TEMPO AND METRICAL ANALYSIS BY TRACKING MULTIPLE METRICAL LEVELS USING AUTOCORRELATION

Olivier Lartillot
RITMO Centre for Interdisciplinary Studies
in Rhythm, Time and Motion
University of Oslo
olivier.lartillot@imv.uio.no

Didier Grandjean
Department of Psychology
Swiss Center for Affective Sciences
University of Geneva
Didier.Grandjean@unige.ch

ABSTRACT

We present a method for tempo estimation from audio recordings based on signal processing and peak tracking, and not depending on training on ground-truth data. First an accentuation curve, emphasising the temporal location and accentuation of notes, is based on a detection of bursts of energy localised in time and frequency. This enables to detect notes in dense polyphonic texture, while ignoring spectral fluctuation produced by vibrato and tremolo. Periodicities in the accentuation curve are detected using an improved version of autocorrelation function. Hierarchical metrical structures, composed of a large set of periodicities in pairwise harmonic relationships, are tracked over time. In this way, the metrical structure can be tracked even if the rhythmical emphasis switches from one metrical level to another.

This approach, compared to all the other participants to the MIREX Audio Tempo Extraction from 2006 to 2018, is the third best one among those that can track tempo variations. While the two best methods are based on machine learning, our method suggests a way to track tempo founded on signal processing and heuristics-based peak tracking. Besides, the approach offers for the first time a detailed representation of the dynamic evolution of the metrical structure. The method is integrated into *MIRtoolbox*, a Matlab toolbox freely available.

1. INTRODUCTION

Detecting tempo in music and tracking the evolution of tempo over time is a topic of research in MIR that has been extensively studied these last decades. Recently approaches based on deep learning have contributed to an important progress in the state of the art [1, 2]. In this paper, we present a method that relates to a more classical approach based on signal processing and heuristics-based data extraction. We previously briefly presented the principles of the approach [3]. This paper offers a more detailed description of the method.

One particularity of the proposed approach is that it enables to track not only one, two or three, but a larger number of metrical levels. This enables to get a detailed description of the dynamic evolution of the metrical structure: not only how the whole structure speeds up or slows down with respect to global tempo, but also how individual metrical levels might be emphasised at particular moments in the music. In order to give an indication of metrical activity that would not reduce solely on tempo but takes into consideration the activity on the various metrical levels, we introduce a new measure, called *dynamic metrical centroid*.

2. RELATED WORK

2.1 Accentuation curve

The estimation of tempo starts from a temporal description of the location and strength of events appearing in the piece of music. This first step consists in inferring an "onset detection curve", also called *accentuation curve* [4]. Musical events are indicated by peaks; the height of each peak is related to the importance of the related event. *Envelope*-based approach globally estimates the energy for each successive temporal frame without considering its spectral decomposition; *spectral flux* methods estimate the difference of energy over successive frames on individual frequencies individually, and further summed together [5, 6]. The envelope approach would work in the case of sequences made of notes sufficiently isolated or accentuated with respect to the background, corresponding to short bursts of energy separated by low-energy transitions, as in simple percussive sequences. Indeed, in such case, the resulting envelope curve would show each percussive event with a peak. On the contrary, for dense musical sequences featuring overlapped notes, such as complex orchestral sounds, the spectral flux method better distinguishes the attack of individual notes, provided that the different notes occupy distinct frequency bands. Minor energy fluctuation along particular frequencies may blur the resulting accentuation curve in the point of making it impossible to detect the actual note attacks. The use of thresholding can filter out energy fluctuation on constant frequency bands (such as *tremolo*) and select only significantly high energy bursts related to note attacks. Still, energy fluctuating in frequency, such as *vibrato*, may still add noise to the resulting accentuation curve.

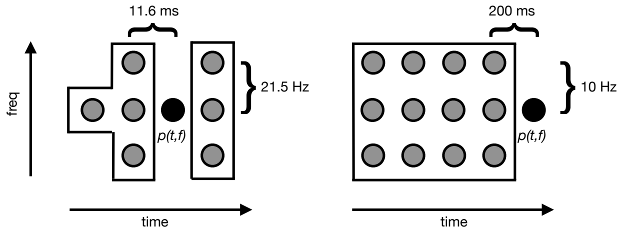

Figure 1. Comparison between the method in [7] (left) and our proposed method (right), for the comparison of a given spectrogram amplitude $p(t, f)$ at time t and frequency f with amplitudes from previous (and next) frames.

In order to detect significant energy bursts on highly localised frequency ranges but still filter out the artifacts due to the possible frequency fluctuation along time of such localised events, it is necessary to add some tracking capability. The approach presented by [7] can be considered as an answer to this problem. It searches for rapid increase of amplitude on particular frequency components, and evaluates for each detected onset its "degree of onset", defined as the rapidity of increase in amplitude. To estimate this increase of amplitude at a given time t for a particular frequency f, the amplitude $p(t, f)$ is compared not only to the corresponding amplitude at the previous frame $p(t-1, f)$, but also to the amplitude at the higher and lower frequency bins $p(t-1, f-1)$ and $p(t-1, f+1)$ as well as $p(t-2, f)$. These previous points form the *contextual background*. The current amplitude $p(t, f)$ is compared with the maximum of the amplitude at those four previous points, as shown in Figure 1. Let $pp(t, f)$ be this maximum. Besides, the corresponding amplitudes at frame $t+1$ are also compared with $pp(t, f)$.

For a given instant t and frequency f, the degree of onset is given by

$$do(t, f) = p(t, f) - pp(t, f) \tag{1}$$

The degrees of onset are summed over the frequency components, leading to the onset curve.

2.2 Periodicity analysis

A pulsation corresponds to a periodicity in the succession of peaks in the accentuation curve. Classical signal-processing methods estimate periodicity using methods such as autocorrelation, the YIN method, bank of comb-filter resonators with a constant half-time [8] or phase-locking resonators. [1] Basically, a range of possible periodicity frequencies is considered, and for each frequency, it is estimated whether there exists any periodicity at that frequency. In the following, we will call *periodicity function* the representation, such as autocorrelation function, showing the periodicity *score* related to each possible *period* (or alternatively each possible frequency).

2.3 Metrical structure

One common approach to extract the tempo from the periodicity function is to select the highest peak, within a range of beat periodicities considered as most adequate, typically between 40 and 200 BPM, with a weighted emphasis on best perceived periodicity range. This approach fails when tracking the temporal evolution of tempo over time, especially for pieces of music where different metrical levels are emphasised throughout the temporal development. For instance, if at a given moment of the piece of music, there is an accentuated quarter note pulsation followed by an accentuated eighth note pulsation, the tempo tracking will switch from one BPM value to another one twice faster, although the actual tempo might remain constant. And as we may imagine, such shift from metrical level to another is very frequent in music.

In [4], three particular metrical levels are considered as core elements of the metrical structure: The *tactus* is considered as the most prominent level, also referred as the foot tapping rate or the *beat*. The tempo is often identified to the tactus level. The *tatum*—for "temporal atom"—is considered as the fastest subdivision of the metrical hierarchy, such that all other metrical levels (in particular tactus and bar) are multiples of that tatum. The *bar-level* or other metrical levels related to change of chords, melodic or rhythmic patterns, etc. The tracking of tempo along time result from a tracking of these three metrical levels using a Hidden Markov Model (HMM).

The tatum is considered (and modeled) as the minimal subdivision such that each other metrical level is a multiple of that elementary level, but this canonic situation does not describe all metrical cases: for instance, binary and ternary subdivisions often coexist, as we will see in section 5.

2.4 Deep-learning approaches

Recent deep-learning approches start from the computation of a spectrogram, eventually followed by a filtering that emphasises the contrast between successive frames, along each different frequency [1]. In [1], the successive frames of the spectrogram are then fed into a Bidirectional Long Short-Term Memory (BLSTM) Recurrent Neural Network (RNN). This network can be understood as performing both the detection of events based on local contrast asnd the detection of periodicity in the succession of events, along multiple metrical levels. This is followed by a Dynamic Bayesian Network that plays a similar role as the HMM, tracking the pulsation along two metrical levels (corresponding to beats and downbeats). In [2], the whole process consists in feeding the spectrogram to a convolutional neural network (CNN).

3. PROPOSED METHOD

The proposed method introduces improvements in the successive steps forming the traditional procedure for metrical analysis that were presented in sections 2.1, 2.2 and 2.3. Those improvements are as follows. A modification of the localised method for accentuation curve estimation enables to better emphasise note onsets in complex polyphony with vibrato and tremolo (section 3.1). Periodicity detection is performed using a modified version of autocorrelation function (section 3.2).

[1] Cf. [4] for a detailed literature review.

Besides, we introduce a new methodology for tracking the metrical structure along a large range of periodicity layers in parallel. The tracking of the metrical structure is carried out in two steps:

1. a tracking of the *metrical grid* featuring a large range of possible periodicities (section 3.3). Instead of considering a fix and small number of pre-defined metrical levels, we propose to track a larger range of periodicity layers in parallel.
2. a selection of core metrical levels, leading to a *metrical structure*, which enables the estimation of metre and tempo (section 3.4).

3.1 Accentuation curve

Our method for the inference of the accentuation curve follows the same general principle of the model introduced in [7], detecting and tracking the apparition of partials locally in the spectrogram, as explained in section 2.1. In our case, the spectrogram is computed for the frequency range below 5000 Hz and the energy is represented in the logarithmic scale in decibel.

We use different parameters for the specification of the temporal scope and the frequency width of the contextual background. In [7], the frequency width is of 43 Hz and the temporal depth of 23 ms. After testing on a range of musical styles, we chose a frequency width around 20 Hz and a temporal depth of .8 second (cf. Figure 1). By enlarging the temporal horizon of the contextual background, this enables to filter out *tremolo* effects and to focus on more prominent increase of energy.

In the proposed model, the second condition for onset detection specified in [7] —namely, that the energy on the frame succeeding the current one should be higher than the contextual background—is withdrawn, for the sake of simplicity. That constraint seems aimed at filtering out bursts of energy that are just one frame long, but bursts two frames long would not be filtered out. And we might hypothesise that short bursts of energy might still be perceived as events.

Finally, the degree of onset is different from the one proposed in [7]. Instead of conditioning the degree of onset to the increase of energy with respect to the contextual background, we propose instead to condition it to the absolute level of energy:

$$do(t, f) = p(t, f) \qquad (2)$$

This is because a burst of energy of a given level $p(t, f)$ might be perceived as strong, and could contribute therefore to the detection of a note onset, even if there was a relatively loud sound in the frequency and temporal vicinity. This modification globally improved the results in our tests.

In our proposed method, the accentuation curve shows more note onsets than in [7]. This leads to a more detailed analysis of periodicity and a richer metrical analysis. This allows sometimes the discovery of the underlying metrical structure that was hidden under a complex surface and was not detected using [7].

3.2 Periodicity analysis

Tempo is estimated by computing an autocorrelogram with a frame length of 5 seconds and hop factor 5%, for a range of time lags between 60 ms and 2.5 s, corresponding to a tempo range between 24 and 1000 BPM. The autocorrelation curve is normalized so that the autocorrelation at zero lag is identically 1.

A peak picking is applied to each frame of the autocorrelogram separately. The beginning and the end of the autocorrelation curves are not taken into consideration for peak picking as they do not correspond to actual local maxima. A given local maximum will be considered as a peak if its distance with the previous and successive local minima (if any) is higher than 5% of the total amplitude (i.e., the distance between the global maximum and minimum) of the autocorrelation function.

One important problem with autocorrelation functions is that a lag can be selected as prominent because it is found often in the signal although the lag is not repeated successively. We propose a simple solution based on the following property: For a given lag to be repeated at least twice, the periodicity score associated with twice the lag should have a high probability score as well. This heuristics can be implemented as a single post-processing operation applied to the autocorrelation function, removing all periodicity candidate for which there is no periodicity candidate at around twice its lag.

3.3 Tracking the metrical grid

3.3.1 Principles

In the proposed approach, we track a large range of possible metrical levels. This is done in two successive steps:

- the construction of a detailed set of periodicities inherent to the metrical structure, leading to what we propose to call a metrical *grid*, where individual periodicities are called *layers*,
- the selection among those metrical layers of core metrical *levels*, whose periods are in multiplicity ratios. All other layers of the metrical grid are simple multiples or submultiples of those metrical levels. One metrical level is selected as the most prevalent, for the determination of tempo.

For each metrical layer i, its *tempo* T_i (meaning the tempo related to the metrical grid by tapping on that particular metrical layer) and period τ_i are directly related to the tempo T_1 and period τ_1 of the reference layer $i = 1$:

$$T_i = \frac{T_1}{i}, \tau_i = \tau_1 i \qquad (3)$$

For instance, the tempo at metrical layer 2 is twice slower than the one at metrical layer 1. Although tempo can change over time, the tempo related to the different metrical periodicities conserve their multiplicity ratio, so that equation 3 remains valid in theory.

The tracking of the metrical grid over time requires a management of uncertainty and noisy data. Periodicity lags measured in the autocorrelogram do not exactly comply with the theoretical lags given by equation 3. For that reason, each metrical layer i is described by both:

- theoretically, the temporal series of periods $\tau_i(n)$ related to metrical layer i knowing the global tempo given by $\tau_1(n)$.
- practically, the temporal series of lags $t_i(n)$ effectively measured at location of peaks in the autocorrelation function for each successive frame n.

In the graphical representation of the metrical structure, both actual and theoretical periods are shown: the temporal succession of the theoretical values at a given metrical layer is shown with a line of dots, whereas the actual periods are indicated with crosses that are connected to the theoretical dot with a vertical line. For instance in Figure 2, we see a superposition of metrical layers, each with a label indicated on the left side, starting from layer 0.25 up until layer 4, with also a layer 4.25 appearing around 30 second after the start of the excerpt.

3.3.2 Procedure

The theoretical periods are inferred based on the measured periods, as we will see in equation 13.

The integration of peak into the metrical grids is done in three steps, related to the extension of metrical layers already registered, the creation of new metrical layers and finally the initiation of new metrical grids.

For each successive time frame n, peaks in the periodicity function (as specified in section 3.2) are considered in decreasing order of periodicity score. This is motivated by the observation that strongest periodicities, corresponding generally to important metrical levels, tend to show a more stable dynamic evolution and are hence more reliable guides for the tracking of the metrical structure. Weaker autocorrelation peaks, on the contrary, may sometimes result from a mixture of underlying local periodicities, hence might tend to behave more erratically. For each frame, the strongest peaks first considered enable to get a first estimation of the tempo $T_1(n)$ at that frame, which will be used as a reference when integrating the weaker periodicities.

Each peak, related to a period (or lag) t is tentatively mapped to one existing metrical layer i. We consider two ways to estimate the distance between current peak t and a given metrical layer i: either by comparing current peak lag t with the actual lag of the peak associated with this metrical layer i at previous frame $n-1$:

$$d_1(t,i) = |t - t_i(n-1)| \tag{4}$$

or by comparing current peak lag t with the theoretical lag at that metrical layer i knowing the global tempo:

$$d_2(t,i) = |t - \tau_i(n)| \tag{5}$$

For low lag values, small difference in time domain can still lead to importance difference in tempo domain. For that reason, an additional distance is considered, based on tempo ratio:

$$d_3(t,i) = \left|\log_2\left(\frac{t}{\tau_i(n)}\right)\right| \tag{6}$$

The distance between current peak t and a given metrical layer i can be then considered as the minimum of the two distances on the time domain:

$$d(t,i) = \min(d_1(t,i), d_2(t,i)) \tag{7}$$

and the closest metrical layer i^* can be chosen as the one with minimal distance:

$$i^* = \arg\min_i d(t,i) \tag{8}$$

If this metrical period has already been assigned to a stronger peak in current frame n, this weaker peak t is discarded for any further analysis. In other cases, its integration to the metrical period i^* is carried out if it is close enough, both in time domain ($d(t,i)$) and in tempo domain ($d_3(t,i)$):

$$d(t,i) < \delta \text{ and } d_3(t,i) < \epsilon \tag{9}$$

In a second step, we check whether the periodicity peak triggers the addition of a new metrical layer in that metrical grid:

- For all the slower metrical layers i, we find those that have a theoretical period that is in integer ratio with the peak lag t:

  $$\min\left(\frac{\tau_i(n)}{t} \bmod 1, 1 - \left(\frac{\tau_i(n)}{t} \bmod 1\right)\right) < \epsilon \tag{10}$$

  where ϵ is set to to .02 if no other stronger peak in the current time frame n has been identified with the metrical grid, and else to .1 in the other case.

  If we find several of those slower periods in integer ratio, we select the fastest one, unless we find a slower one with a ratio defined in equation 10 that would be closer to 0.
- Similarly, for all the faster metrical layers, i we find those that have a theoretical pulse lag that is in integer ratio with the peak lag:

  $$\min\left(\frac{t}{\tau_i(n)} \bmod 1, 1 - \left(\frac{t}{\tau_i(n)} \bmod 1\right)\right) < \epsilon \tag{11}$$
- If we have found both a slower and a faster period, we select the one with stronger periodicity score.
- This metrical layer, of index i_R, will be used as reference onto which the new discovered metrical layer is based. The new metrical index i^* is defined as:

  $$i^* = i_R * \left[\frac{t}{\tau_i(n)}\right] \tag{12}$$

Finally, if the strongest periodicity peak in the given time frame n is strong enough (with periodicity score above a certain threshold θ) and is not associated with any period of the metrical grid(s) currently active, a new metrical grid is created, with a single metrical period (with $i = 1$) related to that peak.

All active metrical grids are tracked in parallel, by tentatively mapping the peaks of the periodicity curve on the periods of each grid.

A metrical grid stops being further extended whenever there in no peak in the given frame that can extend any of the dominant periods. Mechanisms have also been conceived to fuse multiple grids whenever it turns out that they belong to a single hierarchy.

The global tempo associated to the metrical grid is updated based on the actual lags measured along the different metrical periods in the current frame n. For each metrical period i and for the peak lag t_i associated to it, we obtain a particular estimation of the global lag (i.e., the lag at periodicity index 1), namely $\frac{t_i}{i}$. We can then obtain a global estimation of the global lag by averaging these tempo estimation at different periods, using as a weight the autocorrelation score s_i of those peaks:

$$\tau_1(n) = \frac{\sum_{i \in D} s_i \frac{t_i}{i}}{\sum_{i \in D} s_i} \qquad (13)$$

Not all metrical periods are considered, because there can be a very large number of those, and many of the higher periods are only redundant information that tend to be unreliable. For that reason, a selection of the most important—or *dominant*—metrical periods is performed, corresponding to the set D in previous equation. Each time a new metrical grid is initiated, the first metrical period ($i = 1$) is considered as dominant. Any other metrical period i becomes dominant whenever the last peak integrated is strong (i.e., with an autocorrelation score higher than a given threshold θ) and if the reference metrical period upon which layer i is based is also dominant.

The actual updating of the global tempo is somewhat more complex than the description given in the previous paragraph, because we consider the evolution of the tempo from the previous frame to the current frame, and limit the amplitude of the tempo change up to a certain threshold. This enables to add a certain kind of "inertia" to the model such that unrelated periodicities in the signal will not lead to sharp discontinuity in the tempo curves.

Values used for some parameters defined in this section: $\delta = .07$, $\epsilon = .2$, $\theta = .15$.

3.4 Metrical structure

The metrical grids constructed by the tracking method presented in the previous paragraph are so far made of a mere superposition of metrical periods. The ratio number associated with each metrical level should be considered relatively. For instance, the value 1 has no absolute meaning, it is arbitrarily given to the first level detected. Level 1.5 is 3 times slower than level .5. For each metrical grid, one or several of its metrical periods have been characterized as dominant because of their salience at particular instants of the temporal development of the metrical grid, and because such selection offers helpful guiding points throughout the temporal tracking of the metrical grid. Yet these selected dominant metrical periods simply highlight particular articulation of the surface and do not necessarily relate to the core metrical levels of the actual metrical structure.

A metrical structure is composed of a certain number of metrical *levels*: they are particular periods of the metrical grid that are multiple of each other. For instance, in a typical meter of time signature 4/4, the main metrical level is the quarter note, the upper levels are the half note and the whole note, the lower levels are the eighth note, the sixteenth note, and any other subdivision by 2 of these levels. In the same example, dotted half note (corresponding to three quarter notes) is related to one metrical period in the metrical grid, because it is explicitly represented in the autocorrelation function as a possible periodicity, but it is not considered as a metrical *level*.

In the graphical representations shown in Figures 2, 3 and 4, the metrical levels are shown in black while the other metrical layers are shown in gray.

The metrical structure offers core information about meter. In particular, tempo corresponds to beat periodicity at one particular metrical level. In a typical metre, the main metrical level could be used as the tempo reference. In our example, with a typical time signature 4/4, the tempo could be inferred by reporting the period at the metrical level corresponding to the quarter note. However, in practice, there can be ambiguity related to the actual metre, and especially related to the choice of the main metrical level.

For each metrical periodicity i can be associated a numerical score S_i, computed as a summation across frames of the related periodicity score $s_{i,n}$ for each frame n. The metrical periodicities i are progressively considered in decreasing order of score S_i as potential metrical levels.

In a first attempt, we integrate all possible periodicities as long as they form a coherent metrical structure. The metrical structure is initially made of one single metrical level corresponding to the strongest periodicity. Each remaining metrical period P, from strongest to weakest, is progressively compared with the metrical levels of the metrical structure, in order to check that for each metrical level L, P has a periodicity that is a multiple of L, or reversely. In such case, P is integrated into the metrical structure as a new metrical level.

This method may infer incorrect metrical structures in the presence of a strong accentuated metrical period that is not considered as a metrical level. This often happens in syncopated rhythm. For instance, a binary 4/4 metre with strong use of dotted quarter notes could lead to strongest periodicities at the eighth note (let's set this period to $i = 1$), dotted quarter note ($i = 3$) and whole note ($i = 8$). One example is the rhythmical pattern 123-123-12, 123-123-12, etc. In such case, if the periodicities related to dotted quarter note ($i = 3$) is stronger than the periodicities related to whole note ($i = 8$), the first method would consider the metre to be ternary, of the form 6/8 for instance.

In order to solve the limitation of the first method, a more elaborate method constructs all possible metrical structures, with metrical levels taken from the series of metrical periods from the input metrical grid. To each metrical structure is associated a score obtained by summing the score related to each selected level. The metrical structure with highest score is finally selected. In our example, alternative metrical structures are constructed, both for ternary rhythm—

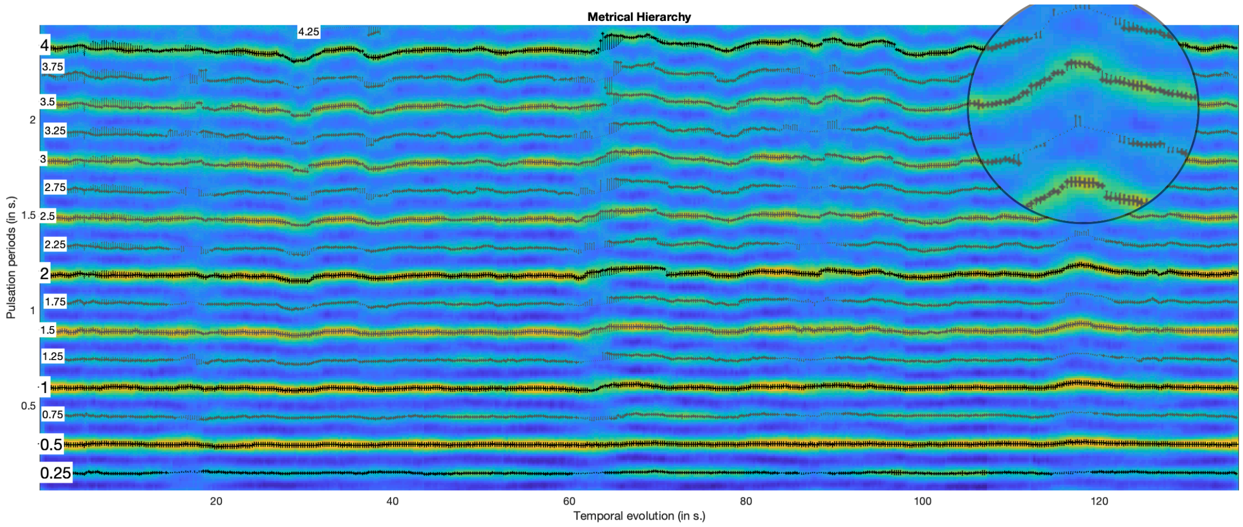

Figure 2. Autocorrelation-based periodogram with tracking of the metrical structure for the first 140 seconds of a performance of the first movement of J. S. Bach's *Brandenburg Concerto No. 2 in F major,* BWV 1047. Each metrical layer is indicated by a line of crosses extending from left to right, and preceded by a number indicating the index of the metrical layer. When the line is interrupted at particular temporal regions, the remaining dotted line represents the temporal tempo at that layer. Metrical levels are shown in black, while other metrical layers are shown in gray. See the text for further explanation.

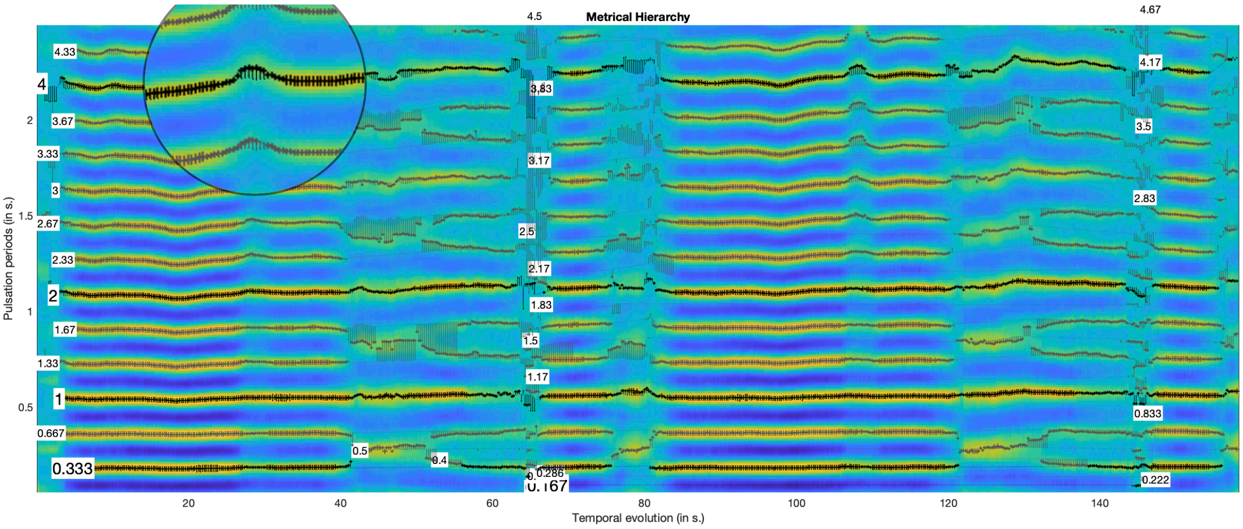

Figure 3. Autocorrelation-based periodogram with tracking of the metrical structure for the first 160 seconds of a performance of the Scherzo of L. van Beethoven's *Symphony No. 9 in D minor*, op.125, using the same graphical conventions as in Figure 2. As before, numbers, indicating metrical layer indices, are displayed where the metrical layers are first detected. For instance, layer 0.5, corresponding to the binary division of layer 1, appears at 40 seconds.

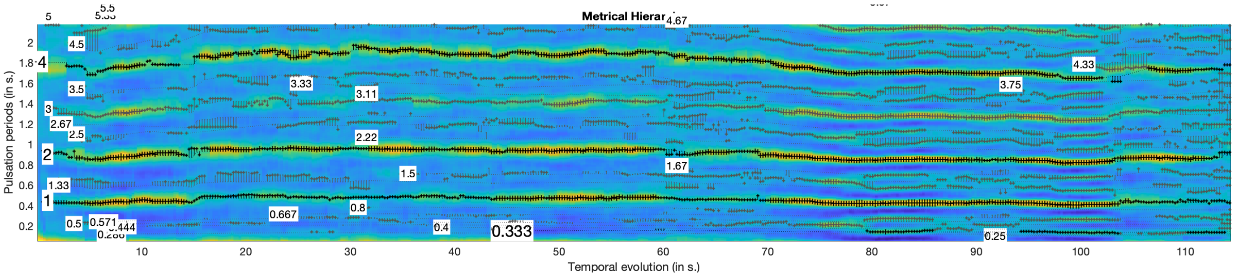

Figure 4. Autocorrelation-based periodogram with tracking of the metrical structure for the first 2 minutes of a performance of the *Allegro con fuoco* of A. Dvorak's *New World Symphony, Symphony No. 9 in E minor*, op. 95, B.178, using the same graphical conventions as in Figure 2.

with metrical levels $(1, 3, 6)$, or $(1, 3, 9)$, etc.—and for binary rhythm—$(1, 2, 8)$, $(1, 2, 4, 8)$, etc. If the periodicity corresponding to $i = 8$ is sufficiently strong, the binary rhythm will be chosen by the model. Although $i = 3$ is stronger than $i = 8$, the combination $(1, 2, 8)$, for instance, can be stronger than the combination $(1, 3, 6)$.

The resulting metrical structure is made of a combination of metrical levels, i.e., a subset $(i_1, i_2, \ldots)$ of the metrical periods of the metrical grid. One metrical level i_R needs to be selected as reference level for the computation of tempo. One simple strategy would consists in selecting the metrical level with highest score, as defined previously. However, those scores are based on purely signal processing method (namely, autocorrelation function), and do not take into account the fact that certain periodicities are more easily perceived than other. Studies have designed so-called "resonance curves" that enable to weight the periodicity score depending on the period, so that periods around typical range of periodicity around 120 BPMs would be preferred [9, 10]. We follow the same method, by weighting the metrical level scores S_{i_j} using the resonance curve proposed by [9], using as input to the resonance curve the median periodicity related to the given metrical level.

4. COMPARATIVE EVALUATION

The original algorithm was submitted to the Audio Tempo Extraction competition under the MIREX (Music Information Retrieval Evaluation eXchange) annual campaign [2] This evaluation is made using 160 30-second excerpts of pieces of music of highly diverse music genres but with constant tempo. Listeners were asked to tap to the beat for each excerpt. From this, a distribution of perceived tempo was generated [11]. The two highest peaks in the perceived tempo distribution for each excerpt were taken, along with their respective heights as the two tempo candidates for that particular excerpt. The height of a peak in the distribution is assumed to represent the perceptual salience of that tempo. Each algorithm participating to this MIREX task should also return two tempo candidates for each excerpt, with corresponding salience. This ground-truth data is then compared with the predicted tempo.

In 2013, our proposed model (OL) obtained the fourth [3] highest P-value, compared to models from 2006 to 2013, as shown in Table 1. It can be noted that these three better models are applicable only to music with stable tempo. Since then, OL has been surpassed by the two aforementioned deep-learning models [1, 2].

The current improved version of OL was submitted to the 2018 competition. The frequency resolution of the spectrogram is decreased without damaging the results. In order to filter out non-relevant peaks, the first peak at the lowest lag in the autocorrelation function is constrained to be preceded by a valley with negative autocorrelation. When comparing pairs of metrical hierarchies, only the most dominant levels of each hierarchy are selected in such a way that we compare hierarchies with same number of levels. Finally, a periodicity that is higher than 140 BPM cannot belong to the two selected metrical levels, except if that fast pulsation is ternary, i.e., if the pulsation at the next level is three times lower. OL 2018 does not offer any improvement in the results compared to the 2013 submission.

5. METRICAL DESCRIPTION

Tracking a large range of metrical levels enables to get a detailed description of the dynamic evolution of the metrical structure. For instance in Figure 3, the meter is initially and for the most part ternary. However between 40 and 50 s. (corresponding to bars 77 to 92), a little before 80 s. as well as between 120 and 130 s., we see that the ternary rhythm is actually perceived as a binary rhythm, as shown by the metrical level 0.5. Reversely in Figure 4, the meter is initially binary, but turns ternary after 80 seconds.

What is particularly interesting in those examples is also that the metrical structure changes, but the tempo remains somewhat constant. This shows that tempo is not a sufficient information for the description of metrical structure.

In order to give an indication of metrical activity that would not reduce solely on tempo but takes into consideration the activity on the various metrical levels, we introduce a new measure, called *dynamic metrical centroid*, which assesses metrical activity based on the computation of the centroid of the periods of a range of selected metrical levels, using their autocorrelation score as weight. The metrical centroid values are expressed in BPM, so that they can be compared with the tempo values also in BPM. High values for the metrical centroid indicate that more elementary metrical levels (i.e., very fast levels corresponding to very fast rhythmical values) predominate. Low values indicate on the contrary that higher metrical levels (i.e., slow pulsations corresponding to whole notes, bars, etc.) predominate. If one particular level is particularly dominant, the value of the metrical centroid naturally approaches the corresponding tempo value on that particular level.

Figure 5 shows the dynamic metrical centroid curve related to the *Allegro con fuoco* of A. Dvorak's *New World Symphony* as shown in Figure 4. The temporal evolution of the dynamic metrical centroid clearly reflects the change of rhythmical activity between the different metrical levels, and the transition between binary and ternary rhythm, which increases the overall perceived rhythmical speed.

6. DISCUSSION

The computational model OL was integrated into the version 1.6 of the open-source *Matlab* toolbox *MIRtoolbox* [12]. It also includes Goto's aforementioned accentuation curve algorithm [7]. The updated version of OL submitted to MIREX 2018 is integrated into version 1.8 of MIRtoolbox.

One main limitation of all current approaches in tempo estimation and beat tracking is that the search for periodicity is carried out on a *percussive* representation of the audio recording or the score, indicating bursts of energies or spectral discontinuities due to note attacks and changes.

[2] http://www.music-ir.org

[3] FW already had a model in 2013 that surpassed OL.

Contestant	SB	HS	EF	FW	GK	**OL**	AK	QH	NW	DP	ES	TL	GP	FK	CD	ZG	AD	SP	MD	DE	AP	PB	GT	CB	ZL	BD
Year 20..	15	18	13	15	11	**13**	06	14	10	06	10	10	12	12	13	11	06	11	14	06	06	06	10	13	18	14
Reference	[1]	[2]					[4]										[5]									
P-score	.90	.88	.86	.83	.83	**.82**	.81	.80	.79	.78	.77	.76	.75	.75	.74	.73	.72	.71	.69	.67	.67	.63	.62	.61	.60	.54
1 tempo	.99	.98	.94	.95	.94	**.92**	.94	.92	.91	.93	.91	.89	.86	.85	.91	.82	.89	.93	.85	.79	.84	.79	.69	.85	.68	.64
both tempi	.69	.66	.69	.57	.62	**.57**	.61	.56	.50	.46	.55	.48	.61	.62	.55	.57	.46	.39	.47	.43	.48	.51	.51	.26	.46	.38

Table 1. Comparison of MIREX results from all contestants of MIREX Audio Tempo Extraction from 2006 to 2018. For each author, only the model yielding best P-score is shown. The model presented in this paper is shown in bold.

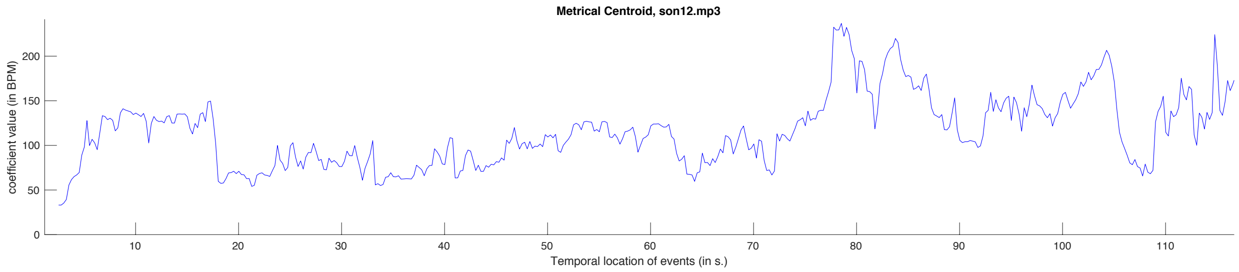

Figure 5. Dynamic metrical centroid curve for the same performance of the *Allegro con fuoco* of A. Dvorak's *New World Symphony* analysed in Figure 4.

Beyond this percussive dimension, other musical dimensions can contribute to rhythm. In particular, successive repetitions of patterns can be expressed in dimensions not necessarily conveyed percussively, such as pitch and harmony. This shows the necessity of developing methods for metrical analysis related not only to percussive regularities, but also to higher-level musicological aspects such as motivic patterns and harmonic regularities.

Acknowledgments

This work was partially supported by the Research Council of Norway through its Centres of Excellence scheme, project number 262762. This work was also partially supported by the Swiss Center for Affective Sciences. The selection of musical materials has benefitted from collaborations with Kim Eliard and Marc-André Rappaz.

7. REFERENCES

[1] S. Böck, F. Krebs, and G. Widmer, "Accurate tempo estimation based on recurrent neural networks and resonating comb filters," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[2] H. Schreiber and M. Müller, "A single-step approach to musical tempo estimation using a convolutional neural network," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[3] O. Lartillot, D. Cereghetti, K. Eliard, W. J. Trost, M.-A. Rappaz, and D. Grandjean, "Estimating tempo and metrical features by tracking the whole metrical hierarchy," in *3rd International Conference on Music and Emotion*, 2013.

[4] A. P. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 11, no. 6, pp. 803–816, 2006.

[5] M. Alonso, B. David, and G. Richard, "Tempo and beat estimation of musical signals," in *International Conference on Music Information Retrieval*, 2004.

[6] J. P. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the use of phase and energy for musical onset detection in complex domain," *IEEE Sig. Proc. Letters*, vol. 11, no. 6, 2004.

[7] M. Goto and Y. Muraoka, "Music understanding at the beat level – real-time beat tracking for audio signals,," in *IJCAI- 95 Workshop on Computational Auditory Scene Analysis*, 1996, pp. 68–75.

[8] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 558–601, 1998.

[9] P. Toiviainen and J. S. Snyder, "Tapping to bach: Resonance-based modeling of pulse," *Music Perception*, vol. 21, no. 1, pp. 43–80, 2003.

[10] L. V. Noorden and D. Moelants, "Resonance in the perception of musical pulse," *Journal of New Music Research*, vol. 28, no. 1, pp. 43–66, 1999.

[11] D. Moelants and M. McKinney, "Tempo perception and musical content: What makes a piece slow, fast, or temporally ambiguous?" in *International Conference on Music Perception and Cognition*, 2004.

[12] O. Lartillot and P. Toiviainen, "Mir in matlab (ii): A toolbox for musical feature extraction from audio," in *International Conference on Music Information Retrieval*, 2007.