TIME VARIABLE TEMPO DETECTION AND BEAT MARKING

Geoffroy Peeters
IRCAM - Analysis/Synthesis Team
1, pl. Igor Stravinsky - 75004 Paris - France
peeters@ircam.fr

ABSTRACT

In this paper, we present a novel approach to the automatic estimation of tempo over time. This method aims at detecting tempo for music with and without percussion. An onset-energy function is first proposed based on the reassigned spectral energy flux. A combination of Discrete Fourier Transform and Frequency Mapped AutoCorrelation Function is used to estimate the dominant periodicities at each time. A Viterbi algorithm is then used in order to detect the most likely tempo and meter/beat subdivision paths over time. The performance of the proposed method is evaluated on three databases.

1. INTRODUCTION

Tempo and beat are very important in the perception of (western) music (a time structured set of sound events). They carry important information that can be used in many applications: query by tempo, processing using tempo information (beat synchronous mixing, beat slicing, segmentation into beat units), musical analysis (interpretation) or more generally sound analysis. For this reason, tempo/beat estimation have been the subject of an increasing number of contributions in the last few years. However, depending on the music genre considered (especially classical and jazz music), their automatic estimation can be very difficult. Tempo/beat estimation methods can be roughly separated into two different approaches: - those that use the signal energy along time (possibly separated through a bank of filters) to measure the periodicity of the signal [14]; - those that detect onsets in the signal and derive from their inter-distances (Inter-Onset-Interval Histogram) the most common periodicity [8] [3]. See [7] for a complete review of tempo/beat detection methods.

1.1. System overview

The system we propose in this paper is a tempo tracker system. It has been designed in order to allow tracking of fast variations of tempo as well as detection for music with and without percussion (classical music). Variations of tempo and changes of meter over time are especially useful in the case of jazz and classical music. The system is non-causal since the tracking of the tempo is done using a non-causal algorithm (Viterbi algorithm).

In the following sections, we detail the various stages of the system. The system relies on a standard schema (see Figure 1). An onset-energy function is first extracted using a proposed reassigned spectral energy flux (part 2). This onset-energy function is then used to estimate the dominant periodicities at each time. This is done using a method based on a combination of DFT and Frequency-Mapped ACF (part 3.1). Finally the most likely tempo and meter/beat subdivision paths over time are estimated using a Viterbi decoding algorithm (part 3.2). We evaluate the performance of the proposed system in part 4.

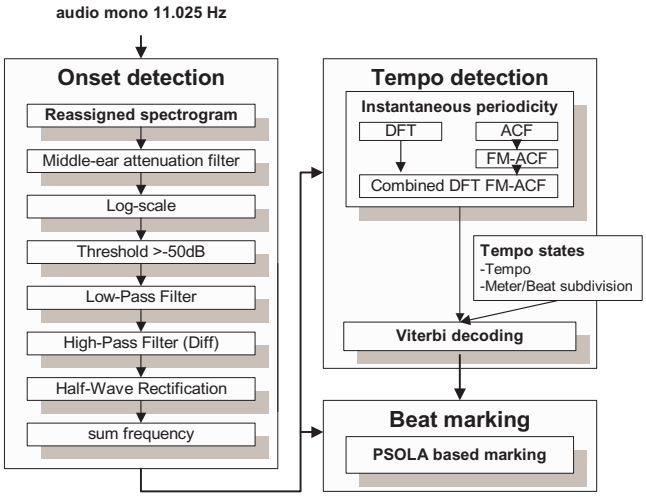

Figure 1. Flowchart of the tempo detection system

2. ONSET-ENERGY FUNCTION

In order to detect the tempo of a piece of music we need to observe its signal through something meaningful in terms of musical periodicity. Most methods use the variation of the signal energy or its variations inside several frequency bands [14]. Since our interest is not only in music with percussion but also in music without percussion, our energy function should also react to any musically meaningful variations (note transitions at constant global energy, slow attacks, ...). These variations are usually visible in a spectrogram representation. [11] proposes a method, called the spectral energy flux, based on the measure of the spectrogram variations over time. Instead, we propose the use of the reassigned spectrogram (see part 2.1) which allows to significantly improve temporal and frequency resolution, therefore avoiding attack blurring, and allows to better differentiate very close pitches.

2.1. Reassigned Spectrogram

The reassigned spectrogram [4] consists in reallocating the energy of the "bins" of the spectrogram (bin=frequency ω_k of the STFT at time t_i of the frame) to the frequency ω_r and time t_r corresponding to their center of gravity. The reassignment of the frequencies is based on the computation of the instantaneous frequency (time derivative of the phase) and can be efficiently computed by:

$$\omega_r(x, t_i, \omega_k) = \omega_k - \Im\left\{\frac{STFT_{dh}(x, t_i, \omega_k)}{STFT_h(x, t_i, \omega_k)}\right\} \quad (1)$$

where $\Im$ stands for the imaginary part, h stands for the analysis window and dh stands for the time derivative of the window: $\partial h(t)/\partial t$. The reassignment of the times is based on the computation of the group delay (frequency derivative of the phase spectrum) and can be efficiently computed by:

$$t_r(x, t_i, \omega_k) = t_i + \Re\left\{\frac{STFT_{th}(x, t_i, \omega_k)}{STFT_h(x; t_i, \omega_k)}\right\} \quad (2)$$

where $\Re$ stands for the real part, h stands for the analysis window and th stands for the frequency derivative of the window ($th = t \cdot h(t)$). Each bin (ω_k, t_i) of the spectrogram is then reassigned to its center of gravity (ω_r, t_r). The bins are accumulated in the time and frequency plane.

2.2. Reassigned Spectral Energy Flux

The signal is first down-sampled to 11.025 Hz and converted to mono (mixing both channels). • The reassigned spectrogram $X(\omega_k, t_i)$ is computed using a hamming window. Depending on the music being considered, a different frequency resolution is required: long window (0.0928 s.) for music without percussion, short window (0.0464 s.) for music with percussion [1] . The hop size is set to 0.0058 s. • In order to simulate the attenuation due to the human middle ear, a filter [12] reinforcing the mid-range frequency is applied to each spectral frame. • As in [10], the energy spectrum is converted to the log scale in order to work on relative variations of energy. • A threshold of -50 dB the maximum energy is applied. • The energy inside each frequency band $e(\omega_k, t_i)$ is low-pass filtered (with an elliptic filter of order 5 and a fc of 10 Hz) and differentiated using a simple $[1, -1]$ differentiator. • $e(\omega_k, t_i)$ is then Half-Wave Rectified. • For a specific time t_i, the sum over all frequency ω_k is computed: $e(t_i) = \sum_k e(\omega_k, t_i)$. The resulting energy function $e(n = t_i)$ has a sampling rate of 172 Hz [2] .

In Figure 2, we compare the resulting onset-energy functions using normal spectral energy flux [top] and reassigned spectral energy flux [bottom]. The analysis parameters are the same for both. Some onsets (around time 4 s., 4.3 s., 5.2 s., 5.6 s.) are missing in the spectral energy flux. This is due to the blurring that occurs in the normal spectrogram but not in the reassigned spectrogram.

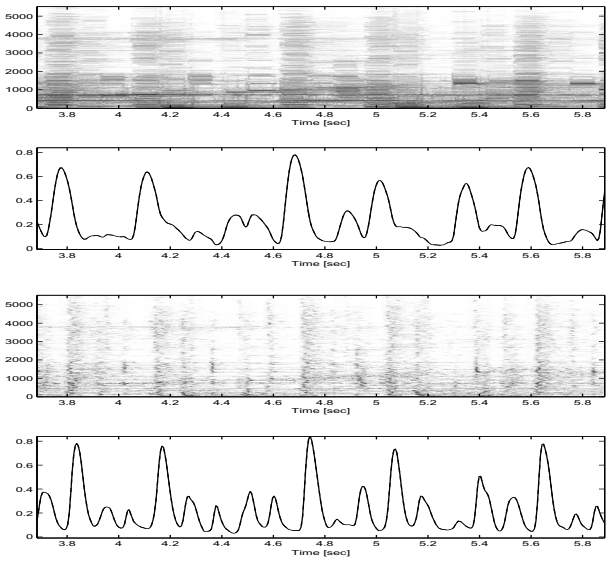

Figure 2. [top] Spectrogram and corresponding spectral flux function [bottom] Reassigned spectrogram and corresponding reassigned spectral energy flux function. [Signal: Carlinhos Brown "Pandeiro Deiro" from ISMIR04 test database]

3. TEMPO DETECTION

From the signal observation $e(n)$ we estimate the tempo. The algorithm we propose works in two stages: - the first estimates the dominant periodicities around a specific time (part 3.1); - the second estimates the tempo and meter/beat subdivision paths that best explain the observed periodicities along time (part 3.2).

3.1. Periodicity estimation: combined DFT and Frequency Mapped ACF

Periodicity estimation of a signal is often done using Discrete Fourier Transform (DFT) or AutoCorrelation Function (ACF). Since $e(n)$ is a periodic signal that can be roughly modeled as a Dirac comb convolved with a LP envelope, the outcome of its DFT is a set of harmonically related frequencies. Depending on their relative amplitude it can be difficult to decide which harmonic corresponds to the tempo frequency. The outcome of its ACF is a set of periodically related lags. Here also it can be difficult to decide which period corresponds to the tempo lag. Algorithms like the Two-Way Mismatch [8] or maximum likelihood try to solve this problem. Because the octave uncertainty of the DFT and ACF occur in inverse domain (frequency domain for the DFT, lag domain or inverse frequency domain for the ACF), we use this property to construct a combined function that reduces these uncertainties. We first make $e(n)$ a zero-mean, unit-variance signal. $e(n)$ is analyzed both by

DFT: We note $F(\omega_k, t_i)$ the amplitude spectrum of $e(n)$ for a frequency ω_k and a frame centered around time t_i (a hamming window is used of length equal to 6 sec, the hop size is set to 0.5 s.).

Frequency Mapped ACF (FM-ACF): We note $A(l, t_i)$ the normalized (in energy and in maximum value) AutoCorrelation Function of $e(n)$ for a lag l and a frame centered around time t_i. The value at lag l of the ACF represents the amount of periodicity at the lag l/sr (where sr is the sampling rate) or at the frequency $\omega_l = sr/l \,\forall l > 0$. Each lag l is therefore "mapped" in the frequency domain. Of course since $A(l, t_i)$ has a constant resolution in time $A(\omega_l, t_i)$ has a decreasing resolution in frequency. In order to get the same linearly spaced frequencies ω_k as for the DFT, we interpolate $A(l, t_i)$ and sample it at the lags $l = sr/\omega_k$. [3] . Finally, Half-Wave Rectification is applied to $A(\omega_k, t_i)$ in order to consider only positive correlation.

Combined function: We now have two measures (the DFT and the FM-ACF) of periodicity at the same frequencies ω_k. We finally compute a combined function $Y(\omega_k, t_i)$ by multiplying the DFT and the FM-ACF at each frequency ω_k: $Y(\omega_k, t_i) = F(\omega_k, t_i) \cdot A(\omega_k, t_i)$.

[1] In the rest of the paper, we will use a long window.

[2] Notes that one could easily derive the onset positions by applying a threshold on $e(n)$. However we found that, for the task of tempo detection, working directly on $e(n)$ is more robust than working on the onsets because it avoids the consequences of false/missed onset detections.

[3] Notes that this doesn't improve the frequency resolution of A.

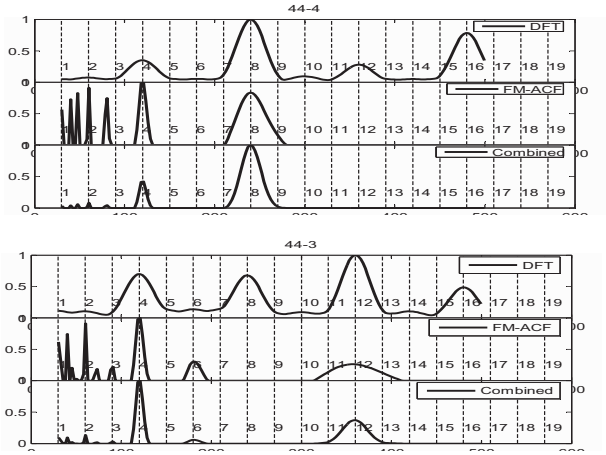

Figure 3. Comparison of DFT, FM-ACF and Combined function for [top three] simple meter in 4/4 [bottom three] compound meter in 4/4

Advantages of the Combined DFT and FM-ACF: In Figure 3 we illustrate the use of the DFT and FM-ACF functions for two characteristic signals: a simple meter in 4/4 (each beat is divided into halves), a compound meter in 4/4 (each beat is divided into thirds). We note 4 the quarter note (tempo), 8 the eighth note, 12 the eighth note triplet, 16 the sixteenth note, ... In the case of a *simple meter* [top], the DFT contains all the harmonics of the tempo: 4, 8, 12, 16, ... despite the fact that the shortest interval is the 8. The FM-ACF contains all the sub-harmonics of the shortest interval: 8, 8/2=4, 8/3, ... Combining both functions allows keeping only the common frequencies: one at the tempo (4) and one at the shortest interval (8). In the case of a *compound meter* [bottom], the DFT contains all the harmonics of the tempo: 4, 8, 12, 16, ... despite the fact that the shortest interval is the 12. The FM-ACF contains all the sub-harmonics of the shortest interval: 12, 12/2=6, 12/3=4, ... Combining both functions allows keeping only the common frequencies: one at the tempo (4) and one at the shortest interval (12). The same can be shown in the case of simple and compound meter in 3/4. The combined use of the DFT and the FM-ACF allows removing most ambiguities (like simple/compound or duple/triple meter confusion) from the spectrum.

3.2. Tempo estimation: Viterbi decoding of "tempo states"

We estimate the dominant periodicities $Y(\omega_k, t_i)$ at each time t_i . We then look for the temporal path of tempo that best explains the observed periodicities over time. Our method shares some similarities with [11] and [6], but in our case the observed periodicities do not only depend on the tempo frequency but also on the meter characteristics. We consider three different *meter/beat subdivision templates (mbst)*: the duple/ simple (noted (2-2)), the duple/ compound ((2-3), example is 6/8 meter) and the triple/ simple ((3-2), example is 3/4 meter). We define a "tempo state" as a specific combination of a tempo frequency b_i and a $mbst$ m_j: $S(i,j) = [b_i, m_j]$ with $i \in I$ the set of considered tempo and $j \in \{1,2,3\}$ the three considered $mbst$. We look for the most likely temporal succession of "tempo state" given our observation. We formulate this problem as a Viterbi decoding algorithm [2].

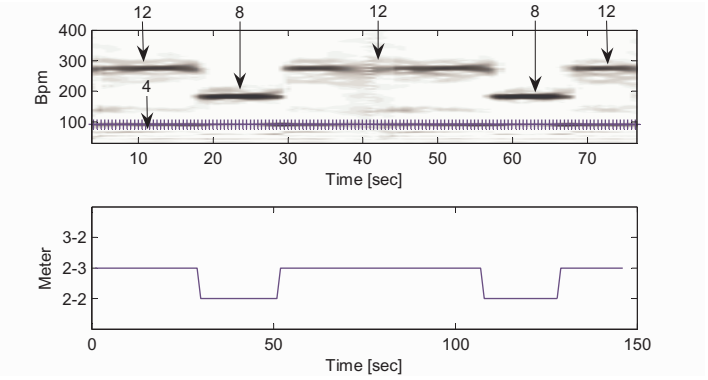

Figure 4. [top] Tempo tracking [bottom] Meter/beat subdivision tracking [Signal: "Standard Of Excellence - accompaniment CD - Book2 - All inst. - 88. Looby Loo"]

Observation probability: $p_{obs}(b_i, m_j) = p_{prior}(b_i, m_j) \cdot p([b_i, m_j]|Y(\omega_k, t_i))$.

• $p_{prior}(b_i, m_j)$ is the prior probability to observe a specific bpm i and a specific $mbst$ j. The goal is to favor the detection of tempo in the range 50-150 bpm but we didn't want to favor any $mbst$ in particular. We set it as a gaussian pdf $p_{prior}(b_i, m_j) = p_{prior}(b_i) = N_{\mu=100,\sigma=150}(b_i)$.

• $p([b_i, m_j]|Y_n(\omega_k, t_i))$ is the probability to observe a specific tempo i and $mbst$ j given our observations $Y_n(x)$. This probability is computed using the following equations (we set $\alpha = \beta$ empirically to 0.5):

- $p([x,(2-2)]|Y_n) = \frac{(\alpha Y_n(x/2)+Y_n(x)+\beta Y_n(2x))}{\sum_y Y_n(y)}$

- $p([x,(2-3)]|Y_n) = \frac{(\alpha Y_n(x/2)+Y_n(x)+\beta Y_n(3x))}{\sum_y Y_n(y)}$

- $p([x,(3-2)]|Y_n) = \frac{(\alpha Y_n(x/3)+Y_n(x)+\beta Y_n(2x))}{\sum_y Y_n(y)}$

Transition probability: $p_{trans}([b_i, m_j][b_k, m_l]) = p_{trans}([b_i],[b_k]) \cdot p_{trans}([m_j],[m_l])$.

• The goal of the first probability is to favor continuous tempo. We set it as a gaussian pdf $N_{\mu=b_i,\sigma=5}(b_k)$.

• The goal of the second probability is to avoid $mbst$ jumps from frame to frame. We set it empirically to 0.1. Finally a standard Viterbi decoding algorithm gives us the best path along time through the states $[b_i, m_j]$ therefore gives us simultaneously the best tempo and the best $mbst$ that explain $Y_n(\omega_k, t_i)$.

Tempo tracking is illustrated into Figure 4. The top of the figure represents the estimated tempo track along time (+) superimposed to the periodicity observation $Y_n(\omega_k, t_i)$ represented as a matrix and annotated by hand (4,8,12). The bottom part represents the estimated $mbst$ along time. The tempo remains constant during all the track duration but depending on the local periodicities (4-12 or 4-8), the $mbst$ is estimated as either (2-3) or (2-2). The tempo and $mbst$ estimations are correct.

4. EVALUATION

Data: The proposed algorithm was evaluated on three databases: - the *"ballroom-dancer" database* (698 tracks of 30 s. long) used for the ISMIR2004 "tempo induction contest" [9] - the *RWC database* [5] which contains many examples of interesting music genre for tempo recognition

(like classical and jazz music) [4] , from which we have extracted the segment from 30 s. to 50 s and annotated it by hand. - a private database consisting mainly of *pop-rock* 'hits' (158 extracts of 20 s. long).

Evaluation method: The tempo and beat markers [5] were extracted automatically over time (the tempo was not considered constant). For each track, we have checked if the tempo was correct during 75% of its duration. We have not applied the systematic "1/2, 2, 1/3, 3" tempo tolerance applied for the ISMIR 2004 "tempo induction" contest. Instead we have considered the following tolerances for the estimated tempo: a) 1/2 or 2 if (2-2), b) 1/3 or 2 if (3-2), c) 1/2 or 3 if (2-3). The results are indicated in Figure 5 for the cases without tolerance (accuracy 1) and with tolerance (accuracy 2).

Results: For the *ballroom* database, the global tempo recognition rates are 63%/92% (accuracy 1/accuracy 2) which is close to the best results obtained during ISMIR 2004. For accuracy 1 (octave errors), most errors occurred in the Jive, Quickstep (fast tempo), Rumba and Waltz (the algorithm follows mainly the tatum). For accuracy 2, most errors occurred in the Waltz category (78%) (because of bad $mbst$ estimation and because onset are difficult to detect in slow chord transitions). For the *RWC* database, the global recognition rate is 79%. For Classical Music (70%), the errors are mainly due to bad onset detections (slow chord transitions) and fuzzy tempo (the tempo is difficult to detect manually). For Jazz Music (83%), the errors are mainly due to bad $mbst$ estimation. The recognition rate for Music Genre (including pop, rock, flamenco, and Indian music) is higher: 85%. For the *pop-rock* 'hits' database, the recognition rates are high: 79% / 98%.

As a general remark, tempo errors mainly occurs because of three reasons: - onsets are difficult to detect in the audio signal (music with slow chord transitions), - bad estimation of $mbst$ (2-2 is often confused with 2-3 in the presence of accentuated dotted-quarter note) - complex rhythm not adequately represented by the $mbst$ (Jazz music). Octave errors often occur when the rhythm is few emphasized; the algorithm then mainly follows the tatum. Examples of beat marking results are available at http://recherche.ircam.fr/anasyn/peeters/tempo/.

5. CONCLUSION

In this paper, we have proposed a method for the automatic estimation of the tempo based on the reassigned spectral energy flux, a combination of DFT and Frequency Mapped-ACF and a Viterbi decoding algorithm. The proposed method has been evaluated on three databases. The recognition rate is high for popular music and for most ballroom music. For the jazz music, when the algorithm fails it was mainly because of the complexity of the rhythm, which is not adequately represented by our three templates. The templates should therefore be extended in the future. In the case of classical music, when the algorithm fails it was mainly because the concept of note onset was unclear. In this case, another kind of signal's observation could be used.

[4] Note that we haven't used the "Traditional Japanese" and "Vocal" tracks of this database.

[5] The beat markers were positioned using a method we previously developed for PSOLA analysis [13]. Two constraints are taken into account: 1) two marks must be separated by the local tempo period 2) the location of the marks must be close to the local maximum of the energy function. The best solution is found using a least-square algorithm.

Ballroom database

	ChaChaCha	Jive	Quick Step	Rumba	Samba	Tango	Viennese Waltz	Waltz	**Total**
# items	111	60	82	98	86	86	65	110	698
Accuracy 1	98%	53%	37%	54%	68%	95%	85%	44%	**63%**
Accuracy 2	100%	98%	91%	94%	93%	94%	91%	78%	**92%**

RWC database

	Classique	Jazz	Music Genre	**Total**
# items	78	59	61	**182**
Accuracy 1				
Accuracy 2	70%	83%	85%	**79%**

Poprock database

	Total
# item	158
Accura	**79%**
Accura	**98%**

Figure 5. Results of the tempo detection evaluation

Acknowledgments

Part of this work was conducted in the context of the European I.S.T. project Semantic HIFI [15] (http://shf.ircam.fr).

6. REFERENCES

[1] M. Alonso, B. David, and G. Richard. Tempo and beat estimation of musical signals. In *ISMIR*, 2004.

[2] R. Bellman. *Dynamic Programming*. Princeton University Press, Boston, 1957.

[3] S. Dixon. Automatic extraction of tempo and beat from expressive perf. *JNMR*, 30(1):39–58, 2001.

[4] P. Flandrin. *Time-Frequency/Time-Scale Analysis*. Academic Press, San Diego, California, 1999.

[5] M. Goto. Rwc (real world computing) music database, 2005.

[6] M. Goto and Y. Muraoka. Music understanding at the beat level real-time beat tracking for audio signals. In *IJCAI*, pages 68–75, 1995.

[7] F. Gouyon and S. Dixon. A review of rhythm description systems. *CMJ*, 29(1), 2005.

[8] F. Gouyon, et al. Pulse-dependent analyses of percussive music. In *AES 22nd*, 2002.

[9] F. Gouyon, et al. An experimental comparison of audio tempo induction algorithms. In *ISMIR*, 2005.

[10] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *ICASSP*, 1999.

[11] J. Laroche. Efficient tempo and beat tracking in audio recordings. *JAES*, 51(4):226–233, 2003.

[12] Moore, et al. A model for the prediction of thresholds loudness and partial loudness. *JAES*, 45:224–240, 1997.

[13] G. Peeters. *Modeles et modelisation du signal sonore adaptes a ses caracteristiques locales*. Phd thesis, Universite Paris VI, 2001.

[14] E. Scheirer. Tempo and beat analysis of acoustic musical signals. *JASA*, 103(1):588–601, 1998.

[15] H. Vinet. The semantic hifi project. In *ICMC*, Barcelona, Spain, 2005.