

B. Generated performances thanks to the previous method

- 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.D 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.
- usage in performance generation with Section IV.A and O. F. B. Katerina Kosta Rafael Ramírez and E. Chew [2]

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.

[19]
A coupled duration-focused architecture for real-time music-to-score alignment
IEEE transactions on pattern analysis and machine intelligence, 2009

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 stated in Proposition II.2.

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.A 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 $\bar{t} = r\Big((t_n)_{n\in\mathbb{N}}\Big)$ are correct representations of a sheet music and performance respectively, as defined in Section II.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 $T_r : t \mapsto T(t^* - t)$.

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

Hence T_r is a formal tempo according to $(\overline{t_n})$ and $\big(\overline{b_n}\big)$.

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 $d(a,b) = 0 \Leftrightarrow \big|\log\big(\frac{a}{b}\big)\big| = 0 \Leftrightarrow a = b$. Finally, let $c \in R_+^*, d(a,c) = k_*\big|\log\big(\frac{a}{c}\big)\big| = k_*\Big|\log\Big(\frac{a}{b}\times\frac{b}{c}\Big)\Big| = k_*\Big|\log\big(\frac{a}{b}\big) + \log\Big(\frac{b}{c}\Big)\Big| \leq d(a,b) + d(b,c)$.

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 $k_* \in R_+^*$ does not affect the result of the process.

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

In order to resist to the tempo octave problem discussed in Appendix A, we choose here to consider a unique candidate within a range $[x, 2x] \subset \big[\frac{1}{2}, 2\big]$, 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 : $\log\big(\frac{1}{x}\big) = \log(2x) \Leftrightarrow -\log(x) = \log(2) + \log(x) \Leftrightarrow \log(x^2) = -\log(2) \Leftrightarrow x^2 = \frac{1}{2} \Leftrightarrow x = \frac{\sqrt{2}}{2}$ since $x > 0$.

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

$$d(T_{n+1}, T_n) = k_*|\log(y)| = d(1, y) \tag{14}$$

where $y = \underset{x'\in[x,2x]}{\operatorname{argmin}}\, d\Big(x', \frac{\Delta t_n}{\Delta t_{n+1}} E\Big(x'\frac{\Delta t_{n+1}}{\Delta t_n}\Big)\Big)$.

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 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.

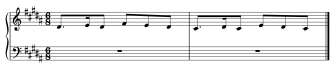


Figure 7: Transcription A

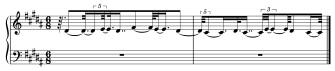


Figure 8: Transcription B

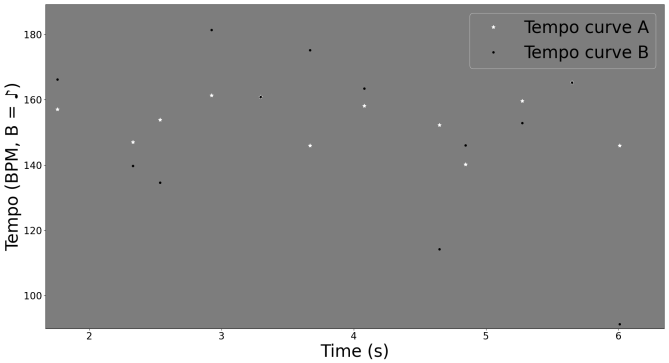


Figure 9: Tempo curves A and B plotted together

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 10.

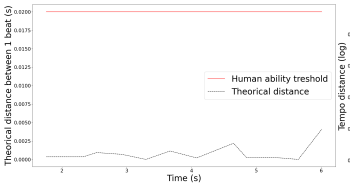


Figure 10: Tempo distance (s)

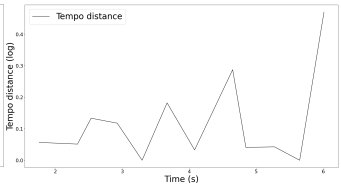


Figure 11: Tempo distance (log)

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 10) 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 :

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

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

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

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

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

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

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

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 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,

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

Hence $A \subset \left\{\frac{t_1 \mp t_2}{n}, (t_1,t_2) \in T^2, n \in \mathbb{Z}^*\right\}$, because $T \subset (R_+^*)^{|T|}$. 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 $f_t : a \mapsto ag(\frac{t}{a})$ 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 maxima 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 $x \mapsto \max\left(\max_{t'\in T\setminus\{t\}} xg\left(\frac{t'}{x}\right), \varepsilon_T(a)\right)$ 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 $N(\frac{t}{a})^+$, $\frac{t}{a}$ 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 $a = \frac{t}{k+\frac{1}{2}}, k \in \mathbb{N}$, since $a > 0$.

2) *Second implication:*
Let $(t,k) \in T \times \mathbb{N}, a = \frac{t}{k+\frac{1}{2}}$.
By definition : $g\left(\frac{t}{a}\right) = g\left(\frac{1}{2}+k\right) = g\left(\frac{1}{2}\right) = \frac{1}{2} = \max_{\mathbb{R}} g$.
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 $\mathrm{tmp} : x \mapsto g\left(\frac{t_1}{x}\right) - g\left(\frac{t_2}{x}\right)$ be a continous function on $N(a)^+$ and A be the set of all $x^* \in N(a)^+$ so that $\mathrm{tmp}(x^*) = 0 \Leftrightarrow g\left(\frac{t_1}{x^*}\right) = g\left(\frac{t_2}{x^*}\right)$.

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 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 $t^* = \underset{t' \in T^*}{\mathrm{argmax}}\, g\left(\frac{t'}{a_1}\right)$ 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, $\varepsilon_T(a) > \varepsilon_T(x)$.
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.C, we proved (12)

■

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, 2, 13

***RTU* – Real Time Unit**: Time unit to represent real events. Here, we usually use seconds as RTU. 1, 2, 13

***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, 13

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

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

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.D 4

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.. 6, 13

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. 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, 5

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... 12

triplet: A triplet is a musical symbol indicating to play a third of the indicated duration. 9

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