Numerical sheet music 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

I. Introduction

We present here some results regarding the analysis of tempo curve of musical performances, with score-based and scoreless approaches extending previously existing models.

The Music Information Retrieval (MIR) community focus on three ways to compute musical information. The first one is raw audio, either recorded or generated, encoded in .wav or .mp3 files. The computation is based on a physical understanding of signals, using audio frames and spectrum, and represents the most common and accessible type of data. The second is a more musically-informed format, that indicates mainly two parameters : pitch (ie the note that the listener hear) and duration, encoded within a .mid (or MIDI) file. Such a file can be displayed as a piano roll, that is a graph whose x-axis is time and y-axis is pitch (hence, the y-axis is discrete). The last way to encode musical information is the computed counterpart of sheet music. A sheet music is a way to write down a musical score, that is usually computed as a .music_xml file, mainly for display purposes. It comes with 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 the pitch of each event. This kind of data is actually the least common and accessible.

To actually play a sheet music, one needs a given tempo, usually indicated as the amoung of beat per minute (BPM). Therefore, the notion of tempo allows to translate symbolic notation (expressed in musical unit, eg : beats) to real time events (expressed in real time unit, eg : seconds). We will discuss later on a formal definition of tempo. However, tempo itself is insufficient to describe an actual performance of a sheet music, ie the 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 in contrast with velocity and timing. The latter represents the delay between the theorical real time onset according to the current tempo, and the actual onset heared 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.

In this study, we shall focus mainly on tempo estimation for a given performance recorded as a MIDI file, on both a local and global level.

II. State of Art

Even though the community studies the four parameters, the hierarchy [1] exposed embodies quite well the importance within the litterature. O. F. B. Katerina Kosta Rafael Ramírez and E. Chew [2] present results pointing that, although velocities don't help to meaningfully estimate tempo, the latter allows to marginally upgrade velocity-related predictions. Actually, velocity appears to be more of a score parameter rather than a performance one : automatic learning methods trained on performances of a single piece showed much better results when asked to predict velocities employed by another performer on the same piece than when trained on other performances of the same performer.

Tempo and related works actually hold a prominent place in litterature. Direct tempo estimation was first computed based on probabilistic models (C. Raphael [3], E. Nakamura, N. Ono, S. Sagayama, and K. Watanabe [4], [5]), and physical / neurological models (E. W. Large and M. R. Jones [6], H.-H. Schulze, A. Cordes, and D. Vorberg [7]) ; before the community tried neural network models [2] and hybrids approaches (K. Shibata, E. Nakamura, and K. Yoshii [8]). As the majority of 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 keep this discussion for section V and VI.

However, note-alignement, that is matching 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 (M. Müller [12]), that can works on raw audio (.wav files) ; and a Hidden Markov Model (E. Nakamura, K. Yoshii, and H. Katayose [13]) that needs more formatted data, such as MIDI files.

In this report, we will present the following contributions :

- a justified proposition for a formal definition of tempo based on C. Raphael [3], [9] and P. Hu and G. Widmer [11] ; and some immediate consequences
- a revision of E. W. Large and M. R. Jones [6] and H.-H. Schulze, A. Cordes, and D. Vorberg [7] to fit a score-based approach
- an extension of G. Romero-García, C. Guichaoua, and E. Chew [14], to fit tempo estimation
- generated data based on F. Foscarin, A. Mcleod, P. Rigaux, F. Jacquemard, and M. Sakai [15] and S. D. Peter *et al.* [16]

III. 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 events $(t_n)_{n\in\mathbb{N}}$, each element of whose indicates the onset of the 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 E. Nakamura, T. Nakamura, Y. Saito, N. Ono, and S. Sagayama [5] represents the limits of human ability to tell two rythmic events appart, and is widely used within the field [8]–[11], [13]–[17].

Likewise, a sheet music will be represented as a strictly increasing sequence of 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 real onset, thus expressed in seconds, whereas (b_n) corresponds to theorical or symbolic onsets, expressed in beats.

With those definitions, let us formally define tempo $T(t)$ so that, for all $n \in \mathbb{N}$, $\int_{t_0}^{t_n} T(t)\,\mathrm{d}t = b_n - b_0$.
Appendix A shows that this definition is equivalent to : $\forall n \in \mathbb{N}, \int_{t_n}^{t_{n+1}} T(t)\,\mathrm{d}t = b_{n+1} - b_n$. However, tempo is only tangible (or observable) between two events *a priori*. We will then define the canonical tempo $T^*(t)$ 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 according to the previous definition. From now on, we will consider the convention : $t_0 = 0$ (s) et $b_0 = 0$ (beat).

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 .music_xml 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], [15], [16].

In our study, we chose to rely on the (n)-ASAP dataset [16] 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.

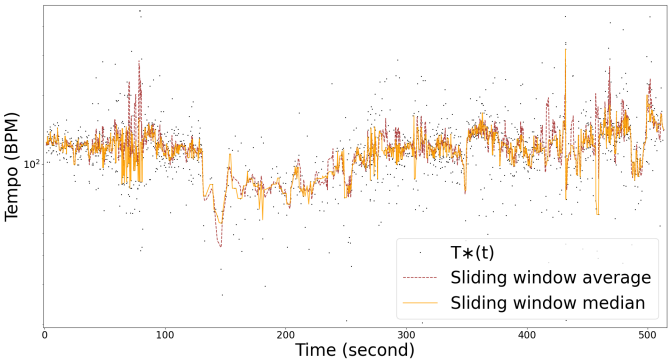

Figure 1: Tempo curve for a performance of Islamey, Op.18, M. Balakirev, with naive algorithm

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 theorical 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 by A. Cont [18].

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:

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

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

Here, (Φ_n) corresponds to the phase, or rather the phase shift 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.* [19]), a theorical study of the system behavior remains quite complex, even in simplified theorical 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 theorical 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 [20]. Here, we used the derandomised version presented in [20], 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 the both rely on a rather stable metronome. In the version displayed hereafter, they were modified to consider score information, in the goal 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 theorical framework of validity. Let $\min_{\text{abs}} : a, b \mapsto \begin{cases} a \text{ if } |a| < |b| \\ b \text{ otherwise} \end{cases}$

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

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

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

$$\Psi_n = -\min_{\text{abs}}(k + b_n - \Phi_n, k+1+b_n-\Phi_n)$$

$$k = \lfloor \Phi_n - b_n \rfloor \quad .$$

Using the analogy presented in [20], we then obtain the following equations for *TimeKeeper* :

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

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

$$K_i = -\min_{\text{abs}}(k\tau + b_i - A_i, (k+1)\tau + b_i - A_i)$$

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

In both of those extensions, the $\min_{\text{abs}}$ 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, ie $a_1 = \max_{n\in\mathbb{N}\ :\ b_n \le B_i} \lfloor b_n \rfloor$ where B_i represents the internal value at time i acting as a beat unit (Φ_i 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, ie $\forall n \in \mathbb{N}, b_n = 0 \bmod 1$, the extension proposed here is equivalent to the original approach, ie $(1), (2) \Leftrightarrow (3), (4)$, 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.

[1] https://metronautapp.com/

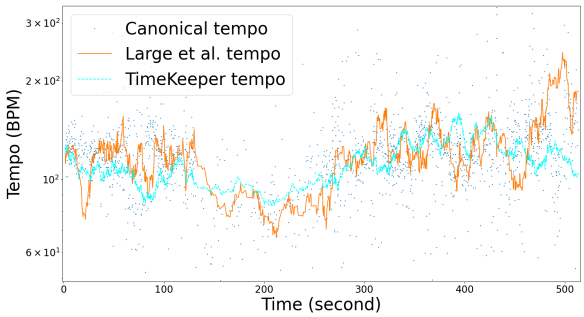

Figure 2: Tempo curve for a performance of Islamey, Op.18, M. Balakirev, according to various models

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, ie 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 theorical 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).

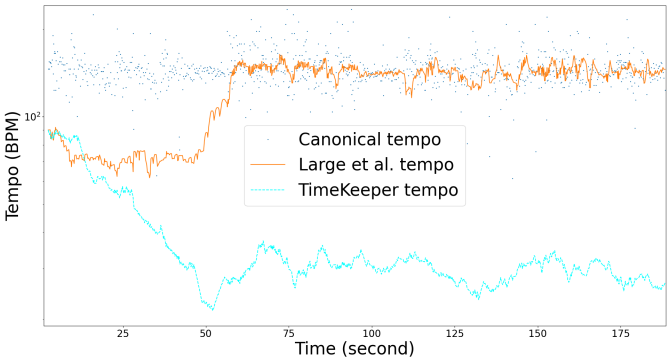

Figure 3: Tempo curve for a performance of Piano Sonata No. 11 in A Major (K. 331: III), W.A Mozart, according to the previous models with irrelevant initialization

IV. Scoreless approaches

A. Motivations

There are three main issues with the previous models, apart from the necessary 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 (ie 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, all corresponding to the same canonical tempo.

We considered here two quantization methods extracted from litterature : D. Murphy [17] and G. Romero-García, C. Guichaoua, and E. Chew [14]. Both of these papers present a theory of approximate division, that is a way to find a rationnal 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 [17] provides an algorithm to find candidate rationnals, 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 [14] 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 $T^*_n > 0$:

$$T^*_{n+1} = T^*_n \frac{T^*_{n+1}}{T^*_n} = T^*_n \frac{\Delta t_n}{\Delta t_{n+1}} \frac{\Delta b_{n+1}}{\Delta b_n} = T^*_n \frac{\Delta t_n}{\Delta t_{n+1}} \times \frac{T^*_{n+1}\Delta t_{n+1}}{T^*_n \Delta t_n}$$

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 (ie 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. Moreover, we 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 :

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

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 theorical 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 theorical 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) :

$$T_{n+1} = T_n \underset{x \in \left[\frac{2}{3}T_n, \frac{4}{3}T_n\right]}{\operatorname{argmin}} d\left(x, \frac{\Delta t_n}{\Delta t_{n+1}} E\left(x \frac{\Delta t_{n+1}}{\Delta t_n}\right)\right) \tag{8}$$

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

In the implementation presented here, the estimator role is more to quantify the ratio in order to output a musically relevant value. In our test, we limited these quantification to accept only regular division (ie 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 (ie 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

Mesure, spectres, résultats sur Asap selon les périodes, les compositeurs, etc..., éventuellement en annexe ?

D. Quantified approach

- 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)

plutôt en annexe je pense

V. Applications

- previous : metronaut, antescofo
- 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
- accompagnement automatique RT avec approche 4 RT ?

VI. Conclusion & perspectives

- intégration pour couplage avec transcription par parsing (+ plus court chemin multi-critère)
- lien approche partie 4 "spectrale" avec Large (amortisseur)

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] 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

[15] 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

[16] 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.

[17] 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.

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

[19] 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.

[20] 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 : FORMALISM GENERAL RESULTS

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, $\int_{t_n}^{t_{n+1}} T(t)\,\mathrm{d}t = \int_{t_0}^{t_{n+1}} T(t)\,\mathrm{d}t + \int_{t_n}^{t_0} T(t)\,\mathrm{d}t = \int_{t_0}^{t_{n+1}} T(t)\,\mathrm{d}t - \int_{t_0}^{t_n} T(t)\,\mathrm{d}t$.
We thus obtain the two implications, hence the equivalence. Rajouter quelques figures éventuellement, expliquer le problème d'octave de tempo + réf quelque part !

B. Going back in time

APPENDIX B : ESTIMATOR MODEL

- d est une distance
- argmin : d est continue (distance), donc bornée et atteint ses bornes sur l'intervalle

Remarque sur la distance log (pas d'importance de k_*).
Pourquoi 2/3 et 4/3 :

- candidat unique pour chaque changement de tempo : résistant aux octaves
- easier to search ($|\frac{2}{3}-1| = |\frac{4}{3}-1|$) : solution of $2x - 1 = 1 - x \Leftrightarrow 3x = 2$: on veut un tempo unique, DONC (à montrer éventuellement) entre x et $2x$, et on prend x "centré" en 1.
- si l'estimateur est l'identité : par hypothèse d'oracle, le tempo est constant, et joué parfaitement (ie fichier midi), et le résultat est le bon
- l'estimateur n'est pas une fonction : l'output à un instant donné peut dépendre des outputs précédents (cas extrême : transcription en temps réel) : très bonne courbe, mais estimateur très complexe, or le problème ici est justement relaxé : ajouter un exemple de fausse transcription et de correcte avec les courbes de tempo correspondantes.

APPENDIX C : QUANTIFIED MODEL

Posons tout d'abord quelques fonctions utiles.
On définit : $g : x \mapsto \min(x - \lfloor x \rfloor, 1 + \lfloor x \rfloor - x)$
On peut vérifier que $g : x \mapsto \begin{cases} x - \lfloor x \rfloor \text{ si } x - \lfloor x \rfloor \leq \frac{1}{2} \\ 1 - (x - \lfloor x \rfloor) \text{ sinon} \end{cases}$ et que g est 1-périodique continue sur $\mathbb{R}$.
Ainsi, on a : $\varepsilon_T(a) = \max_{t \in T}\Big(\min_{m \in \mathbb{Z}} |t - ma|\Big) = \max_{t \in T} \min\big(t - \lfloor \frac{t}{a} \rfloor a, (\lfloor \frac{t}{a} \rfloor + 1)a - t\big) = a \max_{t \in T} \underbrace{\min\Big(\frac{t}{a} - \Big\lfloor \frac{t}{a} \Big\rfloor, \Big\lfloor \frac{t}{a} \Big\rfloor + 1 - \frac{t}{a}\Big)}_{g(\frac{t}{a})} = a \max_{t \in T} g\Big(\frac{t}{a}\Big)$,
donc en particulier, ε_T est continue sur R_+^*.

On remarque de plus, pour $n \in \mathbb{N}^*, T \subset (\mathbb{R}_+^*)^n, a \in R_+^*$: $\varepsilon_T(a) = a\varepsilon_{T/a}(1)$. Hence the intuitive following result : the smaller the tatum, the smaller the bound of the error.

A. Caractérisation des maximums locaux

1) *First implication:*
Let a be a local maxima 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 $t \in \operatorname{argmax}_{t' \in T} g\big(\frac{t'}{a}\big)$, hence $\varepsilon_T(a) = ag\big(\frac{t}{a}\big)$.
For all $\delta \in]-\varepsilon, \varepsilon[, \varepsilon_T(a) = ag\big(\frac{t}{a}\big) \geq \varepsilon_T(a + \delta)$,
and $\varepsilon_T(a + \delta) = (a + \delta) \max_{t' \in T} g\big(\frac{t'}{a+\delta}\big) \geq (a + \delta) g\big(\frac{t}{a+\delta}\big)$.
For $\delta \geq 0, a + \delta \geq a$, so $ag\big(\frac{t}{a}\big) \geq (a + \delta) g\big(\frac{t}{a+\delta}\big) \geq ag\big(\frac{t}{a+\delta}\big)$
Hence, $g\big(\frac{t}{a}\big) \geq g\big(\frac{t}{a+\delta}\big)$ 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 $g\big(\frac{t}{a}\big) = \frac{t}{a} - \big\lfloor \frac{t}{a} \big\rfloor, \varepsilon_T(a) = t - a\big\lfloor \frac{t}{a} \big\rfloor$.

The function ε_T is the maximum of a finite set of continuous functions, with a countable set of A of intersection, ie $A = \Big\{x \in \mathbb{R}_+^* : \exists (t_1, t_2) \in T^2 : t_1 \neq t_2 \wedge xg\big(\frac{t_1}{x}\big) = xg\big(\frac{t_2}{x}\big)\Big\}$. Indeed,
$x \in A \Leftrightarrow \exists (t_1, t_2) \in T^2 : t_1 \neq t_2 \wedge g\big(\frac{t_1}{x}\big) = g\big(\frac{t_2}{x}\big)$
$\Leftrightarrow \exists (t_1, t_2) \in T^2 : t_1 \neq t_2 \wedge \frac{t_1}{x} = \pm\frac{t_2}{x} \bmod 1$
$\Leftrightarrow \exists (t_1, t_2) \in T^2 : t_1 \neq t_2 \wedge x = \frac{t_1 \mp t_2}{n}, n \in \mathbb{Z}^*$
Hence $A \subset \Big\{\frac{t_1 \mp t_2}{n}, (t_1, t_2) \in T^2, n \in \mathbb{Z}^*\Big\}$, because $T \subset (R_+^*)^{|T|}$. Therefore, there is a countable set of closed convex intervalls, which union is R_+^* so that on each of these intervalls, ε_T is equal to $f_t : a \mapsto ag\big(\frac{t}{a}\big)$ 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 intervalls. Since f_t and ε_T are both continuous on $[a - \delta', a]$, $f_t(a) = \varepsilon_T(a)$ and therefore, $t \in \operatorname{argmax}_{t' \in T} g\big(\frac{t'}{a}\big)$. 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 $N\big(\frac{t}{a}\big)^+$ in the following, the previous expression of g remains valid, ie $\forall x \in N\big(\frac{t}{a}\big)^+, g(x) = x - \lfloor x \rfloor$. Moreover, $x \mapsto \lfloor x \rfloor$ is right-continuous, hence by restricting $N\big(\frac{t}{a}\big)^+$, we can assure for all $x \in N\big(\frac{t}{a}\big)^+, \lfloor x \rfloor = \big\lfloor \frac{t}{a} \big\rfloor$. Let then $y = a - \frac{t}{x}$ so that $x = \frac{t}{a-y}$, we then have $\varepsilon_T(a - y) = t - (a - y)\big\lfloor \frac{t}{a} \big\rfloor \leq \varepsilon_T(a) = t - a\big\lfloor \frac{t}{a} \big\rfloor$ 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 $N\big(\frac{t}{a}\big)^+$. Hence, $t - \big\lfloor \frac{t}{a} \big\rfloor a \geq t - (a - y)\big\lfloor \frac{t}{a} \big\rfloor$ ie $a\big\lfloor \frac{t}{a} \big\rfloor \leq (a - y)\big\lfloor \frac{t}{a} \big\rfloor \Leftrightarrow 0 \leq -y\big\lfloor \frac{t}{a} \big\rfloor$ ie $\big\lfloor \frac{t}{a} \big\rfloor =$

0 ie $\left\lfloor \frac{t}{a} \right\rfloor = 0$, since y, t and a are all positive values. Then, $a > t$ and in this case, the local maxima is not strict. Such a maxima is rather ininteresting in our study, since it corresponds to an intervall where ε_T is constant (at least a left neighbourhood of a). Indeed, $\varepsilon_T(a - y) = t = \varepsilon_T(a)$ for $\left\lfloor \frac{t}{a} \right\rfloor = 0$. This constant intervall is then either going on infinitely on the right of a, or else ε_T will reach a value greater than $\varepsilon_T(a) = t$, since ε_T can then be rewritten as $x \mapsto \max\left(\max_{t' \in T \setminus \{t\}} x g\left(\frac{t'}{x}\right), \varepsilon_T(a)\right)$ on $[a, +\infty[$, hence the interesting point, if any, will be a local minimal, that we will consider later. GIVE THE EXPRESSION TO FIND IT OR MODIFY THE ALGORITHM TO FIND IT AS WELL.
On the other hand, on the left of a : BETTER LOCAL MAXIMA TO BE FOUND.

- 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, $\frac{t}{a} = \frac{1}{2} \bmod 1$, ie $\frac{t}{a} = \frac{1}{2} + k, k \in \mathbb{Z}$, 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, $\varepsilon_T(a) = a \max_{t' \in T} g\left(\frac{t'}{a}\right) = a g\left(\frac{t}{a}\right) = \frac{a}{2}$.
For all $x \in]0, a[, \varepsilon_T(x) = x \max_{t' \in T} g\left(\frac{t'}{x}\right) \leq \frac{x}{2} < \frac{a}{2} = \varepsilon_T(a)$, ie $\varepsilon_T(a) > \varepsilon_T(x)$.
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 $\frac{t_1}{a} = \frac{t_2}{a} \bmod 1$, ie $|\frac{t_1}{a} - \frac{t_2}{a}| = k_a \in \mathbb{N}$.

Then, $\frac{t_1}{x^*} = \pm\frac{t_2}{x^*} \bmod 1$ ie $|\frac{t_1}{x^*} \mp \frac{t_2}{x^*}| = k_* \in \mathbb{N}$, and therefore $|t_1 \mp t_2| = a k_a = x^* k_*$, 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, ie $|A| \in \mathbb{N}$.
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 $t^* = \operatorname{argmax}_{t' \in T^*} g\left(\frac{t'}{a_1}\right)$ 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 $\varepsilon_T(x) = x g\left(\frac{t^*}{x}\right)$.
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 $\frac{t^*}{a}$, ie 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, $\varepsilon_T(x) = x g\left(\frac{t^*}{x}\right) = x\left(\frac{t^*}{x} - \left\lfloor \frac{t^*}{x} \right\rfloor\right) = t^* - x\left\lfloor \frac{t^*}{x} \right\rfloor$ 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 $y = \frac{t^*}{\alpha}$, ie $\alpha = \frac{t^*}{y}$, 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 the set of all local maxima of ε_T is $M_T = \left\{\frac{t}{k+\frac{1}{2}}, t \in T, k \in \mathbb{N}\right\}$ ■

B. Caractérisation des minimums locaux

Let a be a local minima of ε_T, ie

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 $]2\tau, t_* + \tau[$

APPENDIX D : MUSICAL GLOSSARY

A. Acronyms

MIR – Music Information Retrieval: Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do. 1

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:
Unité de temps d'une partition, le beat est défini par une signature temps, ou division temporelle. Bien que sa valeur ne soit *a priori* pas fixe d'une partition à une autre, ni même sur une même partition, la notion de beat est en général l'unité la plus pratique quant à la description d'un passage rythmique, lorsque la signature temps est adéquatement définie. 1, 9

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

measure: Une mesure est une unité de temps musicale, contenant un certain nombre (entier) de beat. Ce nombre est indiqué par la time signature 2

online: Définition d'une méthode / d'un algo online 5

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

tatum: Résolution minimal d'une unité musicale, exprimé en beat. Bien que de nombreuses valeurs soit possible, la définition formelle d'un tatum serait la suivante : $\sup\{r \mid \forall n \in \mathbb{N}, \exists k \in \mathbb{N} : b_n = kr, r \in \mathbb{R}_+^*\}$. Pour des raisons pratiques, il arrive que le tatum soit un élément plus petit que la définition donnée, en particulier si cet élément est plus facilement expressible dans une partition, ou a plus de sens d'un point de vue musical. On notera dans la définition de l'ensemble donnée, k n'a pas d'unité, ce qui montre clairement que le tatum s'exprime en beat comme dit précédemment. 4

tempo:
Défini formellement p. 2 selon la formule : $T_n^* = \frac{b_{n+1}-b_n}{t_{n+1}-t_n}$. Informellement, le tempo est une mesure la vitesse instantanée d'une performance, souvent indiqué sur la partition. On peut le voir comme le rapport entre la vitesse symbolique supposée par la partition, et la vitesse réelle d'une performance. Le tempo est usuellement indiqué en beat par minute, ou bpm 1

time signature. 9

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