Simultaneous Beat and Downbeat-Tracking Using a Probabilistic Framework: Theory and Large-Scale Evaluation

Geoffroy Peeters and Helene Papadopoulos

***Abstract*—This paper deals with the simultaneous estimation of beat and downbeat location in an audio-file. We propose a probabilistic framework in which the time of the beats and their associated beat-position-inside-a-bar roles, hence the downbeats, are considered as hidden states and are estimated simultaneously using signal observations. For this, we propose a "reverse" Viterbi algorithm which decodes hidden states over beat-numbers. A beat-template is used to derive the beat observation probabilities. For this task, we propose the use of a machine-learning method, the Linear Discriminant Analysis, to estimate the most discriminative beat-templates. We propose two functions to derive the beat-position-inside-a-bar observation probability: the variation over time of chroma vectors and the spectral balance. We then perform a large-scale evaluation of beat and downbeat-tracking using six test-sets. In this, we study the influence of the various parameters of our method, compare this method to our previous beat and downbeat-tracking algorithms, and compare our results to state-of-the-art results on two test-sets for which results have been published. We finally discuss the results obtained by our system in the MIREX-09 and MIREX-10 contests for which our system ranked among the first for the "McKinney Collection" test-set.**

***Index Terms*—Beat-tracking, downbeat-tracking, beat-templates, Linear Discriminant Analysis, hidden Markov model, reverse Viterbi decoding.**

I. Introduction

Beat-tracking and downbeat-tracking are still today among the most challenging subjects in the music-audio research community. This is due to the complexity of the task. While tempo estimation is mainly a problem of periodicity estimation (with the inherent octave ambiguities), beat-tracking is both a problem of periodicity estimation and a problem of location of the beginning of the period inside the signal (with the inherent ambiguities of the rhythm itself). Downbeat location is mainly a perceptual notion arising from the music construction process [1]. Considering that the best results obtained in the last Audio Beat Tracking contests (MIREX-09 and MIREX-10) are far from being perfect, this problem is far from being solved. If many beat-tracking algorithms achieve good results for rock, pop or dance music tracks (except for highly compressed tracks), this is not the case when considering classical, jazz, world music or recent Western mainstream music styles such as Drum'n'Bass or R'n'B (which use complex rhythms).

Improving the performances of beat and downbeat-tracking is important since those are used in many applications today: beat/ downbeat-synchronous analysis (such as for score alignment or for cover-version identification), beat/ downbeat-synchronous processing (time-stretching, beat-shuffling, beat-slicing), music analysis (beat taken as a prior for pitch estimation, for onset detection or chord estimation) or visualization (time-grid in audio sequencers).

In this introduction, we review related works in beat and downbeat-tracking, we review our previous beat and downbeat-tracking algorithms, we present our new algorithm and compare it to existing works. In the next parts, we detail each part of our new algorithm and perform a large-scale evaluation.

A. Related works

Related works in beat-tracking: This paper deals with beat-tracking from audio signal. We consider tempo period and meter as input parameters of our system and deal with audio data. Numerous good overviews exist in the field of tempo estimation or beat-tracking from symbolic data (see for example [2] [3]). In the following, we only review existing approaches related to beat-tracking from audio signal.

Methods can be roughly classified according to the front-end of the system. Two types of front-end can be used: - *discrete onset* representation extracted from the audio signal (Goto [4] [5], Dixon [6]), or - *continuous-valued* onset function (Scheirer [7], Klapuri [8], Davies [9]).

Methods can also be classified according to the model used for the tracking. Goto [10] and Dixon [6] use a *multi-agents* model. Each agent propagates an assumption of beat-period and beat-phase; a "manager" then determines the best agent. Scheirer [7] and Jehan [11] use *resonating comb-filters* which states provide directly the phase hence the beat information. Klapuri [8] extends this method by using this states as input to a hidden Markov model tracking phase evolution. *Probabilistic formulations* of the beat-tracking problem are also proposed. For example Cemgil [12] proposes a Bayesian framework for symbolic data. This framework is adapted and extended to the audio case by Hainsworth [13]. Laroche [14] proposes the use of dynamic programming to estimate simultaneously beat-period and beat-phase. Dynamic programming is also used by Ellis [15] to estimate beat-phase given tempo as input. Mixed approaches are also proposed. For example Davies [9] mixes a comb-filterbank approach with a multi-agents approach (he uses two agents representing a General State and Context-Dependent State). Most algorithms relying on *histogram methods* for beat-period estimation use a different

G. Peeters and H. Papadopoulos are with the Sound Analysis/Synthesis Team of IRCAM-CNRS STMS, 75004 Paris, France (e-mail: geoffroy.peeters@ircam.fr; helene.papadopoulos@ircam.fr).

algorithm for beat-phase estimation (Seppanen [16], Gouyon [17]). This is because the histogram does not provide phase information. However, recent approaches succeed to use directly the *phase information* to derive beat-phase (Autocorrelation Phase Matrix of Eck [18], mid-level representation of Grosche [19]).

Finally, we can classify the methods according to the way a beat likelihood is associated to a time. Existing algorithms use either - *directly* the values of the discrete onsets (or of the continuous onset function) at the specific time, or - compute a cross-correlation between the local discrete onset sequence (or local continuous onset function) and a *beat-template* representing the theoretical pulses corresponding to the local tempo.

For a long time, the performances of the various approaches have been difficult to compare because authors were using different test-sets and different evaluation rules. Only recently, common test-sets (such as the ones used in [8] and [13]) and evaluation rules (such as the ones collected by [20]) have allowed this comparison. Also, the IMIRSEL team has provided MIREX evaluation frameworks for audio beat-tracking in 2005 [21], 2006 [22], 2009 [23] and 2010 [24] through the MIREX contests. Among the top-ranked participants to these contests are (in alphabetical order): Alonso, Davies, Dixon, Ellis, Eck, Gouyon, Klapuri, Uhle.

Related works in downbeat-tracking: Most of the proposed approaches for downbeat-tracking rely on prior knowledge (such as tempo, time-signature of the piece or hand-annotated beat positions). The system of Allan [25] relies on the assumption that a piece of music contains repeated patterns. He proposes a model that uses autocorrelation to estimate the downbeat locations given beat-positions. This model has been tested on 42 different pieces of music from various genres and achieves a recognition rate of 81% for pieces in 4/4 meter (more testing are needed for pieces on 3/4 meter). The model of Jehan [26] is tempo independent, does not require beat-tracking but requires prior knowledge obtained through listening or learning during a supervised training stage where downbeat locations are annotated. His model has only been applied to music in 4/4 meter. Goto [27] proposes two approaches for downbeat estimation. For percussive music, the downbeats are estimated using rhythmic pattern information. For non-percussive music, the downbeats are estimated using chord change information. Klapuri [8] proposes a full analysis of musical meter into three different metrical levels: tatum, tactus and bar level. The downbeats are identified by matching rhythmic pattern templates to a mid-level representation. Ellis [28] uses a similar "template-based" approach in a drum-pattern classification task. Davies [29] proposes an approach based on spectral difference between band-limited beat-synchronous analysis frames. The sequence of beat positions of the input signal is required and the time-signature is to be known a priori. Gainza [30] proposes a method that segments the audio according to the position of the bar lines. The position of each bar line is predicted by using prior information about the position of previous bar lines as well as the estimated bar length. The model does not depend on the presence of percussive instruments and allows moderate tempo deviations. We also mention the approach proposed by Gouyon [31]. While [31] does not deal with downbeat location, he proposes an approach for the estimation of the meter of an audio signal based on low-level signal features (such as energy, spectral flatness and various energy ratios). Their temporal sequences are then used for meter classification. Using a test-set of 70 sounds, the authors report 95% correct recognition.

B. Presentation of our previous system

1) Tempo/ meter estimation algorithm: This paper concerns the beat and downbeat-tracking problem. For this, we consider as input parameters an onset-energy-function $f(t)$, time-variable tempo $bpm(t) = 60/\text{Tb}(t)$ (where Tb is the length in second of a beat period) and meter (2/4, 3/4 or 6/8). The onset-energy-function has a sampling rate of 172Hz. It is computed using a reassigned-spectral-energy-flux function (RSEF). The system used for the estimation of these input parameters is the one described in [32]. This system has been positively estimated in [32] and in the MIREX-05 contest [21] for tempo estimation[1].

2) Previous beat-tracking algorithm: Our previous beat-tracking algorithm was inspired from a P-sola analysis method for locating the Glottal Closure Instants (GCIs) [33]. This method proceeds in two separated stages. The first stage locates a set of local maxima of $f(t)$ with an inter-distance close to the local estimated tempo period $\text{Tb}(t)$. The second stage performs a least-square optimization in order to satisfy simultaneously two constraints[2]: c-a) "markers close to the local maxima", c-b) "inter-distance between markers close to $\text{Tb}(t)$". We refer the reader to [34] for more details on this method, which we call P-sola in the following.

3) Previous downbeat-tracking algorithm: Our previous downbeat-tracking algorithm was based on a chord-detection algorithm [35]. This algorithm takes as input the location of the beat-markers and computes for each beat a chroma vector using a Constant-Q transform. The chord succession is then obtained using a hidden Markov model given the observed chroma, chord emission and chord transition probabilities. The downbeats are estimated using the assumption that chords are more likely to change on the downbeat positions.

C. Paper contribution and organization

In this paper, we propose a probabilistic framework for the simultaneous estimation of beat and downbeat location given estimated tempo and meter as input.

In part II, we present the probabilistic framework using a hidden Markov model formulation in which beat-times and their associated beat-position-inside-a-bar (bpib) are the hidden states. We give the big picture in II-A, present the HMM formulation in part II-B, the specific reverse Viterbi decoding algorithm in part II-C and the used formulation of the probabilities in part II-D.

[1]In MIREX-05, our tempo evaluation system ranked first with 95.71% in the category "At Least One Tempo Correct".

[2]It should be noted that these two constraints have been also used by Ellis in [15].

We then explain the estimation of the emission probabilities (par III) and of the transition probabilities (part IV). The emission probabilities are estimated using a beat observation probability and two bpib observation probabilities. In part III-A, we propose the use of a machine learning approach to find the "best" beat-templates in order to estimate the beat observation probability. In this, "best" is defined as "such as to maximize the discrimination of the correlation values obtained at the beat and non-beat positions". In part III-B, we propose the use of two functions in order to estimate the bpib observation probabilities: - the first is based on the analysis of chroma vector variation over time (part III-B1) - the second is based on the analysis of the spectral balance (part III-B2). In part IV, we present the transition probabilities which take into account the fact that hidden states represent beats in specific beat-position-inside-a-bar.

Finally in part V, we propose a large-scale evaluation of beat and downbeat tracking using six different test-sets. We compare our results to state-of-the-art results and discuss the results obtained by our algorithm.

D. Comparison to related works

Our algorithm works with a continuous onset-function rather than a series of discrete onsets. The method used to associate a beat likelihood to a time is a beat-template method. We propose a method to train the most discriminative beat-templates using Linear Discriminant Analysis (LDA). This is an important contribution of this paper. As we will see, the use of LDA-trained beat-templates allows improving the results over the ones obtained with simpler beat-templates (such as the beat-template representing the theoretical pulses corresponding to the local tempo [36]).

The simultaneous estimation of beat and downbeat is then formulated as a hidden Markov model in which hidden states are the beat-times and their associated beat-position-inside-a-bar. The concept of beat-position-inside-a-bar and the use of it to derive the downbeat is inspired by the authors previous works [35] [37]. The use of a probabilistic formulation has some links with the Bayesian framework of Cemgil [12] and Hainsworth [13] but the formulation is here very different and used to perform simultaneous beat and downbeat-tracking. The formulation of hidden-states as beat-times can be linked with Laroche [36] and Ellis [15] dynamic programming approaches (especially concerning the decoding algorithm). However, in the present work, we provide a probabilistic formulation using a hidden Markov model which allows the extension of the hidden states to the downbeat estimation problem. It should be noted that our use of hidden Markov model is not related to the way Klapuri [8] uses it. In [8], two independent hidden Markov models, which hidden states represent phase evolution, are used to track separately beat and downbeat phase.

In our system two observation probabilities are used to compute the beat-position-inside-a-bar. They are coming from the analysis of the chroma vector variation over time and of the spectral balance (typical pop/rock rhythm patterns are represented by the time evolution of their spectral distribution). These can be linked to the works of Goto [4] [5], Klapuri [8] or Eronen [38][3]. However, in our case we do not explicitly estimate chords or kick/snare events but model the consequences on the signal of their presence (chroma variation and spectral distribution). Another difference lies in the fact that we estimate the beat-position-inside-a-bar role of each beat. In our model, the second, third and fourth beats are not estimated by propagating the periods from the downbeat estimations but are estimated using their own model[4]. Also this model is based on past and future signal observations of the local bar the beat is located in. This provides us with an inherent local normalization of the probabilities (or in other words with an adaptation to the local properties of the signal).

II. Probabilistic framework

A. Introduction

We define the "beat-position-inside-a-bar" (bpib) [35] as the position of a beat relative to the downbeat position of the bar it is located in. We denote it by β_j with $j \in [1, B]$ where B is the number of beats in a bar (β_1 denotes the downbeat, β_2 the second beat of the bar ...). B can have a fixed value in case of constant meter, or takes the maximum number of allowed beats in a bar in case of variable meters. We will use the estimation of the β_j associated to each beat to derive the downbeats (β_1).

We define $\{\beta\}$ as the set of times being a beat position. We define $\{\beta_j\}$ as the set of times being in a β_j, with $j \in [1, B]$. Of course $\{\beta_j\}$ is a sub-set of $\{\beta\}$ since the bpib are by definition beats. Beat-tracking is the problem of finding the $t \in \{\beta\}$, downbeat-tracking is the problem of finding the $t \in \{\beta_1\}$. In this work, we solve the problem of finding the $t \in \{\beta_j\} \ \forall j$.

Without any prior assumption, any time t of a music track can be considered as a $t \in \{\beta_j\}$. We therefore define a set of hidden states corresponding to each time t of a music track in each possible β_j. For a given track, the number of hidden states is fixed and depends on the track length (through the quantization of the times axis) and on B. We denote by t_i the values of the discretization of the time-axis of a music track: $t_i = iQ \ \ i \in \mathbb{N} \cap [0, \lfloor \frac{T}{Q} \rfloor]$ where Q is the discretization step (we use here $Q = 0.05$s) and T is the total length of the music track. As mentioned above, any time t_i can be considered as a $t_i \in \{\beta_j\}$. We then denote by $s_{i,j}$ the hidden states defined by $t_i \in \{\beta_j\}$. Our goal is to decode the path through the $s_{i,j}$ that best explains our signal observation $\underline{o}(t)$.

For this we consider the observation probabilities:

$$p_{obs}(s_{i,j}|\underline{o}(t)) = p_{obs}(t_i \in \{\beta_j\}|\underline{o}(t)) \quad (1)$$

We also consider the transition probabilities

$$p_{trans}(s_{i',j'}|s_{i,j}) = p_{trans}(t_{i'} \in \{\beta_{j'}\}|t_i \in \{\beta_j\}) \quad (2)$$

In the transition probabilities, we will use the fact that if $t_i \in \{\beta\}$ than the next $t_{i'} \in \{\beta\}$ must be separated by a local

[3]It should be noted however that in [38] Eronen only uses the chroma variation as an accent function in order to estimate the signal periodicity in his K-NN regression approach. He does not deal with the estimation of beat or downbeat location.

[4]It should be noted that a similar approach has been taken by Jehan [11] pg. 86 (in his supervised trained system for downbeat estimation) or Whiteley [39] (in his dynamic bar pointer model).

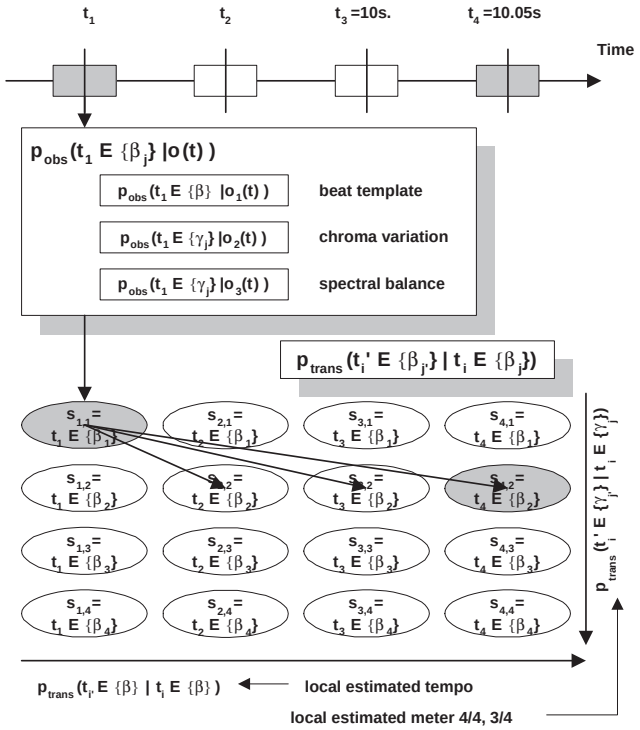

Fig. 1. *Illustration of the observation probability $p_{obs}(t_i \in \{\beta_j\}|\underline{o}(t))$ and transition probability $p_{trans}(t_{i'} \in \{\beta_{j'}\}|t_i \in \{\beta_j\})$. We also illustrate the splitting (see part II-D) - of the observation probability $p_{obs}(t_i \in \{\beta_j\}|\underline{o}(t)) = p_{obs}(t_i \in \{\beta\}|\underline{o_1}(t)) \cdot p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}(t)) \cdot p_{obs}(t_i \in \{\gamma_j\}|\underline{o_3}(t))$, - and of the transition probability $p_{trans}(t_{i'} \in \{\beta_{j'}\}|t_i \in \{\beta_j\}) = p_{trans}(t_{i'} \in \{\beta\}|t_i \in \{\beta\}) \cdot p_{trans}(t_{i'} \in \{\gamma_{j'}\}|t_i \in \{\gamma_j\})$.*

tempo period. We will also use the fact that if $t_i \in \{\beta_j\}$ than the next $t_{i'} \in \{\beta\}$ must be in $\beta_{j+1 mod B}$ (i.e. following the succession of bpib implied by the local musical meter).

In order to clarify this, we provide as short example the illustration of Figure 1 corresponding to the case of a 4/4 meter ($B = 4$). The time $t_3 = 10s$ of the track can be potentially any of the four bpib. It is therefore represented by 4 states $(s_{3,1}, s_{3,2}, s_{3,3}, s_{3,4})$ corresponding to the four possible β_j in a 4/4 meter. The time $t_4 = 10.05s$ is also represented by 4 states $(s_{4,1}, \ldots)$. Given an audio signal, we can estimate the probability to observe each of the state $s_{i,j}$. Given the definition of transition probabilities (taking into account the fact that two beats must be separated by a local period and the fact that there is a high probability that a β_1 well be followed by a β_2) we can decode the states $s_{i,j}$ over the beats, i.e. find the best succession of times (and their associated bpib) over beats that best explain the observations. Suppose during the decoding none of the 4 states attached to time $t_3 = 10s$ have been used. This means that time 10s is not a beat. But during the decoding the state β_2 attached to time $t_4 = 10.05s$ has been used. This means that time 10.05s is the second beat of the local bar.

Figure 1 also illustrates the splitting of the probabilities explained in part II-D. We invite the reader to come back to it when reading part II-D

B. Hidden Markov model

We consider the usual hidden Markov model formulation [40], which models the probability to observe the hidden states s given the observation $\underline{o}(t)$ over time t. This model is defined by - the definition of the hidden states s, - the initial probability $p_{init}(s)$, - the emission probability $p_{emi}(\underline{o}|s)$, - the transition probability $p_{trans}(s'|s)$. The best path through the hidden states s given the observations $\underline{o}(t)$ over time is found using the Viterbi decoding algorithm.

In our formulation, the hidden states $s_{i,j}$ are defined as $t_i \in \beta_j$, i.e. "time t_i is a beat and is in a specific β_j". It should be noted that the time is therefore part of the hidden state definition. This is done in order to be able to apply the periodicity constraint[5] in the transition probabilities. The probabilities are defined as follows:

- The initial probability $p_{init}(s_{i,j}) = p_{init}(t_i \in \{\beta_j\})$ represents the initial probability to be in hidden state [time t_i is a beat and is in a specific β_j]. While in usual Viterbi decoding, "initial" refers to the time t_0 (since the usual decoding operates over time); in our case "initial" refers only to the beginning of the decoding without explicit reference to a time.
  In our system, we do not favor any β_j in particular, but we favor t_i to be a time close to the beginning of the track. $p_{init}(s_{i,j})$ is modeled as a Gaussian function with $\mu = 0, \sigma = 0.5$ evaluated on the t_i of all the states.
- The emission probability $p_{emi}(\underline{o}(t)|s_{i,j}) = p_{emi}(\underline{o}(t)|t_i \in \{\beta_j\})$ represents the probability that the state $s_{i,j}$ (or [time t_i is a beat and is in a specific β_j]) has emitted $\underline{o}(t)$. Note that in this formulation the hidden states $s_{i,j}$ have a non-null emission probability only when $t = t_i$ in $\underline{o}(t)$ (this is because we cannot emit $\underline{o}(t)$ when $t_i \neq t$).
- The transition probability $p_{trans}(s_{i',j'}|s_{i,j}) = p_{trans}(t_{i'} \in \{\beta_{j'}\}|t_i \in \{\beta_j\})$ represents the probability to transit from [time t_i is a beat and is in a specific β_j] to [time $t_{i'}$ is a beat and is in a specific $\beta_{j'}$]. Because we only allow transitions to increasing times t_i, our model is a Left-Right hidden Markov model.

C. Decoding: "reverse" Viterbi algorithm

Because of the introduction of the times t_i in the hidden state definition, the Viterbi decoding is performed over a variable named "beat-numbers" (instead of over time) and noted $bn_k \in \mathbb{N}$. Therefore, we somehow reverse the axis of the Viterbi algorithm since we decode times (the hidden states $s_{i,j} = t_i \in \{\beta_j\}$) over the "beat-numbers" bn_k. We compare the usual Viterbi formulation to the reverse Viterbi formulation in Figure 2 in which we omit the j index for clarity.

In the following, we explain the Forward and specific Backward algorithm we use.

1) Forward: We first remark that the emission probability $p_{emi}(\underline{o}(t)|s_{i,j})$ does not vary over the decoding axis. This is because the decoding operates over the succession of beat number bn_k (and not over the time) over which $p_{emi}(\underline{o}(t)|s_{i,j})$

[5]The periodicity constraint represents the fact that the times t_i associated to two successive beats must be separated by a local tempo period Tb.

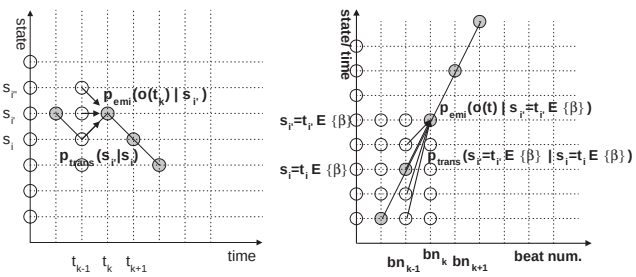

Fig. 2. *(Left) Usual Viterbi decoding: we decode the state s_i over time t_k given a) the probability to emit $\underline{o}(t)$ at time t_k given a state $s_{i'}$: $p_{emi}(\underline{o}(t_k)|s_{i'})$, b) the probability to transit from state s_i to state $s_{i'}$: $p_{trans}(s_{i'}|s_i)$. (Right) Reverse Viterbi decoding: we decode the states s_i (or $t_i \in \{\beta\}$) over beat-number bn_k given a) the probability to emit $\underline{o}(t)$ at beat number bn_k given a state $s_{i'}$ (or $t_{i'} \in \{\beta\}$): $p_{emi}(\underline{o}(t)|s_{i'} = t_{i'} \in \{\beta\})$, b) the probability to transit from state s_i (or time $t_i \in \{\beta\}$) to state $s_{i'}$ (or time $t_{i'} \in \{\beta\}$) : $p_{trans}(s_{i'} = t_{i'} \in \{\beta\}|s_i = t_i \in \{\beta\})$.*

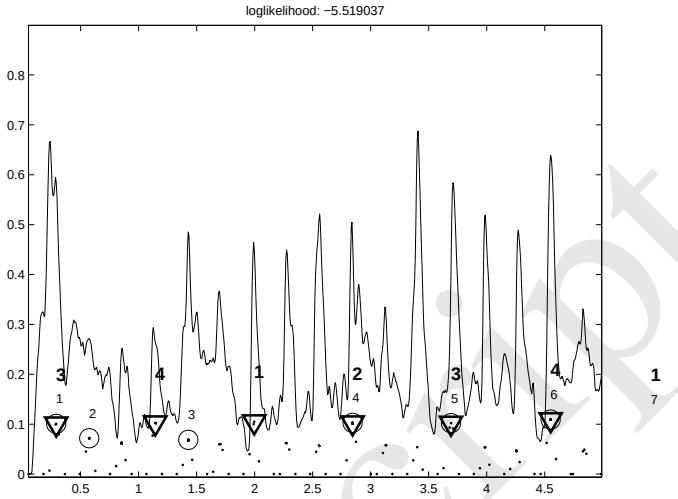

Fig. 3. *Viterbi decoding and backtracking: onset-energy-function (continuous thin line), states $s_{i,j}$ and associated observation probabilities (dots), maximum observation probability of each bn_k (O sign), best Viterbi decoding path ($\triangle$ sign), bn_k (normal number), β_j (bold number). on signal ["Aerosmith - Cryin" from the T-PR test-set].*

remains constant. Because of that, the same $p_{emi}(\underline{o}(t)|s_{i,j})$ is used over the whole decoding (initialization and forward). The Forward algorithm is actually mainly governed by the transition probabilities.

- **Initialization:** We initialize the decoding using $\delta_0(s_{i,j}) = p_{init}(s_{i,j}) \cdot p_{emi}(\underline{o}(t)|s_{i,j})$, i.e. estimating the most-likely $s_{i,j}$ (or $t_i \in \{\beta_j\}$) at beat number bn_0 (at beginning of the track) given their emission probabilities.
- **Forward:** We go on by computing $\delta_k(s_{i',j'}) = p_{emi}(\underline{o}(t)|s_{i',j'}) \max_{i,j}[p_{trans}(s_{i',j'}|s_{i,j}) \cdot \delta_{k-1}(s_{i,j})]$.
- **Ending:** We note τ_k the value of the time t_i associated to the most-likely ending state $s_{i,j}$ for a forward path going until step bn_k. We stop the forward algorithm when τ_k reaches the end of the music track.

2) Backward: In the usual Viterbi algorithm, the final path is found by using the backward algorithm starting from the most-likely ending state. However, in our reverse Viterbi decoding formulation, the last decoded hidden states (which correspond to the last bn_k which is chosen such as with τ_k close to the end of the music track) can correspond to a time τ_k in a silent part (the end of the files can be a silence period) which is not a beat. In other words, we do not know which the best ending state is since we do not know which the last bn_k is. We therefore modified the backward algorithm as follows[6].

Modified backward algorithm: Instead of computing a single backward path, we compute all the backward paths for all the bn_k with τ_k close to the end of the track. Since these various paths can have different (but close) lengths, we normalize the log-likelihood of each path by its length before comparing them. We finally choose the path which has the highest normalized log-likelihood.

3) Result: The decoding attributes to each beat number bn_k the best hidden state $s_{i,j}$ considering the observation $\underline{o}(t)$. It therefore provides us simultaneously the best times t_i for the beat locations and their associated β_j, among which β_1 represent the downbeat locations.

In Figure 3, we illustrate the results of this decoding algorithm on a real signal.

[6]It should be noted that in [15], Ellis also faced this problem in its Dynamic Programming approach and proposed a different solution to this problem.

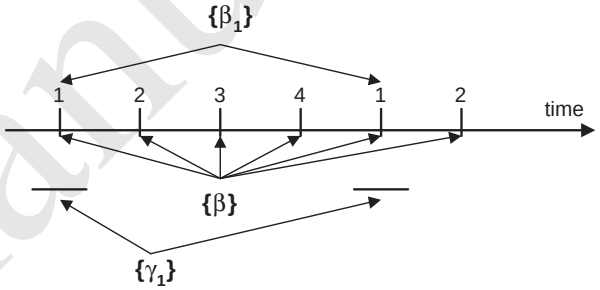

Fig. 4. *$\{\beta\}$ represents the set of beat positions, $\{\beta_j\}$ represents the set of "beat-positions-inside-a-bar" (hence $\{\beta_1\}$ represents the set of downbeat positions), $\{\gamma_j\}$ represents the set of time intervals around the "beat-position-inside-a-bar" (bpib), it is denoted by "vicinity-position-inside-a-bar" (vpib) (hence $\{\gamma_1\}$ represents the set of time intervals around the downbeat positions).*

D. Re-formulation of the probabilities

In practice, in order to estimate the best sequence of hidden states $s_{i,j}$ (or $t_i \in \{\beta_j\}$), we first approximate the emission probability using[7]

$$p_{emi}(\underline{o}(t)|s_{i,j}) \simeq p_{obs}(s_{i,j}|\underline{o}(t)) \tag{3}$$

We denote p_{obs} by "observation probability".

We then split the problem in two. For this, we define $\{\gamma_j\}$ as the set of time intervals around (in the vicinity of) the $\{\beta_j\}$. We therefore have $\{\beta_j\} = \{\beta\} \cap \{\gamma_j\}$. We denoted $\{\beta_j\}$ by "beat-position-inside-a-bar" (bpib), we now denote $\{\gamma_j\}$ by "vicinity-position-inside-a-bar" (vpib). In Figure 4 we illustrate $\{\beta\}$, $\{\beta_j\}$ and $\{\gamma_j\}$.

We therefore have

$$\begin{aligned} p_{obs}(t_i \in \{\beta_j\}|\underline{o}(t)) = & p_{obs}(t_i \in \{\beta\}|\underline{o}(t)) \cdot \\ & p_{obs}(t_i \in \{\gamma_j\}|\underline{o}(t)) \end{aligned} \tag{4}$$

[7]Using Bayes formula, this is equivalent to consider that, - without any information the $p(s_{i,j})$ can be considered equal for all $s_{i,j}$ and - $\sum_{i,j} p(\underline{o}(t)|s_{i,j})p(s_{i,j})$ can be considered as a normalization factor.

Typically, the goal of $p_{obs}(t_i \in \{\beta\}|\underline{o}(t))$ is to estimate precisely the position of the beat. In the opposite, $p_{obs}(t_i \in \{\gamma_j\}|\underline{o}(t))$ uses information surrounding t_i to analyze its local musical context and estimate its vpib role. Because of the use of surrounding information, it's temporal accuracy is lower than the one of $p_{obs}(t_i \in \{\beta\}|\underline{o}(t))$. We therefore require $p_{obs}(t_i \in \{\beta\}|\underline{o}(t))$ to be highly discriminative in terms of beat and non-beat information. In order to do that, we will use different observation functions for $p_{obs}(t_i \in \{\beta\})$ and $p_{obs}(t_i \in \{\gamma_j\})$. These functions are denoted by $\underline{o_1}(t)$, $\underline{o_2}(t)$ and $\underline{o_3}(t)$.

The transition probability is now expressed as

$$p_{trans}(t_{i'} \in \{\beta_{j'}\}|t_i \in \{\beta_j\}) = p_{trans}(t_{i'} \in \{\beta\}|t_i \in \{\beta\}) \cdot p_{trans}(t_{i'} \in \{\gamma_{j'}\}|t_i \in \{\gamma_j\}) \quad (5)$$

The splitting of the observation and transition probabilities is illustrated in Figure 1.

III. Emission probabilities

The emission probability $p_{emi}(\underline{o}(t)|s_{i,j}) = p_{emi}(\underline{o}(t)|t_i \in \{\beta_j\})$ represents the probability to observe $\underline{o}(t)$ given [time t_i is a beat and is in a specific β_j]. As explained i part II-B, this probability has a non-null emission probability only when $t = t_i$. As explained in part II-D, this probability is computed using[8]:

$$p_{obs}(t_i \in \{\beta_j\}|\underline{o}(t)) = p_{obs}(t = t_i) \cdot p_{obs}(t_i \in \{\beta\}|\underline{o_1}(t)) \cdot p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}(t), \underline{o_3}(t)) \quad (6)$$

In this, we have subdivided $\underline{o}(t)$ as three observation vectors $\underline{o_1}(t))$, $\underline{o_2}(t)$ and $\underline{o_3}(t)$. We now explain the two terms in parts III-A and III-B.

A. Beat observation probabilities $p_{obs}(t_i \in \{\beta\}|\underline{o_1}(t))$

$p_{obs}(t_i \in \{\beta\}|\underline{o_1}(t))$ represents the probability to observe [time t_i is a beat] given the observation $\underline{o_1}$ at time t. As explained above, t must be equal to t_i. We therefore use the t_i notation in the following. As in many works, this probability is estimated by computing the correlation between - a beat-template $g(t)$ chosen to correspond to the local tempo Tb(t_i) and - the local onset-energy function starting at time t_i. The beat-template $g(t)$ can be a simple function with values of 1 at the expected beat-position and 0 otherwise (as used in [36]). In [34], we have proposed the use of machine learning to find the beat-template that maximizes the discrimination between the correlation values obtained when $t_i \in \{\beta\}$ and when $t_i \notin \{\beta\}$. We summarize it here using our framework notations and refer the reader to [34] for details and evaluation of it.

[8]In order to split $p_{obs}(t_i \in \{\beta_j\}|\underline{o}(t))$ in two terms we use the assumption that $\underline{o_1}$ and $\underline{o_2}, \underline{o_3}$ are independent, and that $\underline{o_1}$ and $\underline{o_2}, \underline{o_3}$ are independent conditionally to $t_i \in \{\beta_j\}$, i.e. knowing $t_i \in \{\beta_j\}$, the knowledge of $\underline{o_1}$ does not bring information on $\underline{o_2}, \underline{o_3}$.

1) Learning the best beat-template by Linear Discriminant Analysis: We note $f_i(t) = f(t, t \in [t_i, t_i + 4\text{Tb}])$ the values of the local onset-energy function starting at time t_i. The beat-template $g(t)$ must be chosen such as (a) to have the maximum correlation with $f_i(t)$ when $t_i \in \{\beta\}$, (b) to provide the largest discrimination between the correlation values when $t_i \in \{\beta\}$ and when $t_i \notin \{\beta\}$. The condition (b) is needed in our case since the correlation values will be used as observation probabilities in our framework. In the following, we only discuss the case of a "binary subdivision of the beat" and "binary grouping of the beat into bar". Extension to other meters is straightforward.

We note $g(1) \dots g(N)$ the discrete sequence of values of the beat-template $g(t)$ representing a one-bar duration. Considering a 4/4 meter, $g(1)$ represents the value at the downbeat position, $g(1 + \frac{kN}{4})$ with $k \in [0, 1, 2, 3]$ the values at the beat positions. In the same way, we define $F_i(n)$ as the function obtained by sampling the local values of $f_i(t)$ by N value: $F_i(1) = f_i(t_i) \dots F_i(N) = f_i(t_i + 4\text{Tb})$. If t_i is a beat-position, $F_i(1 + \frac{kN}{4})$ with $k \in [0, 1, 2, 3]$ represent the values at the beat positions.

The correlation between $g(n)$ and $F_i(n)$ can be written as (neglecting the normalization terms): $c_i(j) = \sum_{n=1}^{N} F_i(n + j)g(n)$

If we choose t_i as a beat-position, we therefore look for the beat-template (the values of $g(n), n \in [1, N]$) for which

- (a) $c_i(j)$ is maximum at $j \in [0, \frac{N}{4}, \frac{2N}{4}, \frac{3N}{4}]$
- (b) $c_i(j)$ is minimum for all the other values of j

The problem of finding the best values of $g(n)$ is close to the problem of finding the best weights to apply to the dimensions of multi-dimensional observations in order to maximize class separation. This problem can be solved using Linear Discriminant Analysis (LDA) [41]. In our case the weights are the $g(n)$, the dimensions of the observations are the successive values of $F_i(n)$ [9] and the two classes are "beat" and "non-beat". We therefore apply a two-class Linear Discriminant Analysis to our problem.

Creating observations for the two-class LDA problem: In order to apply the Linear Discriminant Analysis, we create observations for the two classes "beat" and "non-beat". These observations are coming from a test-set annotated into beat and downbeat positions. We create for each track l of the test-set and for each annotated bar m of a track, the corresponding $F_{i,l,m}(n)$. We then compute the vector $F_{i,l}(n)$ by averaging the values of $F_{i,l,m}(n)$ over all bars of a track. By shifting (circular permutation is assumed in the following) $F_{i,l}(n)$, we create two sets of observations corresponding to the two classes "beat" and "non-beat": - "beat" class: the four patterns $F_l^b(n) = F_{i,l}(n + j)$ with $j \in [0, \frac{N}{4}, \frac{2N}{4}, \frac{3N}{4}]$, - "non-beat" class: all the remaining patterns $F_l^{nb}(n) = F_{i,l}(n + j)$ with $j \in [1, N]$ $j \notin [0, \frac{N}{4}, \frac{2N}{4}, \frac{3N}{4}]$. We then apply Linear Discriminant Analysis considering the two set of observations $F_l^b(n)$ and $F_l^{nb}(n)$ and their associated classes "beat" and "non-beat".

[9]It should be noted that considering the values of $F_i(n)$ as points in a multi-dimensional features space has been also used in [42] in the framework of rhythm classification.

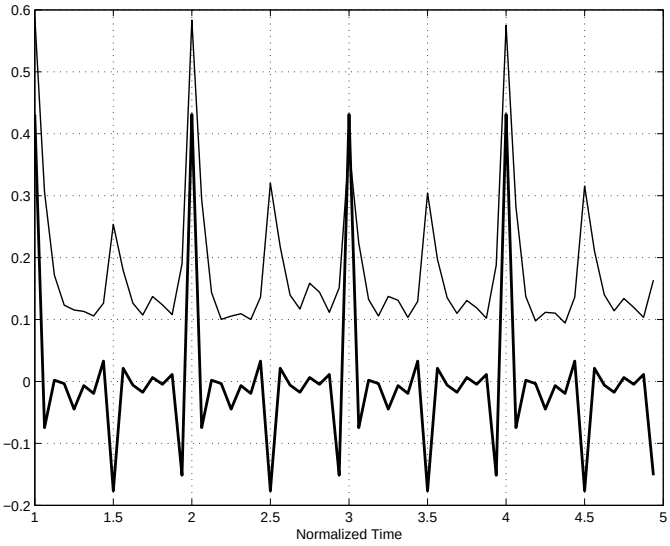

Fig. 5. *Average (over the tracks) value $F(n)$ (thin line) and LDA-trained beat-template $g(n)$ for the T-RWC-P test-set.*

Linear Discriminant Analysis: We compute the matrix $\underline{\underline{U}}$ such that after transformation of the multi-dimensional observation by this matrix, the ratio of the Between-Class-Inertia to the Total-Inertia is maximized. If we note $\underline{u}$ the column vectors of $\underline{\underline{U}}$, this maximization leads to the condition $\underline{\underline{T}}^{-1}\underline{\underline{B}}\underline{u} = \lambda\underline{u}$, where $\underline{\underline{T}}$ is the Total-Inertia matrix and $\underline{\underline{B}}$ the Between-Class-Inertia matrix. The column vectors of $\underline{\underline{U}}$ are then given by the eigen vectors of the matrix $\underline{\underline{T}}^{-1}\underline{\underline{B}}$ associated to the eigen values λ. Since our problem is a two-classes problem, only one column remains in $\underline{\underline{U}}$. This column gives us the weights to apply to $F(n)$ in order to obtain the best separation between the classes "beat" and "non-beat". It therefore defines the best (in terms of discrimination) beat-template $g(n)$.

Result: In Figure 5, we illustrate this for the RWC-Popular-Music test-set [43]. The thin line represents the average (over the 100 tracks) vector $F(n)$, the thick line represents the values of $g(n)$ obtained by Linear Discriminant Analysis. As one can see, the LDA-trained beat-template assigns - large positive weights at the beat-positions (1, 2, 3, 4) and - negative weights at the counter-beat positions (1.5, 2.5, ...) and at the just-before/ just-after beat positions. The use of negative weights is a major difference with the weights used in usual beat-templates (as in [36]) which only use positive or zero weights. The specific locations of the negative weights allow reducing the common counter-beat detection errors (negative weights at the counter-beat positions) and improving the precision of the beat location (negative weights at the just-before/ just-after beat positions). This wouldn't be achieved by using a model where all the positions outside the main beats are set to a constant negative number.

It should be noted that the proposed method does not necessitates that the audio tracks used for training have a steady tempo. This is because their time-axis is re-sampled in the interval $[0, N]$ which can be done even in the case of time-varying tempo. However, the resulting trained beat-templates represent a steady tempo audio signal. This therefore creates a bias when using it to estimate the beat positions of audio signals with time-varying tempo.

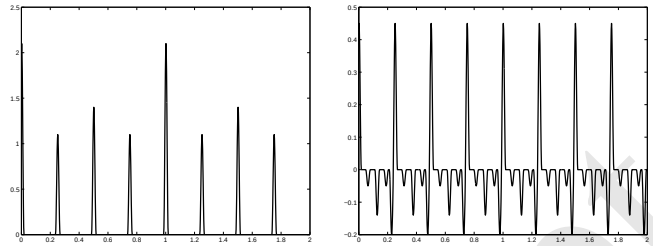

Fig. 6. *Beat-templates used for the computation of the observation probability for a tempo of 120bpm (beat period of 0.5s) and a binary subdivision and grouping of the beat (LEFT) Simple beat template (as used in [36]) (RIGHT]) LDA trained beat-template.*

Use of the LDA-trained beat-templates: In the beat-tracking process, the LDA-trained beat-templates $g(n)$ are used to create the beat-templates corresponding to the local tempo $\mathrm{Tb}(t_i)$. For this, - either $g(n)$ is considered as representing the interval $[0, 4\mathrm{Tb}(t_i)]$ and is interpolated to provide the values corresponding to the sampling rate of $f(t)$[10], - or $g(n)$ is used to create a "model". In both cases, in order to save computation time, one can store in a table the values of $g_{\mathrm{Tb}}(t)$ corresponding to all possible tempi Tb.

For the evaluation of part V-C, we use an LDA-trained beat-template obtained using the "model" approach. This model is manually drawn by keeping only the "salient" points of a trained-template $g(n)$. The "salient" points denote the negative values of the template which were found to be the most discriminative aspect of the template. The use of a model instead of a sampled-template allows an easier adaptation to the various tempi and allows reducing over-fitting to the training-set.

For the template used in part V-C, we have performed the training on the "PopRock extract" test-set. Of course, one can wonder about the applicability of this template to non PopRock music. Ideally, if one knew the type of rhythm of the audio track, one would use the most-appropriate LDA-trained beat-templates (the one trained specifically for this type of rhythm). However, in [34] we have shown that whatever database used for training the template, its use is always better than the use of a "simple" (as used for example in [36]) beat-template. We have shown this by using cross-database validation (one database is used for training, another one for evaluation) using databases representing various types of rhythm (pop, electro, jazz, classical music including rhythm in 4/4, 3/4 with or without swing).

On the right part of Figure 6, we represent the model of the beat-template trained on the "PopRock extract" test-set as will be used in part V-C. For comparison, we represent the "simple" beat-template (as used for example in [36]).

2) Optimization considerations: As mentioned above, the hidden states are defined as $t_i \in \{\beta_j\}$. For this, the time axis of a music track is discretized into $t_i = iQ \;\; i \in \mathbb{N} \cap [0, \lfloor \frac{T}{Q} \rfloor]$ with

[10] $f(t)$ is a 172 Hz function.

$Q = 0.05$s. Large values of Q allows decreasing the number of hidden states but decrease the temporal-precision of the beat-tracking. Because of that, we reassign the time t_i of the state $s_{i,j}$ to the position around t_i which leads to the maximum correlation between the local signal $f(t, t \in [t_i, t_i + 4\text{Tb}])$ and the beat-template $g(t)$. The horizon over which the maximum correlation is searched for is proportional to the local tempo $\text{Tb}(t_i)$ and defined by $L = \text{Tb}(t_i)/\tau$[11]. In [34], we have tested several values of τ. The best results were obtained with τ= 32 which is the value used in the remaining of this paper.

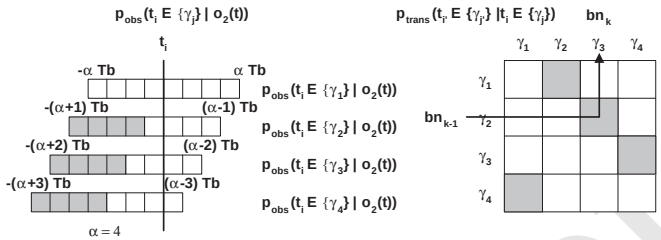

Fig. 7. *(LEFT) Computation of observation probabilities for the vpib from chromagram observation. (RIGHT) Transition probabilities between vpib.*

B. VPIB observation probabilities $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}(t), \underline{o_3}(t))$

$p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}(t), \underline{o_3}(t))$ represents the probability to observe [time t_i is a γ_j] given the observation $\underline{o_2}, \underline{o_3}$ at time t. Any probability derived from signal observations (such as based on harmonic, spectral or loudness/ silence variation) that allows distinguishing between the various γ_j can be used for it. We use here two assumptions to derive the "vpib probability". Each assumption is coupled with a characteristic which is coupled with a signal observation. The first one is based on the chord-change / harmonic-variation / chroma-vector-variation triplet. The second one is based on the rhythm-pattern / low-high-frequency alternation / spectral-distribution triplet. This probability is computed using[12]:

$$p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}(t), \underline{o_3}(t)) = p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}(t)) \cdot p_{obs}(t_i \in \{\gamma_j\}|\underline{o_3}(t)) \quad (7)$$

In this,

- $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}(t)$ is the probability to observe [time t_i is a γ_j] given the observation of chroma vectors variation.
- $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_3}(t))$ is the probability to observe [time t_i is a γ_j] given the observation of spectral distribution.

1) VPIB probability based on chroma variation: We use the *assumption* that chords are more likely to change on γ_1 (at the beginning of the bar). [5] or [35] also used this assumption for downbeat estimation. We use it here to derive the probability of all γ_j at all times t_i. The *characteristic* implied by this assumption is that, if t_i is a γ_1, the harmonic content on its left and on its right should be different. The *observation* we use to highlight this, is the variation of chroma vectors over time. A large variation indicates a potential change in harmony at time t_i hence a higher probability to observe a downbeat at t_i hence a γ_1. The probabilities for the other $\gamma_{j=2,3,4}$ are derived in the same way.

Chroma vector computation: The chroma vectors (or Pitch-Class-Profile vectors) [44] are computed as in [45], i.e. the Short Time Fourier Transform is first computed with a Blackman analysis window of length 0.1856ms and a hop size of 0.0309ms. Each bin is then converted to a note-scale. Median-filtering is applied over time to each note-band in order to reduce transients and noise. Note-bands are then grouped into 12-dimensions vectors. We note $C(l,t)$ the values of the $l \in [1, 12]$ dimension of the chroma vector at time t.

Chroma vector variation: We compare the values taken by $C(l,t)$ on the left of t_i and on its right using two temporal window of duration α. We note $L_{i,1} = [t_i - \alpha\text{Tb}, t_i]$ the left window and $R_{i,1} = [t_i, t_i + \alpha\text{Tb}]$ the right window. α is expressed as a multiple of the local beat duration. In the experiment of part V, we will compare the results obtained with $\alpha = 2$ (assumption that chords change twice per bar) and $\alpha = 4$ (once per bar).

Sliding-window method: In the same way, we compute $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}(t))$ (the probability that t_i is the j^{th} vpib), using the assumption that the harmonic content should be different on the left of $t_i - (j-1)\text{Tb}$ and on its right. This is illustrated in the left part of Figure 7 for the case of a 4/4 meter ($j = 1, 2, 3, 4$). The computation of $p_{obs}(t_i \in \{\gamma_j\}\}|\underline{o_2}(t))$ is therefore obtained by comparing $C(l,t)$ on the intervals $L_{i,j}$ and $R_{i,j}$ defined by

- $L_{i,j} = [t_i - (\alpha + (j-1))\text{Tb}, t_i - (j-1)\text{Tb}]$,
- $R_{i,j} = [t_i - (j-1)\text{Tb}, t_i + (\alpha - (j-1))\text{Tb}]$.

We name this method "sliding-window method" since we slide the analyzed signal according to our β_j assumption.

Distance measures: We study two measures for the computation of the chroma vectors variation. The first measure is the symmetries Mahalanobis distance: $d(L_{i,j}, R_{i,j}) = \frac{1}{2}((\mu_2 - \mu_1)^T \Sigma_1^{-1}(\mu_2 - \mu_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1}(\mu_1 - \mu_2))$ where μ_1 and μ_2 (Σ_1 and Σ_2) are the 12-dimensional mean vectors (12x12dimensional diagonal covariance matrices) of the values of $C(l, t \in L_{i,j})$ and $C(l, t \in R_{i,j})$ respectively. The second measure is a simple "1-cosine" distance between the vectors $\underline{\mu_1}$ and $\underline{\mu_2}$ (it has value of 1 when $\underline{\mu_1}$ and $\underline{\mu_2}$ are in orthogonal directions): $d(L_{i,j}, R_{i,j}) = 1 - \frac{\underline{\mu_1} \cdot \underline{\mu_2}}{||\underline{\mu_1}|| ||\underline{\mu_2}||}$. In the experiment of part V, we will compare both distances.

VPIB probabilities: Both distances have large values when $L_{i,j}$ and $R_{i,j}$ have different harmonic content which indicates a potential downbeat. We therefore use the distances $d(L_{i,j}, R_{i,j})$ as probabilities. For this the probabilities are normalized:

$$p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}(t)) = \frac{1}{\sum_j d(L_{i,j}, R_{i,j})} d(L_{i,j}, R_{i,j}) \quad (8)$$

In Figure 8, we illustrate the computation of $p_{obs}(t_i \in$

[11]Too small values of τ (hence large temporal horizon) leads to reassign several states $s_{i,j}$ to the same time (since the successive horizons overlap), while too large values of τ (hence small temporal horizon) leads to the miss-detection of the real beat location (since the horizons do not overlap anymore)

[12]In order to split $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}, \underline{o_3})$ in two terms we use the assumption that $\underline{o_2}$ and $\underline{o_3}$ are independent, and that $\underline{o_2}$ and $\underline{o_3}$ are independent conditionally to $t_i \in \{\gamma_j\}$, i.e. knowing $t_i \in \{\gamma_j\}$, the knowledge of $\underline{o_3}$ does not bring information on $\underline{o_2}$.

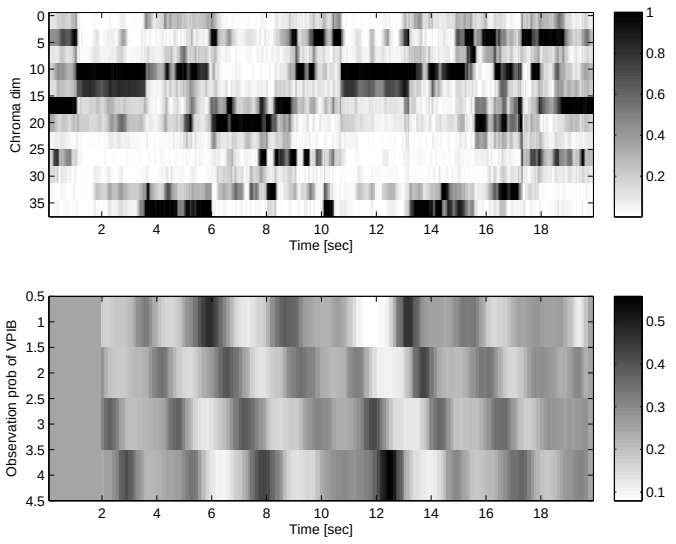

Fig. 8. *(TOP) 12-dimensional chromagram over time, (BOTTOM)* $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}(t))$ *for* $j = 1, 2, 3, 4$ *, on signal ["All Saints - Pure Shores" from the T-PR test-set].*

$\{\gamma_j\}|\underline{o_2}(t))$ on a real signal using $\alpha = 2$ and a "1-cosine" distance.

2) VPIB probability based on spectral distribution: The *assumption* we use is that many music tracks in popular music (pop, rock, electro) use rhythm patterns alternating the presence of kick on $\gamma_{1,3}$ and snare on $\gamma_{2,4}$. [4] or [8] also used this assumption. The *characteristic* implied by this assumption is that the spectral energy distribution will concentrate on lower frequencies for $\gamma_{1,3}$ than for $\gamma_{2,4}$. The *observation* we use to highlight this, is the relative spectral balance between high and low energy content.

Spectral balance computation: At each time t_i, we compute the ratio of the high frequency to the low frequency energy content. For this we use a window centered on t_i of length L and a cutting frequency $kmax$:

$$r(t_i) = \frac{\sum_{t=t_i-L/2}^{t_i+L/2} \sum_{k=kmax}^{N/2} |S(\omega_k, t)|^2}{\sum_{t=t_i-L/2}^{t_i+L/2} \sum_{k=1}^{kmax} \sum_k |S(\omega_k, t)|^2} \quad (9)$$

where N is the number of bins of the Short Time Fourier Transform. L was chosen experimentally to Tb$/2$ and $kmax$ to correspond to 150Hz.

Example: Using the "PopRock extract" test-set annotated into beat and downbeat, we have measured the values of $r(t_i)$ for $t_i \in \{\gamma_{j=1,2,3,4}\}$. For 135 over the 156 titles of this test-set, $r(t_i)$ is larger for the γ_2/γ_4 than for the γ_1/γ_3. We therefore use it to create a probability to observe $\gamma = 1, 3$ or $\gamma = 2, 4$.

BPIB probability: As for the chroma-variation-measure, we use a sliding-window method to derive $r(t_i)$ for all γ_j. At each time t_i, we compute the four values:

$$r_j(t_i) = r(t_i - (j-1)\text{Tb}) \quad (10)$$

r_j is then normalized over the j to sum unit. If $t_i \in \gamma_1$, the following sequence of r_j will be observed [r_1=low, r_2=high, r_3=low, r_4=high]. Since we would like the probability to have high values for γ_1, low values for γ_2, ... we take the negative of $r_j(t_i)$ as probability:

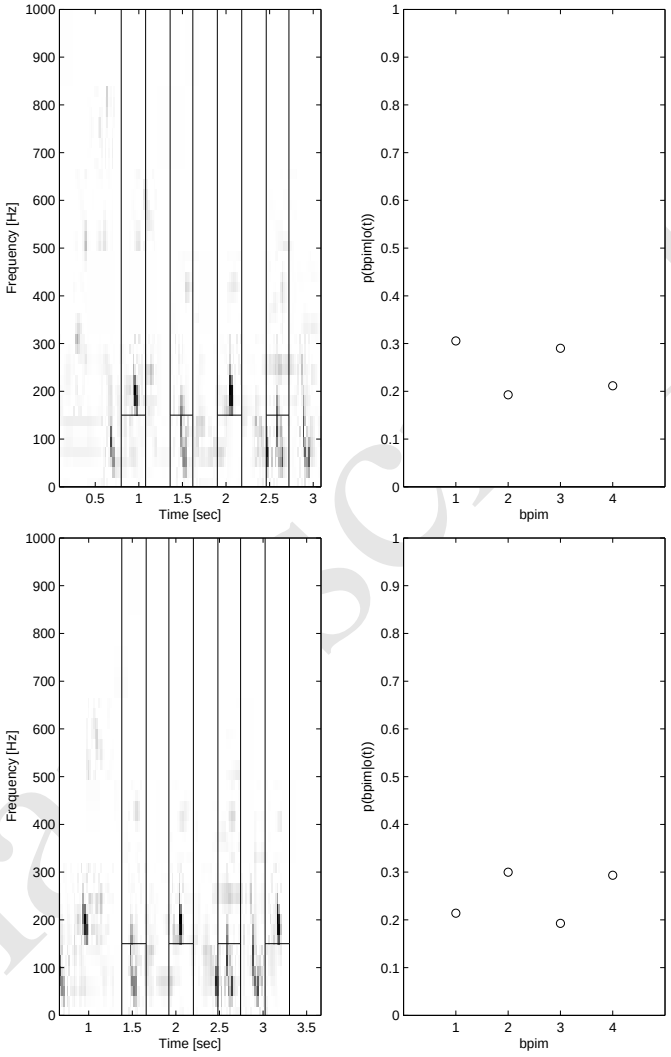

Fig. 9. *(TOP) Spectrogram and* $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_3}(t))$ *for* $j = 1, 2, 3, 4$ *for* t_i *on a* γ_1*, (BOTTOM) Spectrogram and* $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_3}(t))$ *for* $j = 1, 2, 3, 4$ *for* t_i *on a* γ_2 *on signal ["Aerosmith - Walk This Way" from the T-PR test-set].*

$$p_{obs}(t_i \in \{\gamma_j\}|\underline{o_3}(t)) = 1 - r_j(t_i) \quad (11)$$

In Figure 9, we illustrate the computation of $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_3}(t))$ on a real signal. The left parts of each figure represent the spectrogram of the signal and super-imposed to it the four regions used for the computation: $t_i + [-\frac{L}{2}, \frac{L}{2}]$, $t_i - \text{Tb} + [-\frac{L}{2}, \frac{L}{2}]$, $t_i - 2\text{Tb} + [-\frac{L}{2}, \frac{L}{2}]$ and $t_i - 3\text{Tb} + [-\frac{L}{2}, \frac{L}{2}]$. We also indicate the cutting frequency of 150Hz. The right part of each figure indicates the four values of $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_3}(t_i))$ at the given position. The upper figure represents the values obtained when t_i is a γ_1, the lower one a γ_2.

IV. Transition probabilities

The transition probability $p_{trans}(t_{i'} \in \{\beta_{j'}\}|t_i \in \{\beta_j\})$ represents the probability to transit from [time t_i is a beat and is in a specific β_j] to [time $t_{i'}$ is a beat and is in a specific

$\beta_{j'}]$. We compute it using:

$$p_{trans}(t_{i'} \in \{\beta_{j'}\}|t_i \in \{\beta_j\}) = p_{trans}(t_{i'} \in \{\beta\}|t_i \in \{\beta\}) \cdot p_{trans}(t_{i'} \in \{\gamma_{j'}\}|t_i \in \{\gamma_j\}) \quad (12)$$

We also add the condition that only transitions to increasing times t_i (increasing states $s_{i,j}$) are allowed. This makes our model a Left-Right HMM.

A. Beat transition probabilities

$p_{trans}(t_{i'} \in \{\beta\}|t_i \in \{\beta\})$ represents the fact that the successive times t_i associated to the beats must have an inter-distance close to the local tempo period $\text{Tb}(t_i)$. The transition probability models the tolerated departure from this period. We have used a Gaussian function with $\mu = \text{Tb}(t_i), \sigma = 0.05$s evaluated on $\Delta = t_{i'} - t_i$.

B. VPIB transition probabilities

$p_{trans}(t_{i'} \in \{\gamma_{j'}\}|t_i \in \{\gamma_j\})$ represents the probability to transit from a beat in γ_j to a beat in $\gamma_{j'}$. This transition probability constrains the γ_j to follow the circular permutation specific to the considered musical meter: $1 \to 2 \to 3 \to 4 \to 1 \to \ldots$ for a 4/4 meter; $1 \to 2 \to 3 \to 1 \ldots$ for a 3/4 meter. As proposed in [37], a generic formulation of the transition matrix allowing potential meter changes between 4/4 and 3/4 meters over time can be written as

$$M_{trans}(bn_{k-1}, bn_k) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \alpha & 0 & 0 & 1 \\ 1-\alpha & 0 & 0 & 0 \end{pmatrix} \quad (13)$$

where bn_k is the beat-number used for the decoding axis and $\alpha \in [0, 1]$ is a coefficient favoring meter changes ($\alpha \in]0, 1[$) or forcing constant-meter-over-time ($\alpha = 0$ for a 4/4 meter, $\alpha = 1$ for a 3/4 meter). In the experiments done so far, we have obtained better results using $\alpha = 0$ (constant-4/4-meter-over-time). In the experiment of part V, we will therefore only consider the case $\alpha = 0$. The corresponding matrix is illustrated on the right part of Figure 7.

V. Evaluation

In this part, we evaluate the performances of the proposed algorithm for beat and downbeat-tracking and test various configurations of it. We compare it to the results obtained with our previous system and to previously published results representing the state-of-the-art. It should be noted that the evaluation performed here only concerns the performances of beat and downbeat-tracking algorithms. However, because the input of our system are the time-variable tempo and meter estimations coming from the algorithm of [32], the results obtained also depend on the performance of our tempo/meter estimation algorithm.

A. Evaluation rules

Over the years, a large number of measures have been proposed to estimate the performances of beat-tracking algorithms: F-measure of Dixon [6], Gaussian error function of Cemgil [46], set of boolean decisions of Goto [47], perceptual P-score of McKinney [48], continuity based measures CMLc, CMLt, AMLc, AMLt of Goto [47], Hainsworth [49] and Klapuri [8], information based criteria based of Davies [20]. We refer the reader to [20] or to the set of rules used for the MIREX-09 "Audio Beat Tracking" contest [23] for a good and detailed overview of those.

In this evaluation, we indicate the results using two criteria[13]. The first is the F-measure for a relative-tempo-length Precision Window of 0.1. We use it for beat and downbeat evaluation when comparing the performances of the various configurations of our system. The second is the set of CMLc, CMLt, AMLc and AMLt criteria. We use them in order to be able to compare our results to the ones published in previous works on the same test-sets.

1) F-measure at a relative-tempo-length Precision Window of 0.1: Considering a given beat/ downbeat marker annotation and a given track, we note - A: the number of annotated beats (downbeats), - D: the number of detected ones and - CD(PW): the number of correctly detected ones within a given Precision Window (PW). From this we derive the following measures:

- $Recall(PW) = \frac{CD(PW)}{A}$,
- $Precision(PW) = \frac{CD(PW)}{D}$,
- $FMeasure(PW) = \frac{2R(PW) \cdot P(PW)}{R(PW)+P(PW)}$.

Note that the Precision Window is centered on the annotated beat (downbeats) for the Recall and on the estimated beat (downbeat) for the Precision.

Octave errors: Using this measure, we do not consider octave errors as correct[14]. For a correct beat marking but at twice (three time) the tempo, the Recall will be 1 but the Precision 0.5 (0.33). for a correct beat marking at half (one third of) the tempo, the Precision will be 1 but the Recall 0.5 (0.33).

Adaptive Precision Window: In our evaluation the Precision Window is defined as a percentage of the local annotated beat length Tb. This is done in order to avoid drawing misleading conclusions from the results[15]. PW=α means that the estimated beat should be at a maximum distance of $\pm\alpha$Tb the annotated beat. For a given track, we consider the minimum value of $\text{Tb}(t_i)$ over time (the fastest annotated tempo). The values

[13]The results of the experiments using the other criteria (using Dixon, Cemgil, Goto, McKinney ...criteria) can be found at the following URL `http:// recherche.ircam.fr/ equipes/ analyse-synthese/ peeters/ pub/ IEEEbeatdownbeat/`.

[14]Many evaluations consider that estimating twice or half the tempo is correct. Actually, this only makes sens in the case of a binary simple meter. For most test-sets we do not have information about the meter. Therefore, we do not consider doubling or halving the tempo as correct in this evaluation. Moreover, in the case of beat-tracking, - doubling the tempo will require to check that the detected beat-markers correspond to all the tatum (and not only to the tatum corresponding to counter-beats), - halving the tempo will require to check that the detected beat-markers corresponds to the dominant beats (downbeats) in the bar.

[15]Indeed a fixed PW of 0.166s would be restrictive for slow tempi (half-beat duration of 0.5 at 60bpm) but will mean accepting counter-beat as correct for fast tempi (half-beat duration of 0.166s at 180bpm).

given in the following correspond to the average (over all tracks of a test-set) of the F-measure(PW=0.1).

Statistical hypothesis tests: In our evaluation we will compare the values of the F-measure obtained using various configurations of our system. Given a test-set T and a configuration C, we create the vector $\underline{F}_{T,C}$ containing the F-measure(PW=0.1) values for each track of the test-set T. In the following result tables, we indicate the mean value of $\underline{F}_{T,C}$. Because this value is only an estimate of the real mean value of the F-measure, we also perform statistical tests. The goal of these tests is to infer statistical significance of the difference between the estimated mean values. For this, for a given test-set T, we compare the vectors $\underline{F}_{T,C}$ and $\underline{F}_{T,C'}$ using a pairwise Student T-tests with the null hypothesis that the mean of the vectors are equal (we do not assume that the variances of the vectors are equal). We use a 5% significance level[16].

2) CMLc, CMLt, AMLc and AMLt: When comparing our results to previously published results we will use the following measures: - CMLc (Correct Metrical Level with continuity required), - CMLt (same but no continuity required), - AMLc (All Metrical Level with continuity required) and - AMLt (same but no continuity required). We refer the reader to [47] [49] and [8] for more details. For the implementation of CMLc, CMLt, AMLc and AMLt we have used the implementation kindly provided by M. Davies[17]. These measures correspond to the "Correct" and "Accept d/h" criteria and the "Continuity required" and "Individual estimate" categories used in [8]. A precision window of 17.5% as in [8] is used for both estimated marker position and estimated tempo.

B. Test-sets

For the evaluation, we have used six test-sets.

T-PR: The "PopRock extract" is a collection of 155 major top-ten hits of the past decades. Only 20s extract of the tracks are considered. Beat and downbeat annotations have been made by one of the authors[18].

T-RWC-P: The "RWC Popular Music" [50] is a collection of 100 tracks in full-duration of Pop-rock-ballad-heavy-metal popular music.

T-RWC-J: The "RWC Jazz Music" [50] is a collection of 50 tracks in full-duration of Jazz-music with solo piano, guitar, small ensemble or modern-jazz orchestra. The difficulty of this test-set comes from the complexity of the rhythms used in Jazz-music.

T-RWC-C: The "RWC Classical Music" [50] is a collection of 59 tracks in full-duration of Classical-music. The difficulty of this test-set comes from the tempo variations used in Classical-music. Beat and downbeat annotations of the three RWC test-sets are provided by the AIST [51].

T-KLA: "Klapuri" test-set is the one used in [8]. It contains 505 tracks of a wide range of music genre (pop, metal, electro, classical). 474 of them are annotated in beat positions for an extract starting in the middle of the track. 320 of them are annotated in downbeat positions also for an extract starting in the middle of the track. It should be noted that the annotations into downbeats has been made independently from the annotations into beats. Hence, the downbeat positions do not necessarily correspond to beat positions.

T-HAI: "Hainsworth" test-set is the one used in [13], [9] and [52]. It contains 222 tracks, each around 60s length from a large variety of music genres and with time-variable tempo. Because only beat annotations are provided we do not evaluate downbeat-tracking here.

The T-PR, the three RWC and the Klapuri test-sets have been used since they are annotated in beat and downbeat positions. The three RWC test-sets are also freely available to the research community for comparison. The T-KLA and T-HAI[19] have been used in order to provide a comparison with state-of-the-art published results.

C. Beat and Downbeat-tracking results and discussion

In this part, we evaluate the performances of various configurations of our beat and downbeat-tracking algorithm. Table I indicates the results in terms of F-measure with a Precision Window of 0.1 for T-PR, T-RWC-P, T-RWC-J and T-RWC-C using various configurations.

We first distinguish the global model used for the tracking ("Model" column):

- "P-sola" refers to our previous beat-tracking algorithm (see part I-B2). No downbeat estimation is available for this algorithm.
- "Viterbi" refers to all the models proposed in this paper (using only the beat-estimation probabilities $p_{obs}(t_i \in \{\beta\}|\underline{o_1})$ or using the whole beat and bpib probabilities $p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_2})$, $p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_3}$ or $p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_2}, \underline{o_3})$.

We distinguish the beat template used for the computation of $p_{obs}(t_i \in \{\beta\}|\underline{o_1})$ ("Beat-template" column):

- "Simple" refers to the beat-template indicated in the left part of Figure 6,
- "LDA" refers to the use of the LDA-trained beat-template indicated in the right part of Figure 6[20].

We distinguish the algorithm used for downbeat estimation ("Downbeat" column):

- "-" means that the Viterbi model is not used for downbeat estimation. It is only used in beat-tracking configuration $p_{obs}(t_i \in \{\beta\}|\underline{o_1})$.

[16]The detailed values of the Student T-tests (p-value, degrees of freedom) can be found at the following URL `http:// recherche.ircam.fr/ equipes/ analyse-synthese/ peeters/ pub/ IEEEbeatdownbeat/`.

[17]The evalbeat toolbox is accessible at http:// www.elec.qmul.ac.uk/ digitalmusic/ downloads/ beateval/ beateval.zip

[18]The description of this test-set can be found at the following URL `http:// recherche.ircam.fr/ equipes/ analyse-synthese/ peeters/ pub/ IEEEbeatdownbeat/`. The corresponding annotations can be delivered on demand.

[19]We are grateful to A. Klapuri, St. Hainsworth and M. Davies to have let us access these test-sets for the present evaluation

[20]As explained in part III-A1 the LDA-trained beat-template used in all the experiments is a beat-template manually derived from an LDA-training on the T-PR test-set.

TABLE I
BEAT AND DOWNBEAT ESTIMATION RESULTS FOR T-PR, T-RWC-P, T-RWC-J AND T-RWC-C.

			PopRock		RWC Popular		RWC Jazz		RWC Classical	
Model	Beat-Template	Downbeat	beat	downbeat	beat	downbeat	beat	downbeat	beat	downbeat
P-sola			0,87		0,72		0,47		0,41	
Viterbi	Simple	-	0,91		0,84		0,57		0,40	
Viterbi	LDA	-	**0,93**		**0,84**		**0,57**		**0,42**	
Viterbi	LDA	CHRO (α=2 COS)	0,93	0,68	0,84	0,76	0,56	0,46	0,42	**0,35**
Viterbi	LDA	CHRO (α=4 COS)	0,93	0,53	0,84	0,78	0,56	0,40	0,42	0,31
Viterbi	LDA	CHRO (α=2 MAH)	0,91	0,44	0,82	0,49	0,52	0,31	0,39	0,23
Viterbi	LDA	SPEC	0,93	0,49	0,84	0,58	0,56	0,31	0,42	0,22
Viterbi	LDA	CHRO (α=2 cos) + SPEC	**0,93**	**0,74**	**0,84**	**0,80**	**0,55**	**0,47**	**0,41**	**0,34**
Viterbi	LDA	- Chord Detection]		0,64		**0,81**		0,44		0,32

- "CHRO" refers to the use of vpib observation probability based on chroma variation ($p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2})$) to estimate the downbeats,
- "SPEC" refers to the use of vpib observation probability based on spectral balance ($p_{obs}(t_i \in \{\gamma_j\}|\underline{o_3})$) to estimate the downbeats,
- "CHRO+SPEC" refers to the use of vpib observation probability based on chroma variation and spectral balance ($p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}, \underline{o_3})$) to estimate the downbeats,
- "Chord Detection" refers to the results obtained with our previous downbeat-tracking algorithm (see part I-B3). Because this algorithm takes as input the estimation of the beat positions, we have used the best beat estimation ($p_{obs}(t_i \in \{\beta\}|\underline{o_1})$ with an LDA beat-template) to provide it with the beat positions.

It should be noted that when using the Viterbi algorithm, both beat and downbeat estimation are obtained simultaneously. When we mention the use of $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2})$ for downbeat estimation, we actually test the model $p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_2})$ which provides simultaneously beat and downbeat positions.

In the case of the chroma, we also study the influence of the choice parameters used to compute $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2})$:

- "$\alpha = 2$/ $\alpha = 4$" refers to the duration of the window used for the computation of $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2})$.
- "COS/ MAH" refers to the use of the "1-cosine" or "Mahalanobis" distance for the computation of $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2})$.

P-sola against Viterbi: We first compare the P-sola to the Viterbi beat-tracking algorithm. For this we use the baseline Viterbi algorithm, i.e. using the "Simple" beat-template. Results shows a large improvement of the F-measure(PW=0.1) for all test-sets except for T-RWC-C. For T-RWC-P and T-RWC-J, these differences are statistically significant .

Choice of the beat-template (Simple or LDA): We then compare the use of a "Simple" (as used in [36]) to the LDA-trained beat-template. The use of the LDA-trained beat-template leads to a small improvement of beat-tracking results for 2 over 4 test-sets: from FMeas=0.91 to 0.93 for T-PR, 0.4 to 0.42 for T-RWC-C. Remark that the largest improvement is obtained on T-PR which is the test-set used to train the LDA-trained beat template. These differences are however not statistically significant.

We now evaluate the results of downbeat-tracking.

Best parameters for BPIB probability based on chroma variation: For 3 over 4 test-sets, the use of a window duration of $\alpha = 2$ (making the assumption that chords change twice per bar) leads to better results than $\alpha = 4$ (chords change once per bar): FMeas=0.68 and 0.53 for T-PR, 0.76 and 0.78 for T-RWC-P, 0.46 and 0.40 for T-RWC-J, 0.35 and 0.31 for T-RWC-C. For T-PR, the difference is statistically significant. For all test-sets, the use of the "1-cosine" distance leads to better results than the use of the symmetries Mahalanobis distance: FMeas=0.68 and 0.44 for T-PR, 0.76 and 0.49 for T-RWC-P, 0.46 and 0.31 for T-RWC-J, 0.35 and 0.23 for T-RWC-C. These differences are statistically significant for the four test-sets. This result is surprising since the "1-cosine" distance does not take into account the inherent chroma variation inside $L_{i,j}$ and $R_{i,j}$. The bad results obtained with the Mahalanobis distance may be explained by the fact that $L_{i,j}$ and $R_{i,j}$ are too short to reliably estimate the covariance matrices.

BPIB probability based on spectral balance: The results obtained using the spectral balance alone ($p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_3})$) are lower than the ones obtained using chroma variation alone ($p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_2})$): from FMeas=0.68 to 0.49 for T-PR, 0.76 to 0.58 for T-RWC-P, 0.46 to 0.31 for T-RWC-J, 0.35 to 0.22 for T-RWC-C. These differences are statistically significant for the four test-sets (when using $\alpha = 2$ and the "1-cosine" distance for the chroma variation measure). These lower results are not surprising given that the use of spectral balance observation alone does not allow distinguishing the first from the third beat (or the second from the fourth). However, this probability provides a good complement to the chroma variation observation probability as we show now.

Using simultaneously BPIB probability based on chroma variation and spectral balance: For 3 over 4 test-sets, the simultaneous use of the two VPIB ($p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_2}; \underline{o_3})$) allows to further increase the results: FMeas=0.74 for T-PR, 0.8 for T-RWC-P, 0.47 for T-RWC-J, 0.34 for T-RWC-C. However, for none of the test-set, the differences (with the use of chroma alone) are statistically significant. The increase is larger when the file duration is short (T-PR). This can be explained by the fact that BPIB probability based on chroma variation necessitates long duration observation which is not the case of BPIB probability based on spectral balance. Hence a large increase for short duration files. The increase also mainly occurs for files belonging to the Pop and Rock music genre (T-PR and T-RWC-P). This can be explained by the fact that BPIB probability based on "spectral balance" makes the underlying assumption that a "kick/ snare/ kick/ snare" rhythm pattern exists in the signal, which is not the case in Jazz and Classical music.

Downbeat estimation (Viterbi against Chord detection): We

finally compare the results obtained with our complete Viterbi model (Viterbi LDA CHRO ($\alpha = 2$ cos) + SPEC) to the results obtained using the "Chord detection" algorithm of [35]. For 3 over 4 test-sets, the proposed algorithm allows to improve the downbeat-tracking results: FMeas=0.74 and 0.64 for T-PR, 0.8 and 0.81 for T-RWC-P, 0.47 and 0.44 for T-RWC-J, 0.34 and 0.32 for T-RWC-C. Only for T-PR, this difference is statistically significant.

Variations among test-set: As one can observe, the performances of beat-marking are best for the T-PR (FMeas=0.93) and T-RWC-P (0.84) than for the more complex Jazz rhythm of T-RWC-J (0.57) or the time-variable tempo of Classical music of T-RWC-C (0.42). The same can be observed for the downbeat marking (0.74, 0.8, 0.47, 0.34).

TABLE II
BEAT AND DOWNBEAT ESTIMATION RESULTS FOR T-KLA [8] AND T-HAI [13] TEST-SET.

		Cont. Requ. correct	Indiv. Est. correct	Cont. Requ. accept d/h	Indiv. Est. accept d/h
	F-Meas(0.1)	CMLc	CMLt	AMLc	AMLt
BEAT Klapuri Test-Set					
Klapuri et al. (NC)		**59**	64	**73**	80
Viterbi Simple no-DB	0,64	55,45	62,76	68,84	79,95
Viterbi LDA no-DB	**0,67**	**57,03**	**65,50**	**69,96**	**82,86**
Viterbi LDA CHRO/SPEC	**0,66**	55,23	64,64	67,68	81,99
DOWNBEAT Klapuri Test-Set					
Klapuri et al. (NC)		46	47	54	55
Viterbi LDA CHRO/SPEC	0,62	**60,89**	**62,18**	**77,30**	**78,96**
BEAT Hainsworth Test-Set					
Klapuri et al. (NC)		55,7	62,4	**70**	**80**
Davies Plumbey		54,8	61,2	68,1	78,9
Viterbi Simple no-DB	0,60	53,33	59,96	69,90	80,20
Viterbi LDA no-DB	**0,63**	**54,67**	**62,82**	**70,28**	**83,08**
Viterbi LDA CHRO/SPEC	0,62	53,39	61,52	68,69	82,26

D. Comparison to other works

1) Evaluation using Klapuri [8] test-set: In Table II, we present the results of beat and downbeat-tracking using the test-set used in [8]. For comparison, we indicate the results published in [8] for beat and downbeat-tracking.

Concerning beat-tracking, we first compare the use of the "simple" to the "LDA-trained" beat-template using the model $p_{obs}(t_i \in \{\beta\}|\underline{o_1})$. As for the previous test-sets, the use of the "LDA-trained" beat-template provides an improvement of beat-tracking which is statistically significant here: FMeas=0.64 and 0.67. We then compare the performances of beat-tracking obtained using the model $p_{obs}(t_i \in \{\beta\}|\underline{o_1})$ and the model $p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_2}, \underline{o_3})$ (simultaneous estimation of beat and downbeat positions). Surprisingly, there is a small decrease of performances of beat-tracking when using the complete system: FMeas=0.67 and 0.66. We discuss this in details in part V-E2. This difference is however not statistically significant. For the criteria for which temporal continuity is not required (CMLt and AMLt), the performances of our Viterbi-LDA algorithm ($p_{obs}(t_i \in \{\beta\}|\underline{o_1})$ model) are slightly higher than that of [8]: from CMLt= 64 to 65.5, from AMLt= 80 to 83. For the criteria for which temporal continuity is required (CMLc and AMLc), the performances of our algorithm are lower than that of [8].

Concerning downbeat-tracking, our system ($p_{obs}(t_i \in \{\beta\}|\underline{o_1}, \underline{o_2}, \underline{o_3})$) achieves a large improvement over the results published in [8], this for all criteria: CMLc=46 to 61, CMLt=47 to 62, AMLc=54 to 77, AMLt=55 to 79[21].

2) Evaluation using Hainsworth [13] test-set: In Table II, we present the results of beat-tracking using the test-set used in [13], [9] and [52]. We compare our results to the results recently published in [52]: - "Klapuri et al. (NC)" refers to the non-causal algorithm of [8] and - "Davies and Plumbley" refers to the non-causal algorithm of [9].

Again, for this test-set, the LDA-trained beat-template achieves higher results than the "simple" beat-template: FMeas=0.60 and 0.63. However, this difference is not statistically significant. Again, the performances obtained for beat-tracking using the whole system $p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_2}, \underline{o_3})$ are slightly lower than the ones obtained using the beat only part of it $p_{obs}(t_i \in \{\beta\}|\underline{o_1})$. Considering the criteria CMLc, CMLt, AMLc, the results obtained by our system are comparable to the ones by the algorithm of [8]. Our system achieves slightly higher results than [8] for the AMLt criteria: AMTt = 80 and 83.

E. Analysis of errors

1) Beat estimation errors: In this part, we give a short analysis of the errors encountered with the proposed model. For this, we consider the results of the evaluation on the T-KLA test-set (which is the largest test-set and has the most varied content). Considering that our system takes as input an estimated tempo[22], we only consider the tracks for which this tempo has been correctly identified (the ones for which $\left|(\log_2(\frac{estimated\_tempo}{annotated\_tempo})\right| < 0.1$)[23], i.e. we consider 371 over the 474 tracks. For the remaining 103 track, given that the tempo has been wrongly estimated, the following marking process also fails. Among the 371 tracks, we then consider the ones for which the marking has largely failed, i.e. for which the F-measure(PW=0.1) is below 0.6. This corresponds to 93 tracks using the beat-only model $p_{obs}(t_i \in \{\beta\}|\underline{o_1})$ (denoted simply by M-$\{\beta\}$ in the following) and to 100 tracks using the beat/downbeat model $p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_2}, \underline{o_3})$ (denoted by M-$\{\beta_j\}$).

We propose a rough categorization of the type of errors occurring for these 100 tracks. 20 of them belong to the "Electro/Dance" category for which either the signal is highly compressed (hence the hi-hat has a very significant energy resulting in marking the counter-beats) or there is a long-duration break without onsets (the break is only based on synthesizer sounds). 20 of them belong to the "Classical music"

[21]The reader can be surprised to see higher results for downbeat than for beat estimation. This is explained by the fact that - the Klapuri beat-set has 474 tracks, while the downbeat-set has only 320 tracks. - the beat and downbeat annotations have been made independently (annotated downbeats are not necessarily among the annotated beats).

[22]We refer the reader to [32] for an analysis of typical tempo estimation errors of our system.

[23]It should be noted that this measure does not consider the accuracy of the estimation of time-varying tempo over time but only uses the median value of the tempo over time.

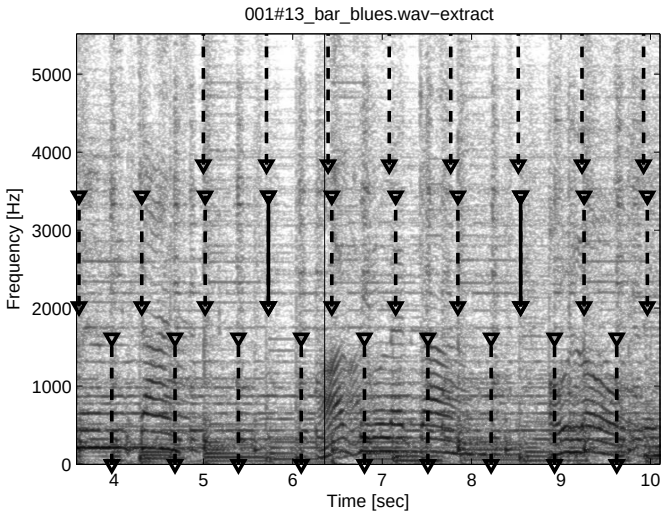

Fig. 10. *Spectrogram-representation and (super-imposed to it) annotated beat positions (dotted vertical lines on the top-part), estimated beat positions using M-$\{\beta_j\}$ (on the middle part) and estimated beat positions using M-$\{\beta\}$ (on the bottom part) on signal ["001 13 bar blues" from the T-KLA test-set].*

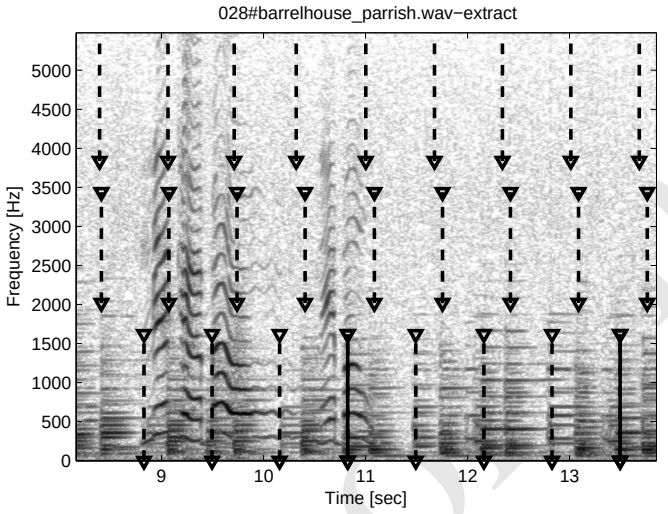

Fig. 11. *Spectrogram-representation and (super-imposed to it) annotated beat positions (dotted vertical lines on the top-part), estimated beat positions using M-$\{\beta\}$ (on the middle part) and estimated beat positions using M-$\{\beta_j\}$ (on the bottom part) on signal ["028 barrelhouse parrish" from the T-KLA test-set].*

category with time-variable tempo and with fuzzy onsets (slow attacks or slow note transitions). 3 belong to the "Expressive performances of piano/guitar" category with clear onsets but very fast tempo variations. 15 belong to the "Jazz Swing" category and 8 to the "Old Blues with shuffle" category. In both cases, the swing or shuffle disturbs the marking either globally or locally in time. Also, in old recordings, the mixing of the vocal part (which is not always clearly gridded to the tempo) in front makes the marking difficult. 9 belong to the Hard-Rock/Metal with dominant electric guitars (creating many onsets) and double-kick. 6 belong to the Latino music category with complex syncopated rhythms. The remaining 16 titles belong to less defined categories.

2) Effect of the simultaneous beat and downbeat estimation on beat estimation: According to Table I and Table II, for a task of beat-tracking, there is no advantage of using the beat/downbeat model M-$\{\beta_j\}$ over using the beat-only model M-$\{\beta\}$. However, a deeper analysis of the results shows that the tracks for which the model M-$\{\beta\}$ fails is not a subset of the one for which M-$\{\beta_j\}$ fails. In other words, they are tracks for which M-$\{\beta_j\}$ succeeds and not M-$\{\beta\}$ and tracks for which M-$\{\beta\}$ succeeds and not M-$\{\beta_j\}$. As we illustrate below, M-$\{\beta_j\}$ uses an underlying music model (of chord changes and spectral balance pattern) which fit or not the content of a given music audio signal.

In Figure 10, we show a track for which the beats have been correctly estimated using M-$\{\beta_j\}$ but not using M-$\{\beta\}$. In many tracks for which M-$\{\beta_j\}$ succeeds and not M-$\{\beta\}$, there exist clear onsets on the counter-beat positions (in the example of Figure 10 the guitar is playing on the counter-beats). The model M-$\{\beta\}$ focuses on those and marks them erroneously as the beat positions. As we see in the figure, the music audio signal has chord changes on the downbeats and the typical kick/snare/kick/snare sequence over beat positions. This corresponds to the music model of M-$\{\beta_j\}$, hence the algorithm succeeds to correctly detect the beat positions.

In Figure 11, we show a track for which the beats have been correctly estimated using M-$\{\beta\}$ but not using M-$\{\beta_j\}$. This is because the audio signal does not correspond to the music model of M-$\{\beta_j\}$. In many tracks for which M-$\{\beta\}$ succeeds and not M-$\{\beta_j\}$, there exist clear onsets on the beat positions but also events of strong energy at other positions which are fuzzy onsets but contain one of the bpib characteristic. Because M-$\{\beta\}$ uses the beat-only probability $p_{obs}(t \in \{\beta\}|\underline{o}(t))$ (which highlights strongly onsets but weakly fuzzy-onsets), M-$\{\beta\}$ succeeds to correctly mark the beats. In the opposite, because the fuzzy onsets creates harmonic changes (the vocal entrances in the figure), and because M-$\{\beta_j\}$ includes the chord change probability $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}(t))$, M-$\{\beta_j\}$ focuses on those and marks them erroneously as the beat positions.

3) Downbeat estimation errors: The large majority of downbeat estimation errors (for the set of tracks for which the beat positions have been correctly estimated) are coming from confusion between the downbeat (β_1) and the 3rd beat (β_3). In most cases, this confusion occurs when the chords change twice per measure or when there is no (or a very small) variation of chords.

VI. Computational requirements and runtime

The proposed algorithm for beat and downbeat-tracking has been integrated into C++ (ircambeat software). This software also includes the tempo/meter estimation algorithm proposed in [32]. Using the beat only model ($p_{obs}(t_i \in \{\beta\}|\underline{o_1})$), the maximum peak of memory load is 14MB by minute of audio to process (56MB to process a 4 minutes track). On an Intel Xeon 2.3 GHz CPU, the computation time is 1.75s by minute of audio to process (7s to process a 4 minutes track). Using the simultaneous beat and downbeat model ($p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_2}, \underline{o_3})$), the maximum peak of

memory load is 19MB by minute of audio to process and the computation time is 1.9s by minute of audio to process.

VII. Conclusion and Future works

In this paper we proposed a probabilistic framework for simultaneous beat and downbeat-tracking from an audio signal given estimated tempo and meter as input.

We proposed a hidden Markov model formulation in which hidden states are defined as "time t is a beat in a specific beat-position-inside-a-bar". Since times are part of the hidden states definition, we proposed a "reverse" Viterbi decoding algorithm which decodes times (and their associated beat-position-inside-a-bar) over beat-numbers. The beat observation probabilities are obtained by using beat-templates. We proposed the use of Linear Discriminant Analysis to compute the most discriminant beat-template. We showed that the use of this LDA-trained beat-template allows an improvement of beat-tracking results for 4 over the 6 test-sets used in our evaluation. For the "Klapuri" test-set, this difference is statistically significant. It is important to note that the "Klapuri" test-set is the largest test-set and was not part of the development of our system.

The beat-position-inside-a-bar (bpib) allows deriving simultaneously beat and downbeat position. We proposed two bpib observation probabilities. The first probability is based on analyzing the variation of chroma vector over time. We studied two window lengths for their computation (corresponding to the assumptions that chords change twice or once per bar) and two distances for their comparison (the "1-cosine" and the symmetries Mahalanobis distances). The best results were obtained using a window length of two beats and a "1-cosine" distance. The second probability is based on analyzing the temporal pattern of the spectral balance. The inclusion of this second probability allows increasing further the downbeat-tracking results.

We compared the results obtained by our new algorithm to the ones obtained with our previous P-sola beat-tracking algorithm (as used in MIREX-05 contest) [34]. Results show a large improvement of the beat-tracking results which is statistically significant for 2 over 4 test-sets. We then compared the results obtained by our new algorithm to the ones obtained with our previous Chord-based downbeat-tracking algorithm [35]. Results show an improvement of the downbeat-tracking results for 3 over 4 test-sets which is statistically significant for the "PopRock extract" test-set.

We compared our results to the one obtained in [8] [13] and [9] using the same test-sets and evaluation measures. For the "Klapuri" test-set, our new algorithm allows to slightly improve the results of beat-tracking for the CMLt and AMLt measures (which do not require temporal contiguity), however this is not the case for the CMLc and AMLc measures (which require temporal contiguity). Our algorithm seems therefore to suffer from temporal discontinuities in the marking. This may be due to the large transition probability assigned to $p_{trans}(t_{i'} \in \{\beta_{j'}\}|t_i \in \{\beta_j\})$ in our experiment. Concerning downbeat-tracking, our algorithm largely improves over the results published in [8] for all criteria. For the "Hainsworth" test-set, the results obtained by our algorithm are close to the ones published in [8]. Our algorithm slightly improves the results considering the AMLt measure (which consider octave errors as correct and do not consider temporal continuity).

We also submitted our tempo and beat-marking system to the MIREX-09 and MIREX-10 Audio Beat Tracking contest [23] [24]. Only beat-tracking performances were measured in these contests. We therefore submitted the system corresponding to $p_{obs}(t_i \in \beta|\underline{o_1}(t))$. We tested four configuration of the tempo estimation stage of [32] (variable-over-time or constant-over-time tempo estimation, meter estimated or forced to 4/4). In 2009, the periodicity measure was the one proposed in [32], in 2010 we tested the use of the hybrid axes DFT/ACF (haDFTACF) periodicity measure proposed in [53]. Two test-sets were used: the "McKinney Collection" [54] [55] and the "Sapp's Mazurka Collection". We refer the reader to the MIREX web sites[24] for details on the test-sets, on the performance measures and on the results. In 2009, for the "McKinney Collection" test-set, our system ranked first for most criteria. In 2010, the performances of our system were lower than in 2009 and the system only ranked second. Comparing the best-results obtained with this test-set in MIREX-06 and MIREX-09 shows an improvement of the results: Dixon (the best result in 2006) reached a P-score of 0.575 in 2006; we reached 0.592 in 2009. The results obtained on the "Sapp's Mazurka Collection" test-set are not as good. These lower results can be partly explained by the large analysis window (8s.) used by our system for periodicity analysis

Considering the results obtained and the adaptability to include new observation probabilities, the proposed probabilistic formulation is promising. The computation time and memory cost is however higher than other methods. However, the method can be highly optimized when implementing it. The C++ version of this algorithm was for example the fastest algorithm in the MIREX-09 contest.

A deep analysis of the results obtained with our proposed model of simultaneous beat and downbeat estimation[25] showed that the simultaneous estimation can be inefficient in case the music model underlying $p_{obs}(t_i \in \{\gamma_j\}|\underline{o_2}, \underline{o_3})$ (chord changes on the downbeats and presence of a low/high/low/high frequency pattern) does not fit to the content of the music audio signal. Future works will therefore concentrate on extending this music model. We will also concentrate on adding new type of observations probabilities for the bpib probability such as the detection of (relative) silences. The LDA-trained beat template used here was the one trained on the PopRock test-set. This PopRock template was applied to Jazz and Classical music. Ideally, one would choose the most appropriate LDA-trained beat-template for the music genre under consideration. Further works will therefore also concentrate on integrating

[24]MIREX 2009: `http:// www.music-ir.org/ mirex/ 2009/ index.php/ Audio_Beat_Tracking_Results` MIREX-2010: `http:// nema.lis.illinois.edu/ nema_out/ mirex2010/ results/ abt/ mck/` and `http:// nema.lis.illinois.edu/ nema_out/ mirex2010/ results/ abt/ maz/`

[25]We compared the beat estimation obtained using the beat-only model $p_{obs}(t_i \in \{\beta\}|\underline{o_1})$ to the ones obtained using the simultaneous beat and downbeat model $p_{obs}(t_i \in \{\beta_j\}|\underline{o_1}, \underline{o_2}, \underline{o_3})$

automatic music genre estimation to our system in order to choose the most appropriate beat-template. Finally, our current system is composed of two independent parts: tempo and meter estimation on one side, beat and downbeat estimation on the other side. Both parts use a hidden Markov model formulation, we will therefore study their simultaneously estimation using a single framework as did for example Laroche in [36].

VIII. Acknowledgments

This work was partly supported by the "Quaero" Program funded by Oseo French State agency for innovation. Many thanks to Frederic Cornu for careful optimization and debugging of the code. Many thanks to Anssi Klapuri, Stephen Hainsworth and Matthew Davies for sharing their test-sets and the beat-tracking evaluation toolbox.

References

[1] J. London, *Hearing in Time: Psychological Aspects of Musical Meter*. New York: Oxford University Press, 2004.
[2] F. Gouyon and S. Dixon, "A review of rhythm description systems," *Computer Music Journal*, vol. 29, no. 1, pp. 34–54, 2005.
[3] A. Marsden, "Special issue on algorithms for beat tracking and tempo extraction," *Journal of New Music Research*, vol. 36, no. 1, pp. 1–60, 2007.
[4] M. Goto and Y. Muraoka, "Music understanding at the beat level real-time beat tracking for audio signals," in *Proc. of IJCAI (Int. Joint Conf. on AI) / Workshop on CASA*, 1995, pp. 68–75.
[5] ——, "Real-time rhythm tracking for drumless audio signals - chord change detection for musical decisions," in *Proc. of IJCAI (Int. Joint Conf. on AI) / Workshop on CASA*, 1997, pp. 135–144.
[6] S. Dixon, "Evaluation of audio beat tracking system beatroot," *Journal of New Music Research*, vol. 36, no. 1, pp. 39–51, 2007.
[7] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 588–601, 1998.
[8] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 342–355, 2006.
[9] M. Davies and M. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Trans on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1009–1020, 2007.
[10] M. Goto and Y. Muraoka, "Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions," *Speech Communication*, vol. 27, pp. 311–335, 1999.
[11] T. Jehan, "Creating music by listening," PHD Thesis, Massachusetts Institute of Technology., 2005.
[12] A. Cemgil and B. Kapen, "Monte carlo methods for tempo tracking and rhythm quantization," *Journal of Artificial Intelligence Research*, vol. 18, pp. 45–81, 2003.
[13] S. Hainsworth and M. Macleod, "Beat tracking with particle filtering algorithms," in *Proc. of IEEE WASPAA*, New Paltz, NY, 2003.
[14] J. Laroche, "Estimating tempo, swing and beat locations in audio recordings," in *Proc. of IEEE WASPAA*, New Paltz, NY, USA, 2001.
[15] D. Ellis, "Beat tracking by dynamic programming," *J. New Music Research*, vol. 6, no. Special Issue on Beat and Tempo Extraction, pp. 51–60, 2007.
[16] J. Seppanen, "Tatum grid analysis of musical signals," in *Proc. of IEEE WASPAA*, New Paltz, NY, USA, 2001.
[17] F. Gouyon, "A computational approach to rhythm description," PHD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2005.
[18] D. Eck and N. Casagrande, "Finding meter in music using an autocorrelation phase matrix and shannon entropy," in *Proc. of ISMIR*, London, UK, 2005.
[19] P. Grosche and M. Muller, "A mid-level representation for capturing dominant tempo and pulse information in music recordings," in *Proc. of ISMIR*, Kobe, Japan, 2009.
[20] M. Davies, N. Degara, and M. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," Queen Mary University of London, Tech. Rep. Technical Report C4DM-TR-09-06, 2009.
[21] MIREX, "Audio tempo extraction," 2005.
[22] ——, "Audio beat tracking contest," 2006.
[23] ——, "Audio beat tracking contest," 2009.
[24] ——, "Audio beat tracking contest," 2010.
[25] H. Allan, "Bar lines and beyond - meter tracking in digital audio," Master Thesis, University of Edinburgh, 2004.
[26] T. Jehan, "Downbeat prediction by listening and learning," in *Proc. of IEEE WASPAA*, New Paltz, NY, 2005.
[27] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, no. 2, pp. 159–171, 2001.
[28] D. Ellis and J. Arroyo, "Eigenrythms: Drum pattern basis sets for classification and generation," in *Proc. of ISMIR*, Barcelona, Spain, 2004.
[29] M. Davies and M. Plumbley, "A spectral difference approach to downbeat extraction in musical audio," in *Proc. of EUSIPCO*, Florence, Italy, 2006.
[30] M. Gainza, D. Barry, and E. Coyle, "Automatic bar line segmentation," in *Proc. of AES 123rd Convention*, New York, NY, USA, 2007.
[31] F. Gouyon and P. Herrera, "Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors," in *Proc. of AES 114th Convention*, Amsterdam, The Netherlands, 2003.
[32] G. Peeters, "Template-based estimation of time-varying tempo," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 158–158, 2007, doi:10.1155/2007/67215.
[33] ——, "Modeles et modelisation du signal sonore adaptes a ses caracteristiques locales," PHD Thesis, Universite Paris VI, 2001.
[34] ——, "Beat-marker location using a probabilistic framework and linear discriminant analysis," in *Proc. of DAFX*, Como, Italy, 2009.
[35] H. Papadopoulos and G. Peeters, "Simultaneous estimation of chord progression and downbeats from an audio file," in *Proc. of IEEE ICASSP*, Las Vegas, USA, 2008.
[36] J. Laroche, "Efficient tempo and beat tracking in audio recordings," *J. Audio Eng. Soc.*, vol. 51, no. 4, pp. 226–233, 2003.
[37] H. Papadopoulos and G. Peeters, "Joint estimation of chords and downbeats from an audio signal," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, 2010.
[38] A. Eronen and A. Klapuri, "Music tempo estimation with k-nn regression," *IEEE Trans on Audio, Speech and Language Processing*, vol. 18, no. 1, pp. 50–57, 2010.
[39] N. Whiteley, A. Cemgil, and S. Godsill, "Bayesian modeling of temporal structure in musical audio," in *Proc. of ISMIR*, Victoria, Canada, 2006, pp. 29–34.
[40] L. Rabiner, "A tutorial on hidden markov model and selected applications in speech," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, 1989.
[41] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.
[42] S. Dixon, F. Gouyon, and G. Widmer, "Towards characterisation of music via rhythmic patterns," in *Proc. of ISMIR*, Barcelona, Spain, 2004, pp. 509–516.
[43] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical, and jazz music databases," in *Proc. of ISMIR*, Paris, France, 2002, pp. pp. 287–288.
[44] G. Wakefield, "Mathematical representation of joint time-chroma distributions," in *Proc. of SPIE conference on Advanced Signal Processing Algorithms, Architecture and Implementations*, Denver, Colorado, USA, 1999, pp. 637–645.
[45] G. Peeters, "Chroma-based estimation of musical key from audio-signal analysis," in *Proc. of ISMIR*, Victoria, Canada, 2006, pp. 115–120.
[46] A. Cemgil, B. Kappen, P. Desain, and H. Honing, "On tempo tracking: Tempogram representation and kalman filtering," *Journal of New Music Research*, vol. 29, no. 4, pp. 259–273, 2001.
[47] M. Goto, "Issues in evaluating beat tracking systems," in *Proc. of IJCAI*, 1997.
[48] M. McKinney, D. Moelants, M. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," *Journal of New Music Research*, vol. 36, no. 1, pp. 1–16, 2007.
[49] S. Hainsworth, "Techniques for the automated analysis of musical audio," PHD Thesis, Cambridge University, 2004.
[50] M. Goto, "Rwc (real world computing) music database," 2005.
[51] ——, "Aist annotation for the rwc music database," in *Proc. of ISMIR*, Victoria, Canada, 2006, pp. pp.359–360.
[52] A. Stark, M. Davies, and M. Plumbley, "Real-time beat-synchronous analysis of musical audio," in *Proc. of DAFX*, Come, Italy, 2009.
[53] G. Peeters, "Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal," *submitted to IEEE. Trans. on Audio, Speech and Language Processing*, 2010.

[54] M. McKinney and D. Moelants, "Deviations from the resonance theory of tempo induction," in *Conference on Interdisciplinary Musicology*, Graz, Austria, 2004.

[55] 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*. Evanston, IL, 2004.

Geoffroy Peeters Geoffroy Peeters received his Ph.D. degree in computer science from the Universite Paris VI, France, in 2001. During his Ph.D., he developed new signal processing algorithms for speech and audio processing. Since 1999, he works at IRCAM (Institute of Research and Coordination in Acoustic and Music) in Paris, France. His current research interests are in signal processing and pattern matching applied to audio and music indexing. He has developed new algorithms for timbre description, sound classification, audio identification, rhythm description, automatic music structure discovery, and audio summary. He owns several patents in these fields. He has also coordinated indexing research activities for the Cuidad, Cuidado, and Semantic HIFI European projects and is currently leading the music indexing activities in the Quaero Oseo project. He is one of the co-authors of the ISO MPEG-7 audio standard.

Helene Papadopoulos Helene Papadopoulos was born in Paris, France, in 1983. She graduated in 2006 from ENSEA in Paris ("Ecole Nationale Superieure de l'Electronique et de ses Applications"), a leading French Engineering School specialized in Electronics, Signal Processing, Computer Science and Communication. The same year, she received the M.Sc. degree in Image, Signal Processing and Artificial Intelligence from the University of Cergy-Pontoise. She is currently pursuing a Ph.D. degree in the Analysis/Synthesis team at IRCAM in Paris. In parallel to her scientific studies, she pursues musical studies at the professional level. Her research interests include signal processing, machine learning, music perception and cognition, music content processing, classification and music information retrieval.