A CONDITIONAL RANDOM FIELD SYSTEM FOR BEAT TRACKING

Thomas Fillon[⋆] *Cyril Joder*[†] *Simon Durand*[‡] *Slim Essid*[‡] [*]

[⋆] Parisson, 16 rue Jacques Louvel-Tessier, 75010 Paris, France
[†] European Patent Office, 80298 Munich, Germany
[‡] Institut Mines-Telecom, Telecom-ParisTech, CNRS LTCI, 37/39 rue Dareau, 75014 Paris, France

ABSTRACT

In the present work, we introduce a new probabilistic model for the task of estimating beat positions in a musical audio recording, instantiating the Conditional Random Field (CRF) framework. Our approach takes its strength from a sophisticated temporal modeling of the audio observations, accounting for local tempo variations which are readily represented in the CRF model proposed using well-chosen potentials.

The system is experimentally evaluated by studying its performance on 3 datasets of 1394 music excerpts of various western music styles and comparatively to 4 reference systems in the light of 6 reference evaluation metrics. The results show that the proposed system tracks perceptively coherent pulses and is very effective in estimating the beat positions while further work is needed to find the correct salient tempo.

Index Terms— Music Information Retrieval, Beat tracking, Conditional Random Fields.

1. INTRODUCTION

Beat tracking is a commonly addressed yet still challenging task for researchers in Music Information Retrieval (MIR) and related fields. It entails inferring, from the audio signal observation, subjective musical notions such as instantaneous tempo and beat positions inside the musical piece. It holds an important and essential position for many automatic audio and music analysis tasks where the extraction and exploitation of information related to the musical rhythm is crucial.

In fact, the problem of estimating beat positions in a musical audio recording has been extensively addressed through diverse solutions. A good overview of these different approaches can be found in [1, 2]. The first stage of most beat-tracking systems consists in extracting the rhythmic information through a discrete estimation of the musical onset positions or through the computation of a continuous onset likelihood function, here referred to as the *onset detection function*. Given that observation, the beat-period and the beat position can be estimated either successively or jointly, based on diverse deterministic or probabilistic methods.

A common approach for the estimation of the beat-period is to construct an Inter-Onset-Interval histogram or to use a spectral analysis method. Another popular approach is to rely on a bank of resonating comb-filters [3, 4, 5], which has the advantage of providing both beat-period and beat-phase information.

When probabilistic models are considered, the tracking of the beat-period and beat-phase information can be done through dynamic programming [6, 7], particle filtering [8], hidden-Markov model[5, 2].

In this paper, we introduce a novel probabilistic model for the beat tracking problem, instantiating the powerful Conditional Random Field (CRF) framework. Our approach is largely inspired by the audio-to-score alignment system described in [9, 10], which is here adapted to handle beat labels.

This model has the essential advantage of providing a proper formulation of the task which enables us to directly solve the problem of tracking the beat-period and beat-position as a probabilistic sequence labeling problem, where the goal is to determine the most probable sequence of beat-period and beat-position labels at every time instant. Furthermore, the proposed CRF beat-tracking system can be easily improved and extended by incorporating new observations

The outline of the paper is the following. In the next Section, we formalize the beat tracking problem and formulate it as a sequence labeling problem. Then Section 3 briefly recalls the general CRF framework and describes our model for beat tracking. Subsequently, we present an experimental study that we have conducted to assess the performance of our proposal, comparing it to 4 different reference systems, before we suggest some conclusions in Section 5.

2. THE BEAT TRACKING PROBLEM

As previously indicated, a beat-tracking system has to both estimate the beat-periods and the beat-positions. We propose a new probabilistic system for their joint estimation given the observation of features relating to onset and period likeli-

[*]The first two authors performed the work while at Telecom-ParisTech and this work was supported under the research programme Quaero, funded by OSEO, the French State agency for innovation.

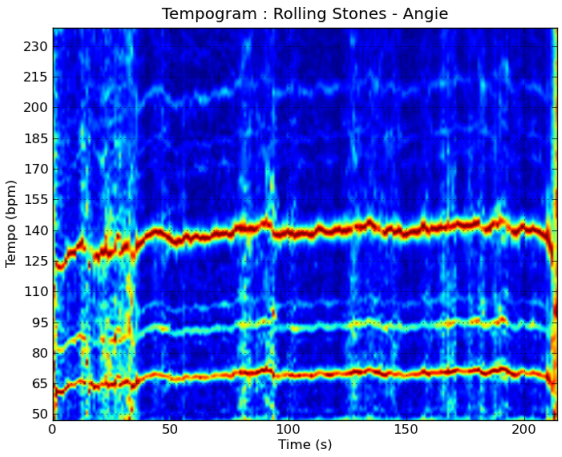

Fig. 1. Example of tempogram for the song "Angie" by the *Rolling Stones*. The tempo is fluctuating around 65 beats per minute (BPM).

hoods over time.

The front-end system extracting these feature observations relies on the method proposed by Alonso in [11, 12]. The features are extracted on a frame-by-frame basis and consist of the following:

The Onset detection function which is computed from the spectral energy flux of the input audio signal over time. This produces a vector of onset likelihoods (one value per time frame).

The Tempogram which gives an estimation of the salience of periodicity candidates extracted from the detection function using pitch detection techniques. This produces a matrix of periodicity likelihoods over time (a vector of values per time frame) as illustrated in Figure 1.

For each signal frame, given the feature observations, we propose to handle the beat-tracking problem as a *sequence labeling* problem: to each frame is assigned a discrete beat-period label and a discrete beat-phase label (*i.e.* the beat position inside a period), also named occupancy, as illustrated and further detailed in Figure 2.

Let $\mathbf{Y}$ denote the sequence of period and occupancy values, and $\mathbf{X}$ the sequence of feature observations. Assuming an appropriate model $p(\mathbf{Y}|\mathbf{X};\theta)$ of the posterior probabilities of target labels $\mathbf{Y}$ given features $\mathbf{X}$, parametrised by θ, the beat-tracking problem can be seen as the one of determining the label sequence $\hat{\mathbf{Y}}$ such that:

$$\hat{\mathbf{Y}} = \underset{\mathbf{Y}}{\operatorname{argmax}}\, p(\mathbf{Y}|\mathbf{X};\theta) \tag{1}$$

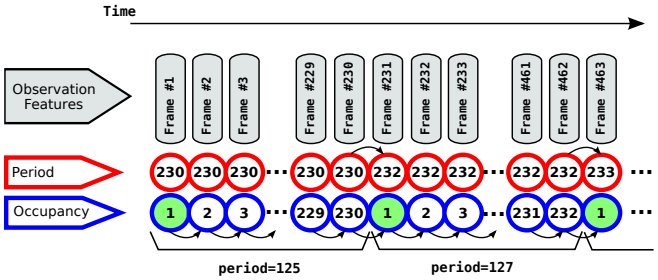

Fig. 2. Beat tracking as a labeling problem. In this example, the observations are extracted at a frame rate of 250Hz and the candidate tempo is around 65 BPM (i.e. a period of 230 frames). Each frame is assigned both a *period* label and an *occupancy* label characterizing the relative frame position inside that period. The frames labeled with occupancy #1 correspond to the beat instant.

3. CRF FOR BEAT TRACKING

We solve the sequence labeling problem inherent to beat tracking using Conditional Random Fields. CRF [13] are a powerful class of discriminative classifiers for structured input–structured output data prediction, which have proven successful in a variety of real-world classification tasks [14, 15]. Compared to Hidden Markov Models (HMM) and more general Bayesian networks, CRFs draw their power from both their discriminative nature and their non-oriented graphical model structure as they model directly the posterior probabilities of a sequence of N labels given feature observations according to:

$$p(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \prod_{n=1}^{N} \underbrace{\Psi(\mathbf{Y},\mathbf{X},n)}_{Transition} \underbrace{\Phi(\mathbf{Y},\mathbf{X},n)}_{Observation};$$

where $\Psi(\mathbf{Y},\mathbf{X},n)$ and $\Phi(\mathbf{Y},\mathbf{X},n)$ are respectively *transition* and *observation potentials* at each time position n, which play a similar role to transition and observation probabilities found in HMM, except that the former are not proper probabilities, hence the need for the normalizing term $Z(\mathbf{X})$ that guarantees that $p(\mathbf{Y}|\mathbf{X})$ is a well defined probability, which sums to 1 [14].

In this work, we use a Markovian form of CRF, where the transition potentials are defined on consecutive labels, in a linear-chain fashion, and observation potentials depend on single labels, so that:

$$p(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \Phi(Y_1,\mathbf{X}) \prod_{n=2}^{N} \Psi(Y_n,Y_{n-1},\mathbf{X})\Phi(Y_n,\mathbf{X}). \tag{2}$$

The output variables (labels)

In our CRF system for beat-tracking, the output variables, $\mathbf{Y}$ are composed of: i) the *occupancy* variables: $\mathbf{D}$, which are

discrete : $D_n \in [1, ..., d_{max}]$; and ii) the *period* variables: $\mathbf{T}$, also discrete : $T_n \in [d_{min}, ..., d_{max}]$);
where d_{min} and d_{max} are positive integers.

The observation sequences are the ones described in Section 2 : i) the onset detection function sequence $\mathbf{S}$ and ii) the tempogram matrix $\mathbf{G}$ (a sequence of vectors of periodicity salience stored column-wise in $\mathbf{G}$). For simplicity, let $G_n[T]$ denote the salience of periodicity T at time frame n.

The observation potentials

The observation potentials, $\Phi(Y_n, \mathbf{X})$, of equation (2) are defined for the period and occupancy labels occurring at frame n, given the whole sequence of observations $\mathbf{X}$.

In the proposed system, we factorize the observation potentials in two distinct terms, $\Phi(Y_n, \mathbf{X}) = \Phi_T(T_n, G_n) \cdot \Phi_D(Y_n, \mathbf{S})$ where $\Phi_T(T_n, G_n) = G_n[T_n]$ is directly derived from the tempogram. For the second term, we infer the beat location by correlating the onset detection function with a parametrized periodic pulse $\Pi_T(t)$ function whose definition is derived from [16, 17] such as:

$$\Pi_T[t] = 1 + \tanh \gamma \left(\cos \left(2\pi \frac{t}{T} \right) - 1 \right), \text{ for } t \in [-2T; 2T] \tag{3}$$

where $\gamma = 2$, T is the considered musical period expressed in samples at the frame rate and t is a time index in samples. This simple operation favors the estimation of beat position at location where the onset detection function is in phase with the periodic pulse.

The second observation potential term is thus:

$$\Phi_D(Y_n, \mathbf{S}) = \Phi_D((T_n, D_n), \mathbf{S}) = S[n - D_n + 1] \star \Pi_{T_n}[n]$$

where $\star$ is the convolution operator.

The transition potentials

The transition potentials, $\Psi(Y_n, Y_{n-1}, \mathbf{X})$, of equation (2) do not depend on the observations $\mathbf{X}$ in the following model and thus equals $\Psi(Y_n, Y_{n-1})$. These potentials are constructed as follows.

First, tempo transitions are only allowed at the beat positions (*i.e.* when occupancy is labeled as 1). Following [18], we assume that the tempo changes are relative rather than absolute and that for example, the probability is the same for doubling the tempo and for halving it. Tempo are thus constrained by the following transition penalty:

$$h_t(T_{n+1}, T_n) = \begin{cases} \exp\left(-\gamma_t \left| \log \frac{T_{n+1}}{T_n} \right|^2 \right), & \text{if } \frac{T_{n+1}}{T_n} \leq 2 \\ \exp\left(-\gamma_t \left| \log 2 \right|^2 \right), & \text{if } \frac{T_{n+1}}{T_n} > 2 \end{cases} \tag{4}$$

controlled by the parameter $\gamma_t > 0$.

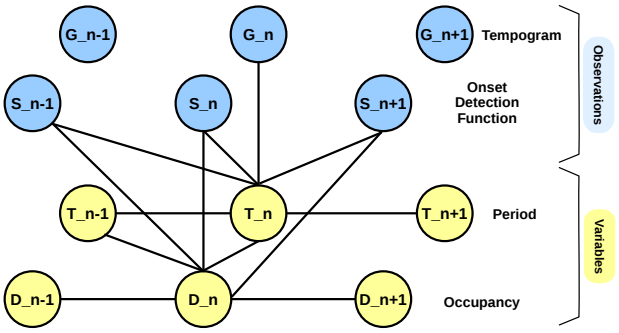

Fig. 3. Undirected graphical model representation for the proposed CRF beat-tracking system at the time frame n. For clarity, the dependencies between the variables (D_n, T_n) and observations S_p outside the range $[n-1, n+1]$ are omitted.

As for the occupancy labels, the transition potentials chosen constrain the occupancy to be incremented by 1 during the whole period and to switch to 1 at the end the period. This very simple constraint forces the coherence between the occupancy and period label values. We thus have:

$$\Psi(Y_n, Y_{n-1}, \mathbf{X}) = \begin{cases} h_t(T_{n+1}, T_n) \cdot \mathbb{1}(D_n = 1) \\ \quad \text{if } D_{n-1} = T_{n-1} \\ \mathbb{1}(D_n = D_{n-1} + 1) \cdot \mathbb{1}(T_n = T_{n-1}) \\ \quad \text{otherwise} \end{cases} \tag{5}$$

where $\mathbb{1}$ is the *indicator function*. The dependency structure of the probabilistic model thus specified is represented in Figure 3.

Decoding

The most probable label sequence for the model (2) can be calculated by the Viterbi algorithm with the same complexity as in a HMM (see [9] for further details about the decoding implementation). However, one of the main advantages of CRF is that they directly model the conditional distribution $p(\mathbf{Y}|\mathbf{X})$ *i.e.* the probability of the target labels given the observation sequence. Also, the CRF framework relaxes the conditional independence assumption of HMMs and thus X_n is not supposed to be independent of all other variables given Y_n and the whole feature sequence $\mathbf{X}$ can be safely used in the calculation of the observation potential $\Phi(Y_n, \mathbf{X})$. For further details on CRF in general see [13, 19] and in the context of MIR research see [9, 20].

Given the above-mentioned properties of the CRF, the most substantial advantage of the proposed system in the context of beat tracking is the good fit between the theoretical model and the problem and the ability to extend this model to incorporate other musical information or priors about the data, in a more flexible way compared to HMM and Bayesian networks.

4. EXPERIMENTAL EVALUATION

We evaluate our system on three different reference datasets. The first is the Ballroom dataset that contains 698 30-second long excerpts of various Dance styles, such as Cha Cha, Quickstep, Samba, Jive, Tango, Rumba, Viennese Waltz and Slow Waltz [1]. The tempo is rather stable within each song but varies a lot across music styles; ranging from an average of 85 BPM in the Slow Waltz subset to a average of 205 BPM in the Quickstep subset. The second dataset is the Klapuri dataset that contains 474 60-second long excerpts of Classical, Electronic/Dance, Jazz/Blues, Hip Hop/Rap, Rock/Pop and Soul/RnB/Funk music styles [5]. The tempo can change considerably inside a song, especially in Classical and Jazz music. The third dataset is the Hainsworth dataset that contains 222 60-second long excerpts of Dance, Rock/Pop, Jazz, Folk, Classical and Choral music styles [8]. The tempo can here also change considerably inside a song and has a wide range inside each music style, ranging from 60 BPM to 160 BPM and even 200 BPM in Jazz music.

We use two sets of standard evaluation measures, the F-measure [21] and the continuity-based evaluation CMLc, CMLt, AMLc and AMLt [5] [8] [22]. The F-measure is the harmonic mean of precision and recall rates. The precision is the ratio of correctly detected beats among all detected beats and the recall is the ratio of correctly detected beats on annotated beats. Correct detections occur when an estimated beat position falls within a tolerance window centered around a ground-truth beat position. We use two types of tolerance window. The first is the standard fixed 70 ms tolerance window as used in the MIREX audio beat tracking evaluation initiative for example[2]. The second is a relative beat period precision window of 0.1 [2]. The tolerance window is thus a percentage of the local beat period. This is interesting as in the case of a wide range of tempo in the datasets, a fixed tolerance window can be too restrictive for slow paced songs and too permissive for fast paced songs. The continuity-based evaluation measures are used with a 17.5% precision window. While the F-measure tracks the instantaneous performance of a beat tracker, those measures track the correct estimations of continuous parts of the signal. CMLc tracks the longest continuous sequence of correctly estimated beats at the correct metrical level, while CMLt tracks the sum of all those sequences. AMLc tracks the longest continuous sequence of correctly estimated beats but allows octave errors[3]. AMLt also allows octave errors and tracks the sum of continuous sequences correctly estimated. These measures are computed from the evaluation toolbox in [23].

We compare the performance of our beat tracking system to 4 other reference methods [7, 24, 25, 2]. Results are given in Table 1 and show that our proposal is competitive with the others. Further, one can see a trend across datasets, that is [25] produces the best results in terms of F-measure and correct metrical level, while our system produces the best results with the continuity based evaluation measures that allow octave errors (AMLc and AMLt). This indicates that our system is able to finely follow the tempo variations and track the beats in a continuous fashion. However, it is prone to errors relating to situations where the tempo is estimated to be twice as fast or twice as slow as the one annotated.

Method	Fmeas 70 ms	Fmeas 0,1	CMLc	CMLt	AMLc	AMLt
Ballroom dataset						
CRF	74.6	72.9	49.8	50.2	**87.3**	**88.6**
[24]	73.9	69.9	55.9	57.7	84.7	87.3
[25]	**80.8**	**77.7**	**60.0**	**61.6**	82.7	85.8
[7]	69.7	62.8	27.9	31.8	67.6	81.8
[2]	77.4	71.4	53.8	55.5	85.9	87.9
Klapuri dataset						
CRF	69.6	67.1	46.8	50.5	**76.4**	**87.1**
[24]	67.9	64.5	50.8	58.4	66.2	79.2
[25]	**72.9**	**69.9**	**54.1**	60.6	67.8	79.7
[7]	61.8	54.6	14.7	20.1	38.1	69.4
[2]	70.2	66.0	51.8	**61.7**	65.8	81.3
Hainsworth dataset						
CRF	64.6	61.4	45.0	49.6	**74.4**	**85.8**
[24]	66.9	62.4	53.4	61.0	69.1	79.2
[25]	**71.5**	**67.6**	**57.8**	**64.3**	72.9	82.5
[7]	60.5	51.6	14.9	20.5	39.1	69.6
[2]	66.4	61.0	52.0	60.6	68.6	81.6

Table 1. Beat tracking results for the three datasets. CRF stands for the method proposed in this article.

5. CONCLUSION

In this paper, we have proposed a novel probabilistic approach to automatic beat tracking. Considering the problem as a sequence labeling one, our proposal exploits a sophisticated temporal model on top of features composed of a tempogram and an onset detection function. In this model, local tempo variations are captured through well-chosen potentials specifying a conditional random field that allows for locating beats at the granularity of short-time analysis windows.

Compared to reference beat tracking systems, ours performs competitively and obtains the best scores on all datasets on two evaluation measures (AMLc and AMLt), while being perfectible in terms of F-measure.

Given the simplistic nature of the features used in this work, our proposal holds a great promise towards improved beat tracking, and by extension downbeat detection, as higher-level features are considered. These are some of the directions that will be explored in future work.

[1] www.ballroomdancers.com

[2] http://www.music-ir.org/mirex/wiki/2014:Audio_Beat_Tracking

[3] Where beats are detected at half or twice the annotated rate

6. REFERENCES

[1] F. Gouyon and S. Dixon, "A review of automatic rhythm description systems," *Computer Music Journal*, vol. 29, 2005.

[2] G. Peeters and H. Papadopoulos, "Simultaneous beat and downbeat-tracking using a probabilistic framework: theory and large-scale evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1754–1769, Aug 2011.

[3] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[4] T. Jehan, *Creating music by listening*, Ph.D. thesis, Massachusetts Institute of Technology, 2005.

[5] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, 2006.

[6] J. Laroche, "Efficient tempo and beat tracking in audio recordings," *Journal of the Audio Engineering Society*, vol. 51, no. 4, pp. 226–233, 2003.

[7] D. P. W. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[8] S. Hainsworth and M. D. Macleod, "Particle filtering applied to musical tempo tracking," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 2385–2395, 2004.

[9] C. Joder, S. Essid, and G. Richard, "A conditional random field framework for robust and scalable audio-to-score matching," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2385 – 2397, nov. 2011.

[10] C. Joder, S. Essid, and G. Richard, "Hidden discrete tempo model: a tempo-aware timing model for audio-to-score alignment," in *ICASSP*, May 2011.

[11] M. Alonso, G. Richard, and B. David, "Tempo estimation for audio recordings," *Journal of New Music Research*, vol. 36, no. 1, March 2007.

[12] M. Alonso, G. Richard, and B. David, "Accurate tempo estimation based on harmonic+noise decomposition," *EURASIP Journal on Advances in Signal Processing*, 2007.

[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields : Probabilistic models for segmenting and labeling sequence data," in *ICLM*, 2001.

[14] C. Sutton and A. McCallum, "Dynamic Conditional Random Fields : Factorized Probabilistic Models for Labeling and Segmenting Sequence Data," in *ICML*, 2004.

[15] C. Sutton and A. McCallum, *An introduction to Conditional Random Fields for relational learning*, ir 4, pp. 93–128, MIT Press, 2006.

[16] E. W. Large and J. F. Kolen, "Resonance and the perception of musical meter," *Connection science*, vol. 6, no. 2-3, pp. 177–208, 1994.

[17] E. W. Large, "Beat tracking with a nonlinear oscillator," in *Working Notes of the IJCAI-95 Workshop on Artificial Intelligence and Music*, 1995, vol. 24031.

[18] A. T. 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, 2000.

[19] H. M. Wallach, "Conditional random fields: An introduction," *Technical Reports (CIS)*, p. 22, 2004.

[20] S. Essid, "A tutorial on conditional random fields with applications to music analysis," presented at the 14th International Society for Music Information Retrieval Conference, nov 2013.

[21] S. Dixon, "Evaluation of audio beat tracking system beatroot," *Journal of New Musical Research*, vol. 36, pp. 39–51, 2007.

[22] M. Goto, "Issues in evaluating beat tracking systems," in *IJCAI*, 1997.

[23] M. E. P. Davies, N. Degara, and M. D. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," *Queen Mary University, Centre for Digital Music, Tech. Rep. C4DM-TR-09-06*, 2009.

[24] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1009–1020, 2007.

[25] N. Degara, E. Argones Rua, A. Pena, S. Torres-Guijarro, M. E. P. Davies, and M. D. Plumbley, "Reliability-informed beat tracking of musical signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 290–301, 2012.