A Conditional Random Field Viewpoint of Symbolic Audio-to-Score Matching

Cyril Joder
Institut Télécom - Télécom ParisTech - CNRS/LTCI
37 rue Dareau
75014 Paris, France
cyril.joder@telecom-paristech.fr

Slim Essid
Institut Télécom - Télécom ParisTech - CNRS/LTCI
37 rue Dareau
75014 Paris, France
slim.essid@telecom-paristech.fr

Gaël Richard
Institut Télécom - Télécom ParisTech - CNRS/LTCI
37 rue Dareau
75014 Paris, France
gael.richard@telecom-paristech.fr

ABSTRACT

We present a new approach of symbolic audio-to-score alignment, with the use of Conditional Random Fields (CRFs). Unlike Hidden Markov Models, these graphical models allow the calculation of state conditional probabilities to be made on the basis of several audio frames. The CRF models that we propose exploit this property to take into account the rhythmic information of the musical score. Assuming that the tempo is locally constant, they confront the neighborhood of each frame with several tempo hypotheses.

Experiments on a pop-music database show that this use of contextual information leads to a significant improvement of the alignment accuracy. In particular, the proportion of detected onsets inside a 100-ms tolerance window increases by more than 10% when a 1-s neighborhood is considered.

Categories and Subject Descriptors

H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Methodologies and techniques*

General Terms

Algorithms, Experimentation

Keywords

Conditional Random Fields, Music Information Retrieval, Audio/Score Alignment, Indexing

1. INTRODUCTION

Audio-to-score matching aims at finding the correspondence between a musical score and a recording of the piece. This matching allows for the use of the score as precise indexing information about the audio. Thus, applications such as score-controlled audio browsing or automatic musical analysis can be associated to this task. We deal with the case where the score is a "perfect scanned sheet", that is a symbolic score where the relative note durations (according to the tempo) are indicated, but where the tempo is unknown.

Many work on audio-to-score alignment rely on Hidden Markov Models (HMM) [8, 6] whose hidden states account for the notes or chords of the score. The dependencies between different audio frames are modeled by the state transition model. However, the HMM structure is limited in that it can only render a specific form of temporal dependencies. In particular, the correlation between the note lengths —the idea of *tempo*— cannot be modeled.

Different approaches have been proposed in previous works to incorporate this concept of tempo into the models. Raphael [9] and Cont [1] consider an additional random variable representing tempo. Thus, additional dependencies between variables at different times can be modeled. However, in such semi-Markov models, an audio observation is supposed to depend only on the current state.

In [7], a "filtering" of the similarity matrix is introduced to include contextual information in the comparison between points of an audio stream for musical structure analysis.

Following a similar idea, we propose in this work to take into account the neighborhood of each audio frame in our symbolic audio-to-score alignment problem. This is made possible in the Conditional Random Fields [5] (CRFs) framework. CRFs are probabilistic models for labeling and segmenting sequential data, which have been designed for natural language processing [5]. This framework allows for the consideration of audio frames from an arbitrary past or future for the calculation of each state probability.

The rest of this paper is organized as follows: our CRF model for symbolic audio-to-score alignment is introduced in Section 2. Section 3 provides results of alignment experiments and finally some conclusions are drawn in Section 4.

2. CONDITIONAL RANDOM FIELDS FOR MUSIC-TO-SCORE MATCHING

2.1 Conditional Random Fields

Conditional Random Fields (CRFs) are a form of undirected graphical models that can be seen as the discriminative counterpart of HMMs, which are said to be *generative*. Whereas in HMMs, the observations are conditioned on the hidden variables, CRFs condition these hidden variables on the observation sequence.

Figure 1 compares both model structures. Double nodes

MM'10, October 25–29, 2010, Firenze, Italy.

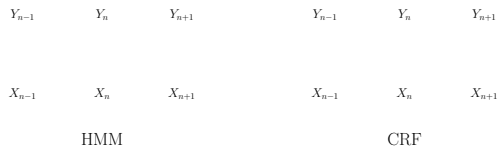

Figure 1: Comparison between HMM and CRF graphical representations.

Figure 2: Example of the score segmentation into a sequence of chords.

represent observed variables and shaded nodes (in CRF) correspond to variables that the model conditions on. Since these latter variables are observed, no conditional independence property is assumed, contrary to HMMs, in which the observed variables are independent given the hidden variables. Thus, features from an arbitrary past or future may be considered without a significant increase of the decoding complexity. For more information, we refer to [5].

2.2 Structure of the Alignment Model

The audio recording is split into short overlapping frames. Let $\mathbf{x}_1^N = x_1 \dots x_N$ be this observation sequence of length N. Our aim is to find the positions in the recording (as frame indexes) of the notes indicated by the score. However, it can also be seen as the problem of assigning a location in the score to each frame. More precisely, we segment the score into *chords*, which are sets of notes sounding simultaneously. This is illustrated in Figure 2. Then, we search for the chord label sequence that best matches the audio data.

Let $\mathbf{Y}_1^N$ be the (unobserved) random process representing the chord labels. In the following, the boundary indexes $_1^N$ will be omitted when not needed to clarify the presentation. We also consider other hidden variables than the chord label to model the random process. Let T_n be a (discrete) random variable symbolizing the current tempo value, and D_n the location within the current chord at time n ($\{D_n = d\}$ means that the current chord started at time $n-d$).

Our goal is to find the most probable sequences of chord labels $\hat{\mathbf{y}}$, duration indexes $\hat{\mathbf{d}}$ and tempo values $\hat{\mathbf{t}}$, given the observation sequence $\mathbf{x}$:

$$\left(\hat{\mathbf{y}}, \hat{\mathbf{d}}, \hat{\mathbf{t}}\right) = \underset{\mathbf{y},\mathbf{d},\mathbf{t}}{\operatorname{argmax}}\, p\left(\mathbf{Y} = \mathbf{y}, \mathbf{D} = \mathbf{d}, \mathbf{T} = \mathbf{t} | \mathbf{X} = \mathbf{x}\right). \quad (1)$$

The returned chord label sequence is then $\hat{\mathbf{y}}$.

Let $V_n = (Y_n, D_n, T_n)$ be the vector of hidden variables at time n. In CRFs, the assumption is made that the process $\{V_n\}$ verifies the Markov property, conditioned on the observation sequence. Thus, the conditional probability of (1) can be factorized as

$$p\left(\mathbf{v}|\mathbf{x}\right) = p(v_1|\mathbf{x}) \prod_{n=2}^{N} p(v_n|v_{n-1}, \mathbf{x}). \quad (2)$$

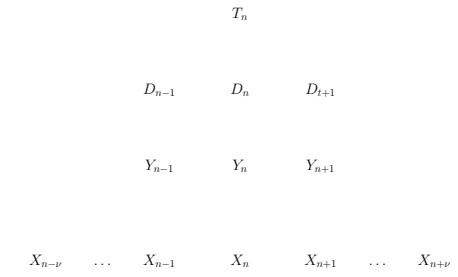

Figure 3: Graphical representation of our CRF models. Solid lines: *unconstrained model*. Dashed lines: additional edges of the *constrained model* (only edges concerning variables at time n are drawn).

2.2.1 Unconstrained Model

In our first model, which we call the *unconstrained model*, a simplifying hypothesis allows us to obtain a problem which is very close to a HMM. We assume that the probability $p(v_n|v_{n-1}, \mathbf{x})$ can be separated into a *transition function* $\psi(y_n, y_{n-1})$ controlling the transitions between the chords, and an *observation function* $\phi_\nu(v_n, \mathbf{x}_{n-\nu}^{n+\nu})$ which links the current hidden variables with the observations within a neighborhood of ν frames. The probability of (2) can then be written

$$p\left(\mathbf{v}|\mathbf{x}\right) = \frac{1}{Z(\mathbf{x})} \psi_0\left(y_1\right) \prod_{n=2}^{N} \psi\left(y_n, y_{n-1}\right) \prod_{n=1}^{N} \phi_\nu\left(v_n, \mathbf{x}_{n-\nu}^{n+\nu}\right) \quad (3)$$

where $Z(\mathbf{x})$ is a normalizing factor. The CRF framework allows us to take into account an arbitrary neighborhood (thanks to the parameter ν) in a simple manner. Note that the case $\nu = 0$ boils down to a hidden Markov model.

2.2.2 Constrained Model

In the previous model, the transition function ψ does not depend on the variables D_n, indicating that no prior information on the sequence $\mathbf{D}$ is modeled. This choice is explained by complexity considerations, however this is a very coarse approximation. Indeed, this process is strongly structured since it is in fact a counter: we have either $D_n = D_{n-1} + 1$ (if the chord is the same), or $D_n = 0$ (if $Y_n \neq Y_{n-1}$).

This constraint on the process D is added in the *constrained model*. In this model, the probability factorizes as:

$$p\left(\mathbf{y}, \mathbf{d}, \mathbf{t}|\mathbf{x}\right) = \frac{1}{Z(\mathbf{x})} \psi_0\left(y_1\right) \prod_{n=2}^{N} \psi'(y_n, d_n, y_{n-1}, d_{n-1}) \prod_{n=1}^{N} \phi_\nu\left(y_n, d_n, t_n, \mathbf{x}_{n-\nu}^{n+\nu}\right). \quad (4)$$

The transition function ψ' takes four arguments, resulting in a more complex model than the previous one. The graphical representations of both models are displayed in Figure 3.

2.3 Transition and Observation Functions

2.3.1 Transition Functions

In the alignment process, we suppose that the order of the chords is the one indicated in the score. Hence, there are only two possible chord transitions: either a continuation of

the same chord, or the beginning of the next one. We then set the transition functions of (3) to:

$$\begin{aligned}\psi_0(y_1) &= \delta_{\{y_1,1\}} \\ \psi(y_n, y_1) &= \delta_{\{y_n,y_{n-1}\}} + \delta_{\{y_n,y_{n-1}+1\}},\end{aligned}$$

where $\delta_{\{\cdot,\cdot\}}$ is the Kronecker function. Thus, only sequences of chords in the right order have non-zero probabilities.

For the *constrained model*, in order to express the structural constraint on D_{n-1} (see Sec. 2.2.2), the transition function becomes:

$$\begin{aligned}\psi'(y_n, d_n, y_{n-1}, d_{n-1}) &= \delta_{\{y_n,y_{n-1}\}}\delta_{\{d_n,d_{n-1}+1\}} \\ &\quad +\delta_{\{y_n,y_{n-1}+1\}}\delta_{\{d_n,0\}}.\end{aligned}$$

2.3.2 Observation Functions

The features that we use in this work are *chroma vectors* which are extracted according to [10], with a time resolution of 50 Hz. These features provide an efficient representation of the pitched content of a musical signal for our task [4].

For a chord label y, we build a chroma vector template g_y from the content of this chord, as in [3]. This template is normalized so that it can be regarded as a probability distribution over the chroma bins. In the case of silence, we use a flat template, in order to model the background noise. As a matching measure $f(x, y)$ between a chroma vector x and the chord y, we use the opposite of the Kullback-Leibler divergence:

$$f(x, y) = -\sum_{i=1}^{12} \bar{x}(i) \log\left(\frac{\bar{x}(i)}{g_y(i)}\right) \quad (5)$$

where $\bar{x}$ is a normalized version of x so that it sums to 1, and $\bar{x}(i)$ is the i-th component of this vector.

In order to take into account the observations extracted from several frames, we make the assumption that the tempo changes slowly over time and can be considered as *locally constant* over windows of length $2\nu + 1$ frames. Thus, it is possible to confront a local feature sequence with a chord sequence corresponding to a score position and a (constant) tempo hypothesis.

Formally, let $v_n = (y_n, d_n, t_n)$ be the hidden variable vector, let $\mathbf{s}(v_n)$ be the chord label sequence corresponding to a theoretical rendition of the score with the hypothesis: H_{v_n}: {*The tempo is constant and equals t_n; the score position at time n is* (y_n, d_n)}. We define the observation function as:

$$\phi_\nu\left(v_n, \mathbf{x}_{n-\nu}^{n+\nu}\right) = \exp\left(\sum_{k=-\nu}^{\nu} \lambda_k f\Big(x_{n+k}, s_{n+k}(v_n)\Big)\right), \quad (6)$$

where the λ_k are parameters which control the importance given to the features at the different time-shifts. Intuitively, the weights λ_k should be decreasing with $|k|$, in order to emphasize the current feature. We set $\lambda_k = e^{-\frac{2|k|}{F_s}}$, where F_s is the feature sampling rate.

Note that in the case $\nu = 0$, the value of $\phi_0(v_n, x_n)$ does not depend on the tempo and duration variables, since we always have $\phi_0\left(v_n, x_n\right) = \exp\left(\lambda_0 f(x_n, y_n)\right)$.

When the hypothesis H_{v_n} is inconsistent, *i.e.* when d_n is greater than the length of the chord y_n at tempo t_n, we set $\phi_\nu\left(v_n, \mathbf{x}_{n-\nu}^{n+\nu}\right) = 0$.

2.4 Decoding Process

In the case of the *unconstrained model*, the optimization problem of (1) can be factorized. Indeed, the maximization over $\mathbf{d}$ and $\mathbf{t}$ can be done "at the feature level". We define:

$$\tilde{\phi}_\nu\left(y_n, \mathbf{x}_{n-\nu}^{n+\nu}\right) = \max_{d,t}\left\{\phi_\nu\left(y_n, d, t, \mathbf{x}_{n-\nu}^{n+\nu}\right)\right\}. \quad (7)$$

This maximum can be computed by an exhaustive search, since all our hidden variables are discrete. The equation (1) can then be written:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\left\{\psi_0(y_1)\prod_{n=2}^{N}\psi(y_n, y_{n-1})\prod_{n=1}^{N}\tilde{\phi}_\nu\left(y_n, \mathbf{x}_{n-\nu}^{n+\nu}\right)\right\},$$

The Markov property of the hidden variables allows for a decoding with the Viterbi algorithm.

For the *constrained model*, the optimal sequence is:

$$\begin{aligned}\left(\hat{\mathbf{y}}, \hat{\mathbf{d}}\right) = \underset{\mathbf{y},\mathbf{d}}{\operatorname{argmax}}\Bigg\{&\psi_0\left(y_1\right)\prod_{n=2}^{N}\psi\left(y_n, y_{n-1}\right) \\ &\prod_{n=1}^{N}\max_t\left\{\phi_\nu\left(y_n, d_n, t, \mathbf{x}_{n-\nu}^{n+\nu}\right)\right\}\Bigg\}.\end{aligned} \quad (8)$$

The Viterbi algorithm can also be used here. However, the decoding is more complex than for the former model, since the process $\hat{\mathbf{d}}$ needs to be stored as well as $\hat{\mathbf{y}}$.

3. ALIGNMENT EXPERIMENTS

3.1 Database and Setting

We apply the presented models to an offline audio-to-score alignment task, on a database of 94 pop songs, from the RWC database [2]. These songs are polyphonic multi-instrumental pieces of length 2 to 6 minutes. The alignment ground-truth is given by the synchronized MIDI files provided with the recordings. The same MIDI files are exploited as target scores, however we do not use the timing information of these ground truth files, as mentioned above.

The chosen evaluation measure is the recognition rate, defined as the fraction of onsets which are correctly detected less than a threshold θ away from the real onset time. We test two values of θ: 300-ms and 100-ms.

The set of possible tempo values Σ has to be a (not too large) finite set, in order to solve the tempo maximization of eq. 7 and 8. Two tempo sets have been selected, based on musical motivations, given here in beat per minute (bpm):

$$\begin{aligned}\Sigma_1 &= \{40, 64, 88, 120, 160, 200, 240\}, \\ \Sigma_2 &= \{40, 48, 56, 64, 72, 80, 88, 96, 104, 112, 120, \\ &\quad 132, 146, 160, 176, 192, 208, 224, 240\}.\end{aligned}$$

Finally, we consider three values of the parameter ν: 0, 25 and 50 frames, corresponding to 0-s, 0.5-s and 1-s.

3.2 Results and Discussion

Table 1 displays the recognition rates obtained in these experiments. From these results, we can first notice that better scores are obtained with the tempo set Σ_2 than with Σ_1 (except in the case $\nu = 0$, where the tempo set has no effect). Consistently with one's intuition, the feature sequences considered in Sec. 2.3.2 can be more accurately modeled when more tempos are tested.

It can also be observed that the additional variable dependencies rendered by the *constrained model* (compared to the *unconstrained* one) improve the accuracy of the model, since the former system always performs better than the latter, for the same parameter settings. This improvement also occurs in the case $\nu = 0$, where no neighborhood is considered.

	Unconstrained Model			*Constrained Model*		
Tempos	$\nu=0$	$\nu=25$	$\nu=50$	$\nu=0$	$\nu=25$	$\nu=50$
Σ_1	65.1	69.3	74.4	79.1	76.4	78.9
Σ_2	65.1	69.9	75.7	79.1	76.8	**79.7**

(a) $\theta = 300$ ms

	Unconstrained Model			*Constrained Model*		
Tempos	$\nu=0$	$\nu=25$	$\nu=50$	$\nu=0$	$\nu=25$	$\nu=50$
Σ_1	36.6	44.3	53.7	46.4	49.3	57.2
Σ_2	36.6	45.1	55.8	46.4	49.9	**59.0**

(b) $\theta = 100$ ms

Table 1: Recognition rates of our systems (in %).

(a) $\nu = 0$ **(b)** $\nu = 50$ (1-s)

Figure 4: "Matching Matrices" of a musical excerpt for two values of the neighborhood parameter ν.

This is due to the fact that the model limits the note lengths to values depending on their duration in beat. As explained in Sec. 2.3.2, the value of ϕ_ν is set to 0 when a note length is incoherent with the tempo hypothesis. Thus, the maximal possible length of a note is its duration under the hypothesis of the slowest considered tempo (40 bpm here).

A disappointing result can be seen in Table 1 (a): the scores of the *constrained model* are lower with $\nu = 25$ than with no contextual information. This can be explained by a limitation of our observation features. Indeed, the chroma templates used in the calculation of the function $f(x, y)$ (see Sec. 2.3.2) are approximations which do not take into account note harmonics nor percussion sounds, and the best match of an observation sometimes does not correspond to the "real chord". Hence, since an observation is taken into account in several consecutive neighborhoods, this bias can be amplified by our CRF models.

However, this problem can be alleviated by increasing the neighborhood length ν. This way, more frames are considered and the proportion of "biased observations" inside each neighborhood decreases, resulting in a more robust localization if the tempo grid is fine enough to render these longer dependencies. Thus, the results are improved by a 1-s neighborhood, with the tempo set Σ_2. Moreover the "precise alignment" recognition rate (with a 100-ms threshold) increases with ν, for all the tested settings. This indicates that taking into account contextual information does help to match an audio observation with its position in the score.

This is further illustrated on Figure 4, where "matching matrices" storing the values of the function $\tilde{\phi}_\nu$ of eq. (7) are displayed for two values of ν. Whereas with $\nu = 0$, the matching matrix is hard to read, its counterpart with a 1-s neighborhood clearly brings to light high-score paths, corresponding to repeating phrases. Thus, the position in the score can be more accurately estimated.

4. CONCLUSION

We present a new approach of symbolic audio-to-score alignment, with the use of the Conditional Random Fields framework. These models allow us to take into account a neighborhood of each frame for the matching of an audio observation with its symbolic description (*e.g.* chord). We propose two CRF models which take into account the rhythmic information given by the score thanks to a hidden variable representing tempo.

Experiments show that the use of contextual information does improve the accuracy of the obtained alignments. In spite of a known limitation of the chroma features used which can affect in particular recordings containing much percussion, the 100-ms recognition rate increases by more than 10% when a 1-s neighborhood is considered.

This approach, applied here to an offline alignment task, can very easily be transposed to the real-time case, by considering only the past neighborhood. The use of the CRF framework could allow for a learning of the parameter λ_k in a discriminative fashion. We believe that there is also much room for improvement in the design of the dependency structure of the model. In our experiment, the *constrained model*, which takes into account more variable dependencies than the simpler *unconstrained model*, obtains better results. More complex structures could be conceived in order to more accurately render the process dynamics.

5. REFERENCES

[1] A. Cont. A coupled Duration-Focused architecture for Real-Time Music-to-Score alignment. *IEEE Trans. on PAMI*, 32(6):974–987, June 2010.

[2] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. Rwc music database: Popular, classical, and jazz music databases. In *Proc. of ISMIR*, 2002.

[3] N. Hu, R. B. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proc. of WASPAA*, 2003.

[4] C. Joder, S. Essid, and G. Richard. A comparative study of tonal acoustic features for a symbolic level music-to-score alignment. In *Proc. of ICASSP*, 2010.

[5] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, 2001.

[6] N. Montecchio and N. Orio. A discrete filterbank approach to audio to score matching for score following. In *Proc. of ISMIR*, 2009.

[7] M. Müller and F. Kurth. Enhancing similarity matrices for music audio analysis. In *Proc. of ICASSP*, 2006.

[8] B. Pardo and W. Birmingham. Modeling form for on-line following of musical performances. In *Proc. of National Conference on Artificial Intelligence*, 2005.

[9] C. Raphael. Aligning music audio with symbolic scores using a hybrid graphical model. *Machine Learning Journal*, 65:389–409, 2006.

[10] Y. Zhu and M. Kankanhalli. Precise pitch profile feature extraction from musical audio for key detection. *IEEE Trans. on Multimedia*, 8(3):575–584, June 2006.