HIDDEN DISCRETE TEMPO MODEL: A TEMPO-AWARE TIMING MODEL FOR AUDIO-TO-SCORE ALIGNMENT

Cyril Joder, Slim Essid, Gaël Richard

Institut Télécom - Télécom ParisTech - CNRS/LTCI
37 rue Dareau, 75014 Paris, France
{joder, essid, grichard}@telecom-paristech.fr

ABSTRACT

In this paper, we present the Hidden Discrete Tempo Model, an effective Dynamic Bayesian Network for audio to score matching. Its main feature is an explicit modeling of tempo, which directly influences the timing model of the musical performance. Thanks to a discretization of the tempo set, it allows for an efficient decoding by the Viterbi algorithm, and facilitates the introduction of features which directly depend on the local tempo. We take advantage of this property by using the cyclic tempogram descriptor in addition to chroma vectors and onset detection features. Experiment run on both classical piano and pop music show the very high accuracy of this model for audio to score alignment, as well as the usefulness of the tempo feature used.

***Index Terms*—** music information retrieval, automatic alignment, dynamic Bayesian networks, acoustic features

1. INTRODUCTION

Audio-to-score alignment, which is the task of matching a musical performance with the corresponding score, can lead to several kinds of applications. In a real-time context, it can be used for the tracking of a live performance (see for example [1]), which then allows for interactions between the musicians and a computer.

Other applications, which do not require the real-time constraint can also be found in the field of Music Information Retrieval (MIR). Indeed, a music-to-score synchronization provides a precise and meaningful indexing of the audio content, with high-level musical information. Consequently, it allows for an intuitive browsing in a musical piece, musicological analyses such as chord transcription or even score-informed source separation. The presence of a large number of freely available scores on the Internet makes the use of music-to-score alignment for these indexing applications possible.

In the recent literature, many real-time score following systems have employed probabilistic models which belong to the Dynamic Bayesian Network (DBN) class [2]. In such systems, hidden random variables represent the current position in the score, in order to take into account the uncertainty of the matching. The most widely used of these models is probably the Hidden Markov Model (HMM) (for example [3]).

However, the Markovian property of HMMs can be a weakness for the modeling of the note durations. Indeed, in an HMM, the note lengths are supposed to be independent, and their prior distributions that can be introduced are "absolute" (in seconds). This model does not always correspond to the reality of western music since the timing of most musical pieces is given relatively to a *tempo* process (in beats), which can be both unknown and variable.

For this reason, more elaborate models for audio-to-score matching [4, 5, 1] introduce another random process representing the tempo. In such systems the note duration probabilities are then dependent on the current tempo value. In [4], the duration model is quite rudimentary, since the tempo variable can only take three values ('fast', 'medium' and 'slow'). As a result, the tempo process does not strongly constrain the note duration. In [5] and [1], the tempo is modeled by a continuous variable. This potentially allows for a flexible modeling of the tempo. However, the introduction of this continuous variable prohibits the use of dynamic programming methods for exact decoding of the probabilistic model. Cont [1] uses an adaptive framework which updates a tempo estimate at each step of the algorithm. Raphael [5] takes advantage of the specific transition probabilities of his model to calculate the current tempo probability corresponding to each partial path in the score.

In this work, we introduce the Hidden Discrete Tempo Model (HDTM), which exploits a discretization of the tempo set. This has two main advantages compared to a continuous model. First, it allows for a practical inference of all the model variables thanks to dynamic programming techniques. This model also allows for the use of acoustic features characterizing the local tempo, in addition to the pitch and onset descriptors. We show that, thanks to this model, the use of the *cyclic tempogram* features [6] improves the alignment precision on a large database of both popular and classical music.

The rest of this paper is organized as follows: the Hidden Discrete Tempo Model is introduced in Section 2. We then detail in Section 3 the observation models. Experiments on the alignment precision obtained with this model and the influence of the latter feature are presented in Section 4 before suggesting some conclusions in Section 5.

2. THE HIDDEN DISCRETE TEMPO MODEL (HDTM)

2.1. Timing Model

A polyphonic musical score can be segmented into *chords*, defined as sets of notes that sound at the same time. Every time a note appears or disappears, a new chord is created. This segmentation provides a "linear" representation of the score as a sequence of chords, as displayed in Figure 1. Given this representation, aligning an audio recording to the score boils down to finding the time indexes of the chord onsets, or equivalently the length of these chords.

Let c be a chord index and let T^c be a random variable representing the current tempo (expressed in seconds per beat). The "theoretical" length (if performed perfectly in time), in seconds, of the

This work has been partly supported by the Quaero Program, funded by OSEO, French State agency for innovation.

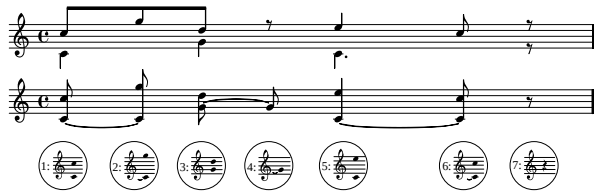

Fig. 1. Score Representations. Top: The original graphical score. Middle: "Homophonic" version of the same score. Bottom: Chord indexes.

given chord is $\lambda^c T^c$, where λ^c is the duration, in beats, indicated by the score. However, in order to account for the discretization of the tempo set and for the performance imprecision (due to interpretation choices or possible human errors), a deviation is allowed. More formally, let L^c denote the random variable representing the length of chord c (in seconds). We set a gaussian-like probability distribution:

$$P(L^k|T^k) = \frac{1}{Z_1} \exp\left\{-\frac{(L^k - \lambda^k T^k)^2}{2(\lambda^k T^k)^2 \sigma_l^2}\right\} \quad (1)$$

where σ_l^2 is a parameter which controls the tolerance to the timing deviations and Z is a normalizing factor. As in [7], we expect the possible deviation to increase with the note duration. That is why the variance is proportional to the theoretical length.

As implicitly used in equation (1), we suppose that the tempo is constant over a chord's duration and only changes at chord transitions. In Raphael's model [5], the tempo values are modeled by a Gaussian random walk process. Instead of that, 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. Thus, similarly to [8] the transition probabilities are Gaussian with respect to the logarithm tempo values:

$$P(T^{k+1}|T^k) = \frac{1}{Z_2} \exp\left\{-\frac{1}{2\sigma_t^2}\left(\log\frac{T^{k+1}}{T^k}\right)^2\right\} \quad (2)$$

where σ_t controls the tempo variation tolerance and Z_2 is a normalizing factor. In practice, we consider that in the case of strong, abrupt tempo changes, any tempo can be reached with the same probability. Hence, we limit the tempo ratio in equation (2) to 2.

2.2. Dynamic Bayesian Network Representation

For a practical representation of the HDTM, we use the Dynamic Bayesian Network (DBN) formalism [2]. We suppose that the recording is divided into a discrete sequence of short-time frames. Let N be the number of frames. For each time frame n, let C_n and T_n be random variables representing respectively the current chord and the current tempo. Note that, if $C_n = c$, we have $T_n = T^c$. For notation simplicity, this will be denoted by T^{C_n}. We also use an *occupancy* variable D_n whose value is equal to the number of frames since the beginning of the current chord. Hence, we have $D_n = 1$ iff the frame n corresponds to a chord onset. The relations between the variables are:

$$D_{n+1} = \begin{cases} 1 & \text{if } L^{C_n} = D_n \\ 1 + D_n & \text{otherwise,} \end{cases}$$

$$C_{n+1} = \begin{cases} 1 + C_n & \text{if } D_{n+1} = 1 \\ C_n & \text{otherwise,} \end{cases}$$

$$T_{n+1} = \begin{cases} T^{C_{n+1}} & \text{if } D_{n+1} = 1 \\ T_n & \text{otherwise,} \end{cases}$$

where L^{C_n} denotes the length of chord C_n. The corresponding probabilities are calculated thanks to equations (1) and (2).

In order to characterize the variations that can occur inside a chord, in particular between the *attack* and *sustain* phases, we also introduce a Bernoulli *attack indicator* variable A_n. The event $A_n = 1$ indicates an *attack* phase. We assume that the first frame of an "attacking chord" (whose beginning corresponds to a newly entering note) is always in an *attack* phase. The second frame can be either in an *attack* or in a *sustain* phase. Since a note attack is supposed to be short, we assume that all the following frames of the chord correspond to the *sustain* phase. Thus, the probability to be in the *attack phase* is:

$$P(A_n = 1|C_n, D_n) = \begin{cases} 1 & \text{if } C_n \text{ attacking and } D_n = 1 \\ \frac{1}{2} & \text{if } C_n \text{ attacking and } D_n = 2 \\ 0 & \text{otherwise.} \end{cases}$$

3. OBSERVATION MODELING

Similarly to [9], we use chroma features to characterize the pitch content, and an onset feature derived from the spectral flux to detect the note attacks. These features are denoted respectively by V_n and F_n. However the HDTM allows for the consideration of yet another kind of information, regarding the current tempo. Hence we introduce the use of the cyclic tempogram feature, denoted by G_n in an alignment system. The complete dependency structure of the model is represented in Figure 2.

We suppose that these observations only depend on their "corresponding" hidden variables, so that the conditional probability of the observations given the hidden variables is:

$$P(V_n, F_n, G_n|C_n, A_n, T_n) = P(V_n|C_n)P(F_n|A_n)P(G_n|T_n)$$

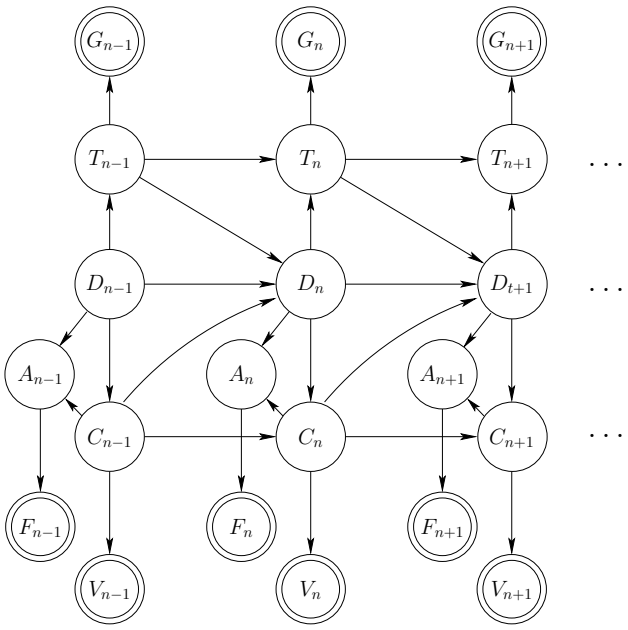

Fig. 2. Graphical model representation of the presented models. Simple and double contour lines indicate respectively hidden and observed variables.

3.1. Chroma Vectors

Chroma vectors provide a compact, yet efficient representation of the harmonic content of a musical signal for audio-to-score alignment [10]. We use here Zhu's chroma features [11], with a time resolution of 50 Hz. For each chord label c, a chroma vector template u_c is built, in the same way as in [10]. This template can be considered as a "theoretical chroma synthesis" of the chord. The chroma model is then given by

$$P(V_n = v|C_n = c) = \frac{1}{Z_3} e^{\alpha D(\bar{v}\|\bar{u_c})},$$

where $D(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence and the operator $\bar{\cdot}$ represents a normalization so that a vector can be considered as a probability distribution (the components sum to 1). α is a positive parameter and Z_3 is a normalization factor.

3.2. Onset Feature

The onset feature used is a straightforward onset detector function based on spectral flux [12]. It is obtained by subtracting a local threshold (calculated with a 67% rank filter of length 200-ms) to the spectral flux values. A logistic model is used in order to calculate the probability of an onset:

$$P(A_n = 1|F_n) = \frac{e^{\beta F_n}}{1 + e^{\beta F_n}}$$

where β is a positive parameter controlling the "confidence" on the onset detector. The probability $P(F_n|A_n)$ can then be calculated by Bayes' rule.

3.3. Cyclic Tempogram

The *cyclic tempogram* feature has been introduced in [6] for a musical structure analysis application. It provides a mid-level representation of the tempo which allows for a robust tempo analysis, since it takes into account not only the time-lags corresponding to a beat length, but also the ones corresponding to half or twice this beat length.

In order to calculate this feature, we first compute the local (normalized) autocorrelation of the *spectral flux* feature over sliding 5-s windows, for time-lags between $\tau_{\text{min}} = 200$ ms and $\tau_{\text{max}} = 3.2$ s. Let $h_n(\tau)$ be the value of this autocorrelation function computed over a window centered on frame n.

Similarly to a *chromagram*, the time lags are separated into *octave equivalence classes*: two time-lags τ_1 and τ_2 are octave equivalent iff there is a $k \in \mathbb{Z}$ s.t. $\tau_1 = 2^k \tau_2$. The value $g_n(\tau)$ of the cyclic tempogram for a time-lag τ is calculated by adding all the values of this autocorrelation function corresponding to the same equivalence class:

$$g_n(\tau) = \sum_{k \in \mathbb{Z}} h_n(2^k \tau).$$

In practice, the autocorrelation function is "blurred" by a Gaussian filter in order to account for imprecisions induced by the discretization of the time lag set.

The observation model is then given by

$$P(T_n = t|G_n = g_n) = \frac{1}{Z_4} e^{\gamma g_n(t)},$$

where γ is a positive parameter and Z_4 is a normalization factor. The probabilities $P(G_n|T_n)$ are then calculated by Bayes' rule.

4. EXPERIMENTS

4.1. Database and Settings

For our experiments, we use two databases. The first one contains 59 classical piano pieces (about 4h15), from the MAPS database [13]. These recordings are the rendition of MIDI files played by a Disklavier piano. The second corpus is composed of 90 pop songs (about 6h) from the RWC database [14]. A ground-truth is provided as aligned MIDI files. The target scores are built from the same files. However, we do not consider the tempo values of the MIDI files, in order to simulate the use of graphical scores (sheet music). Moreover, we discarded the possible percussion parts, because of the variable quality of their transcription.

A learning database has been built using one hour from each of these sets. The parameters of the models are set thanks to a coarse grid search on this learning database. The evaluation is then run on the rest of both MAPS and RWC datasets.

The chosen evaluation measure is the onset recognition rate, defined as the fraction of onsets which are correctly detected less than a tolerance threshold θ away from the real onset time of each note of the score. The value $\theta = 300$ ms is based on the MIREX contest[1]. For a more precise alignment evaluation, we use two other thresholds: 100-ms and 50-ms.

The alignment with the HDTM is performed by estimating the *Maximum a Posteriori* (MAP) path in the model, defined as

$$\text{argmax}\, P(\mathbf{C}_1^N, \mathbf{A}_1^N, \mathbf{D}_1^N, \mathbf{T}_1^N \,\big|\, \mathbf{V}_1^N, \mathbf{F}_1^N, \mathbf{G}_1^N),$$

where $\mathbf{C}_1^N = C_1, \ldots, C^N$. The MAP sequence is computed thanks to the Viterbi algorithm, along with the pruning strategy exposed in [9]. The used set of possible tempo values is, in beats per minute:

$$\mathcal{T} = \{28, 30, 34, 40, 48, 56, 64, 72, 80, 88, 96, 104,$$
$$112, 120, 132, 146, 160, 176, 192, 208, 224, 240\}$$

The values of the model parameters, estimated on the learning database, are displayed in Table 1.

Parameter	α	β	γ	σ_t^2	σ_l^2
Value	10	10	1	$\frac{1}{20}$	$\frac{1}{200}$

Table 1. Estimated model parameter values.

4.2. Alignment Results

In these experiments, we test two different systems in order to assess the usefulness of exploiting a feature characterizing the tempo. The first system uses the model described before, whereas the other does not exploit the cyclic tempogram descriptor (the parameter γ is set to 0). The obtained recognition rates are summed up in Table 2.

The alignment obtained by the HDTM are very precise. Indeed, more than 99% of the chords are recognized less than 300-ms away from the ground truth, on both databases. As a comparison, the recognition rate of the (rudimentary) HMM system used in [9] is about 86%.

The performance is also very high at a finer precision level since on the MAPS corpus, the recognition rates are higher than 91% for a 50-ms threshold. Lower scores are obtained on the RWC dataset,

[1] Music Information Retrieval Evaluation eXchange 2010, score following task: `http://www.music-ir.org/mirex/wiki/2010:Real-time_Audio_to_Score_Alignment_(a.k.a_Score_Following)`

	MAPS Corpus		RWC Corpus	
Model	$\gamma = 0$	$\gamma = 1$	$\gamma = 0$	$\gamma = 1$
300-ms	99.34%	99.40%	99.00%	99.22%
100-ms	97.91%	98.03%	94.21%	94.56%
50-ms	91.08%	91.25%	75.18%	75.50%

Table 2. Recognition Rates of the Hidden Discrete Tempo Model, with the use of the tempo feature ($\gamma = 1$) and without ($\gamma = 0$). As a comparison, a HMM system obtains about 86%.

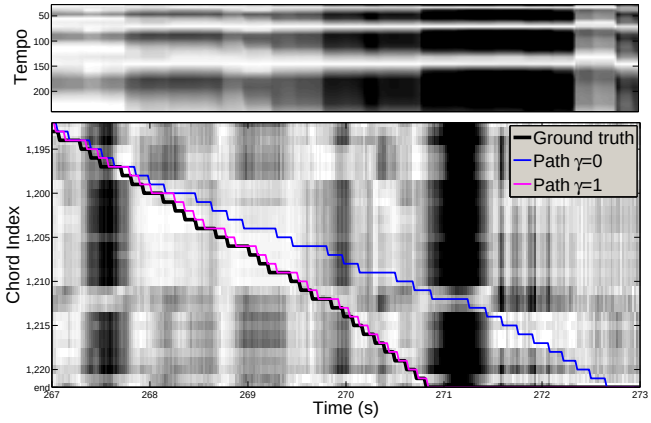

Fig. 3. Alignments obtained on an example piece. Up: tempo likelihoods. Down: chord likelihoods and alignment paths. White indicates high values.

which can be explained by the higher complexity of the music content in terms of number of instruments, but also by some annotation errors.

The benefit of using the tempo feature is not always marked. Indeed the improvement is not statistically significant on the MAPS database. However, the recognition rate increases obtained on the RWC corpus are larger than the radii of the 95% confidence intervals (which are respectively equal to 0.07%, 0.15% and 0.28% for the three thresholds). This can be explained by the steadier tempi and the percussive contents (mainly drums), which consitute strong indications about the tempi, in this corpus.

Figure 3 displays the example of the end of a particular pop song where the benefit of the tempo feature can be seen. On this extract, the likelihood of the ground truth path is relatively low, because of a discrepancy between the chroma model and the observation (visible near 271 s). Thus, the system which does not consider the cyclic tempogram features drifts to a slower tempo path, whose chroma templates better fit the observations. However, in this part, the tempo feature strongly favors the ground truth tempo (and its octaves). This forces the alternative system to follow a steady tempo, and the resulting alignment is more accurate. Even though this situation is not frequent (it only happens one three songs out of 75), the tempo feature does not harm the performance in the other pieces. Thus we consider it to be worthwhile.

5. CONCLUSION

In this paper, we present the Hidden Discrete Tempo Model, a tempo-dependent model for audio to score matching. Its main feature is a hidden tempo variable, allowing for an explicit modeling of the timing of the musical performance. The representation as a Dynamic Bayesian Network with discrete hidden variables makes an efficient decoding possible through the Viterbi algorithm, and allows for the consideration of features characterizing the local tempo.

Experiments run on both classical piano and pop music show the very high accuracy of this model for audio to score alignment, as well as the usefulness of the cyclic tempogram as a tempo feature. The tempo precision can be adjusted by changing the characteristics of the discretization grid, which can be non-linear. It is also worth mentioning that, although we performed off-line alignment, this model could straighforwardly be applied to a real-time context. Indeed, the complexity of this model can be reduced by pruning methods such as beam search.

6. REFERENCES

[1] Arshia Cont, "A coupled Duration-Focused architecture for Real-Time Music-to-Score alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 974–987, June 2010.

[2] Kevin P. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Computer science division, UC Berkeley, July 2002.

[3] Nicola Montecchio and Nicola Orio, "A discrete filterbank approach to audio to score matching for score following," in *Proc. of ISMIR*, 2009.

[4] Paul Peeling, A. Taylan Cemgil, and Simon Godsill, "A probabilistic framework for matching music representations," in *Proc. of ISMIR*, Vienna, Austria, 2007, pp. 267–272.

[5] Christopher Raphael, "Aligning music audio with symbolic scores using a hybrid graphical model," *Machine Learning Journal*, vol. 65, pp. 389–409, 2006.

[6] Peter Grosche, Meinard Müller, and Frank Kurth, "Cyclic tempogram – a mid-level tempo representation for music signals," in *Proc. of ICASSP*, Mar. 2010.

[7] H. Takeda, T. Nishimoto, and S. Sagayama, "Rhythm and tempo analysis toward automatic music transcription," in *Proc. of ICASSP*, Apr. 2007, vol. 4, pp. IV–1317 –IV–1320.

[8] A. T. Cemgil, H. J. Kappen, P. Desain, and H. Honing, "On tempo tracking: Tempogram Representation and Kalman filtering," *Journal of New Music Research*, vol. 28:4, pp. 259–273, 2001.

[9] Cyril Joder, Slim Essid, and Gaël Richard, "An improved hierarchical approach for music-to-symbolic score alignment," in *Proc. of ISMIR*, Utrecht, Holland, Aug. 2010.

[10] Cyril Joder, Slim Essid, and Gaël Richard, "A comparative study of tonal acoustic features for a symbolic level music-to-score alignment," in *Proc. of ICASSP*, 2010.

[11] Yongwei Zhu and M.S. Kankanhalli, "Precise pitch profile feature extraction from musical audio for key detection," *IEEE Trans. on Multimedia*, vol. 8, no. 3, pp. 575–584, June 2006.

[12] Miguel Alonso, Gaël Richard, and Bertrand David, "Extracting note onsets from musical recordings," in *Proc. of ICME*, 2005.

[13] Valentin Emiya, Roland Badeau, and Bertrand David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 18, no. June, pp. 1643 –1654, Aug. 2010.

[14] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "Rwc music database: Popular, classical, and jazz music databases," in *Proc. of ISMIR*, 2002, pp. 287–288.