SECTION-LEVEL MODELING OF MUSICAL AUDIO FOR LINKING PERFORMANCES TO SCORES IN TURKISH MAKAM MUSIC

Andre Holzapfel[1]*, Umut Şimşekli*[1]*, Sertan Şentürk*[2]*, Ali Taylan Cemgil*[1]

1: Dept. of Computer Engineering, Boğaziçi University, 34342, Bebek, İstanbul, Turkey
2: Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

ABSTRACT

Section linking is an important task that is closely related to audio-to-score alignment, where the aim is to relate certain important structural boundaries in a reference score of a piece to their occurrences in the recording of the piece. The problem becomes more challenging when the performances differ substantially from the reference score due to interpretation and improvisation, which is very common in non-western musics such as the Turkish makam music. In this paper, we address the section linking task and present a score-informed hierarchical Hidden Markov Model for modeling musical audio signals from a coarser temporal level, where the main idea is to explicitly model the long range and hierarchical structure of music signals. In addition to having low computational complexity and achieving a transparent and elegant model, the experimental results show that our method outperforms the state-of-the-art on a Turkish makam music corpus.

Index Terms— Audio-to-score alignment, Section linking, Hierarchical hidden Markov models, Turkish makam music

1. INTRODUCTION

The problem of matching sections (section linking) to a symbolic representation is closely related to a task commonly referred to as audio-to-score alignment [1]. In audio-to-score alignment, the goal is to align each time slice in a performance recording to a note position in a symbolic musical notation of the performed piece. Instead of such a detailed alignment at the note level, section linking attempts to relate certain important structural boundaries in a reference score of a piece to their occurrences in the recording of the piece [2]. The concentration on coarse section boundaries enables a computationally lighter approach, yet section linking is still a challenging problem when the performances differ substantially from the reference score due to interpretation and improvisation, which is very common in non-western musics such as the Turkish makam music. Section linking is a key task in computational musicology, that can be used to discover music recordings in semantically meaningful and structured ways. It also renders a useful application for non-western music education, where matching scores and performances is not straightforward for the students due to the non-standard notation. Furthermore, it can also be regarded as a preprocessing step for a subsequent finer note-to-note alignment. This way exact alignment can be computed only in sections where it is demanded by the user.

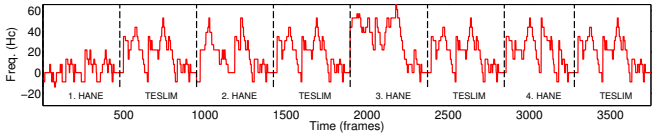
State-of-the-art approaches for audio-to-score alignment can be roughly categorized into two classes. The first group approaches the problem by means of Dynamic Time Warping (DTW), which applies dynamic programming in order to minimize a matching function between a score and an audio representation. Recently, such approaches were refined to cope with structural deviations from the notation by the performer(s) [3]. The second group of approaches tackle the problem using a probabilistic framework. In [4], a hidden Markov model (HMM) is proposed, where the tempo and score position are represented as latent variables, and the inference of the tempo-dependent score position is performed using Viterbi decoding. In [5], timed events are modeled using a hierarchical hidden Markov model (HHMM) with notes as states. The duration of timed events interacts with the estimation of the tempo that is preformed by an oscillator based model. Inference in this model is performed using causal inference, since the goal is real-time score following in live performances. A perspective on audio-to-score alignment using Conditional Random Fields is taken by the authors of [6]. They propose models of various complexities, with the most performant model resembling a HHMM with the duration of note events influenced by an additional tempo variable. Similar to [5], note events are modeled as states with related duration variables. They propose a set of observations that influence the various hidden variables of the model, and suggest pruning methods in order to be able to cope with the exact inference in the models.

In this paper we present a score-informed hierarchical Hidden Markov Model for modeling musical audio signals from a coarser temporal level, where the main idea is to explicitly model the long range and hierarchical structure of music signals. Since we aim to link the scores and the performance in the section-level and not directly aim at a note-to-note alignment, we avoid modeling strategies as presented in [5, 6] and come up with a computationally lighter but precise model for section linking.

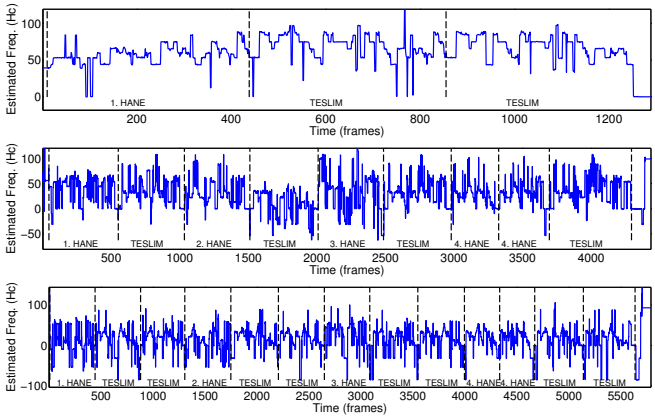
As for note-to-note alignment, section linking is applicable in musical contexts that make use of notation. In the Music Information Retrieval (MIR) literature, the context of alignment tasks has predominantly been Eurogenetic classical and popular music. However, here, as in [2] we wish to focus on Turkish makam music. This music, as we shall detail in the following section, deviates significantly from the notation on the note level by introducing a manifold of ornamentations. The large amount of ornamentations is likely to cause problems for systems targeted at note-to-note alignment, since they typically assume transitions from one note in the score to the next, something that is frequently violated for Turkish makam music. Hence, apart from reducing complexity and achieving a transparent and elegant model, proposing a probabilistic approach for pursuing alignment on a high level, *section linking*, is further motivated by the musical properties of the repertoire.

The rest of the paper is structured as follows; Section 2 explains the music collection used for evaluation and the applied preprocessing steps. Thereafter, the model is introduced in Section 3, and the

This work is supported by a Marie Curie Intra-European Fellowship (grant number 328379).



(a) The reference score with annotated sections.



(b) Different performances of the same piece. The fundamental frequencies are estimated by using [9].

Fig. 1: An example piece from the corpus: Uşşak Saz Semaisi by Neyzen Aziz Dede. The dashed vertical lines represent the section boundaries.

experimental results along with the applied evaluation methods are explained in Section 4. Section 5 concludes the paper.

2. MUSIC CORPUS

We derive the evaluation data used in this paper from the dataset described in [2]. The evaluation data consists of 166 complete performances of instrumental pieces from the Turkish makam repertoire. For each performance a machine-readable notation is available from the collection presented in [7]. In each notation, the onsets of sections in the compositions are annotated. Typically, the compositions consist of four non-repeating sections called *hane*, with a repeating section referred to as *teslim* in between them. The notations are strictly monophonic, and describe the core melody of the piece. The performances containing more than one instrument, however, cannot be considered as strictly monophonic but represent a typical example of heterophony; usually one instrument takes a higher degree of freedom to ornament the basic melody. In pieces with one instrument, the basic (notated) melody is enriched by using additional notes, too. For this reason, the number of played notes is usually significantly higher than the number of notes found in a score. While notation in Eurogenetic music divides an octave into 12 equal steps, Turkish makam music commonly conceptualized with a division of the octave into 53 steps [8]. One of these steps is referred to as Holderian comma (Hc), and the notation makes use of this resolution with the tonic of the pieces notated as 0Hc.

As detailed in [8], a performance of a piece of Turkish makam music makes use of one of 12 different transpositions, with the choice of the transpositions depending on the preferences of the musicians. For that reason the pitch of a note in the score is not related to a unique frequency value in Hz. We apply a fundamental frequency estimation to the recording [9] and convert the frequency values in Hz to a Hc-scale, with the tonic frequency again taking the value 0Hc. This way we eliminate the influence of transpositions ensure comparability with the notation. In our paper, we use manually annotated frequency values, but automatic approaches can be applied as well as discussed by [10]. An example piece from our corpus is shown in Figure 1. As the figure demonstrates, the performances often differ significantly from the reference score and from each other, making the linking problem challenging.

In the following sections, we will refer to the estimated fundamental frequencies in Hc as x_n, with n being the index of the analysis window of length $46.6ms$, without overlaps between the windows. The sequence of pitch values derived from the score is derived at the same frame rate for compatibility. The annotations that will be used for evaluation relate each section transition played in a performance to a position in the score.

Typically a performance is not played at the tempo denoted in the score. Therefore we apply a simple and accurate method to derive an initial value for the factor to correct for the tempo deviation between performance and notation. To this end, we follow [2] and compute a point-wise distance matrix between the pitch values of the initial 20% of the performance and the pitch values describing the first section in the score. Since a performance usually starts with the first section, this distance matrix will have some strong diagonal line segments. These are then detected using a Hough transform, and the angle of the longest continous line segment is determined. From this angle we obtain a factor F_{dur} by which the durations in the score are multiplied to arrive at an intial hypothesis of the durations of the sections according to the performance tempo. This hypothesis serves as a starting point for the model described in Section 3.

It is important to point out here that other signal representations such as Pitch-Class-Profiles are considered to be a more robust signal representation for alignment tasks than features based on fundamental frequency estimation. However, in [2] it was shown that in the targeted repertoire this does not hold, and for that reason we choose the fundamental frequencies as our signal representations.

3. THE MODEL

In this section, we present a novel probabilistic model for section-level modeling of musical audio. Our aim is to infer the section boundaries by making use of the score information. The main idea in our model is to incorporate section-level sequential and hierarchical structure of music signals into a single dynamic Bayesian network. We explicitly model different layers of the hierarchy by using a HHMM. The proposed model is flexible and can be applied to a wide range of musical genres.

We define the following discrete hidden variables:

- **Section variable:** $s_n \in D_s = \{1, \ldots, S\}$: represents all individual sections that are defined in the score, with S being the number of sections in the score. In our corpus, the typical set of sections is $D_s \equiv \{1.\text{HANE}, 2.\text{HANE}, 3.\text{HANE}, 4.\text{HANE}, \text{TESLIM}\}$. In the performances, the order of these sections and the number of times that they are played often vary. Our ultimate aim is to find the most-likely sequence of sections that are present in a performance.
- **Duration variable:** $d_n \in D_d = \{1, \ldots, D\}$: determines the 'ideal' duration of a section in time frames. Due to tempo changes, the duration of a section varies during the performance, therefore we allow a section to have D different durations within a piece.
- **Counter variable:** $c_n \in D_c = \{1, \ldots, C\}$: begins at value d_n at the beginning of a section and decrements until it hits 1 during the presence of the section. It also roughly determines which note of the given section is played at the time-frame n.

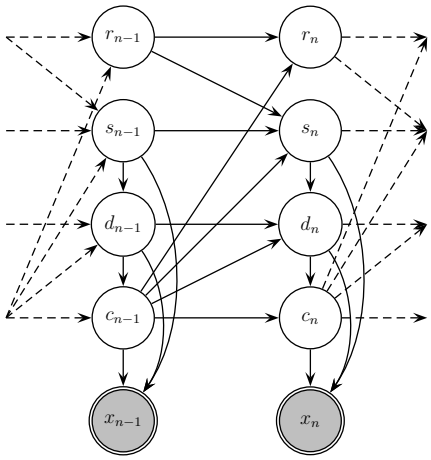


Fig. 2: Dynamic Bayesian network; The gray nodes are observed, the white nodes represent the hidden variables, and the arrows represent the conditional independence structure.

- **Repetition variable:** $r_n \in \{1, \ldots, R\}$: counts the number of consequent repetitions of a section s_n. When a section s_n starts at time n, r_n is set to 1 and r_n is incremented by 1 if the same section is performed subsequently. In our corpus, each section is allowed to repeat at most once, therefore we set $R = 2$.

The graphical model for the proposed model is given in Figure 2.

3.1. Transition Model

We start by defining the transition distribution for the counter variable as follows:

$$p(c_n|d_n, c_{n-1}) = \begin{cases} 1, & c_{n-1} = 1 \text{ and } c_n = d_n \\ 1, & c_{n-1} = 2 \text{ and } c_n = 1 \\ 1 - \omega_c, & c_{n-1} > 2 \text{ and } c_n = c_{n-1} - 1 \\ \omega_c, & c_{n-1} > 2 \text{ and } c_n = c_{n-1} - 2 \\ 0, & \text{otherwise} \end{cases}$$

This distribution chooses a step of size -1 with a probability of $1 - \omega_c$, and a step of size -2 with a probability ω_c as long as the counter has not yet reached the value 1. When it hits 1, a section boundary is reached and c_n is set to d_n, the current duration of the section s_n. This distribution enables the model to compensate for the coarse grid of the duration variable d_n, and helps to model intermediate tempo values as well as tempo instabilities within a section.

Next, we assume the following transition distribution on the repetition variables:

$$p(r_n|\cdot) = \begin{cases} 1, & c_{n-1} \neq 1 \text{ and } r_n = r_{n-1} \\ 1, & c_{n-1} = 1 \text{ and } r_{n-1} = R \text{ and } r_n = 1 \\ \omega_r, & c_{n-1} = 1 \text{ and } r_{n-1} < R \text{ and } r_n = r_{n-1} + 1 \\ 1 - \omega_r, & c_{n-1} = 1 \text{ and } r_{n-1} < R \text{ and } r_n = 1 \\ 0, & \text{otherwise} \end{cases}$$

This allows for a transition of the repetition counter only at the section boundaries ($c_{n-1} = 1$). It limits the number of section repetitions to $R - 1$ (1 in our case), and allows for a repetition with a probability of ω_r.

The transition distribution of the duration variable is defined as follows:

$$p(d_n|s_n, d_{n-1}, c_{n-1}) = \begin{cases} \delta(d_n - d_{n-1}) & , c_{n-1} \neq 1 \\ p_d(d_n|s_n) & , c_{n-1} = 1 \end{cases}$$

where $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ otherwise. Here the duration variable stays the same until c_n hits 1 and transitions to another 'duration' depending on the current section s_n. This transition is governed by $p_d(d_n|s_n)$, that is a uniform distribution over the D possible duration states.

Finally, we define the transition distribution of the section variable as follows:

$$p(s_n|s_{n-1}, c_{n-1}, r_{n-1}) = \begin{cases} \delta(s_n - s_{n-1}) & , c_{n-1} \neq 1 \\ p_s(s_n|s_{n-1}, r_{n-1}) & , c_{n-1} = 1 \end{cases}$$

This distribution is similar to the one of the duration variable: the section variable stays the same until c_n hits 1 and transitions to another section depending on the previous section s_{n-1} and the number of repetitions r_{n-1}. These transitions are governed by the distribution $p_s(s_n|s_{n-1}, r_{n-1})$ that specifies the structural properties of the musical idiom. In our case, we allow a self transition only if $r_{n-1} = 1$. Otherwise, we force a transition to a different section that is subsequent in the score. More sophisticated rules could be introduced, but this was found not to significantly improve model performance with the given data.

3.2. Observation Model

Given the current section s_n, its duration d_n, and the counter c_n, we have sufficient information to determine which note is supposed to be played at time n. We define the mapping $f(s_n, d_n, c_n)$ in such a way that it determines the true frequency of the note in the score (in Hc) played at time n. We will briefly call this mapping as f_n.

In order to compensate for octave errors that occur in the estimation of the fundamental frequency from the recording, we assume the following mixture of Gaussians as the observation model:

$$p(x_n|s_n, d_n, c_n) = \frac{1}{3} \sum_{i=1}^{3} \mathcal{N}(x_n; \mu_i, \sigma)$$

where $\mathcal{N}$ denotes the Gaussian distribution. Here $\mu_1 = f_n$, $\mu_2 = f_n - 53$, and $\mu_3 = f_n + 53$ (where 53Hc corresponds to one octave).

Note that, since all the hidden variables are discrete, we can reduce this model to an ordinary HMM and we can perform an exact inference by using the Viterbi algorithm. The most-likely state sequence provides us with the information regarding the section linking.

4. EXPERIMENTS

4.1. Methodology

We evaluate the proposed model on our annotated data corpus following evaluation procedures applied for note-to-note alignment [5], and the evaluation as performed in [2]. The Precision Pr, Recall Rc, and F-measure F are defined as follows:

$$Pr = \frac{N_{TP}}{N_{ANN}},\ Rc = \frac{N_{TP}}{N_{EST}},\ F = \frac{2 * Pr * Rc}{Pr + Rc}$$

where N_{TP} denotes the number of correctly detected section boundaries, and N_{ANN} and N_{EST} denote the number of annotated and estimated section boundaries, respectively. A section boundary detection is counted as correct only when it predicts a transition to the correct section label, and if it happens within a certain tolerance window. The size of this window T_{tol} was set to $\pm 3s$ in [2]. We chose the same size in the default setting, but we will determine how demanding higher accuracy affects system performance.

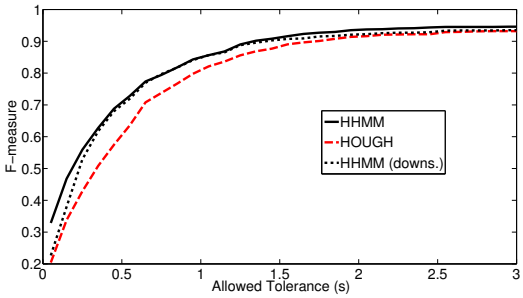


Fig. 3: Illustration of the F-measure depending on the allowed temporal tolerance.

In our experiments, the values of ω_c and ω_r were not found critical, and we arbitrarily chose $\omega_c = 0.1$ and $\omega_r = 0.5$. The value of σ was set to 0.5, which approximates a tolerance of ± 1 Hc and represents a musically meaningful tolerance value [8]. For our corpus, we allow d_n to deviate in $D = 5$ discrete steps of $[-16\%, -8\%, 0\%, 8\%, 16\%]$ from $F_{dur} * \mathbf{d}(m)$, where $\mathbf{d}(m)$ denotes the length of the m-th section in the score (see Section 2 for the duration correction factor F_{dur}).

We compare our model with the approach presented in [2]. This approach applies the same input features, but proceeds with the alignment in two steps that differ significantly from our approach. In the first stage, they obtain a list of section candidates by applying Hough transforms to similarity matrices derived from all notated sections individually compared with the performance. In a second stage, the approach proceeds with a heuristic procedure to choose between these candidates in rule based manner. While this system performed well on the Turkish makam repertoire [2] it is not straightforward to adapt it to any other repertoire. We will refer to the proposed system as HHMM and to the system presented in [2] as HOUGH in the remainder of the text.

4.2. Results

In Table 1 the performance measures of the section linking of the two compared methods are shown. With $T_{tol} = 3s$ both systems achieve performance values larger than 90%, with the differences between the two systems being statistically not significant in a pairwise t-test at a 5% significance level. When demanding, however, a higher accuracy in time, the performance of the HHMM suffer a smaller decrease than the performance of the HOUGH method. The performance at $T_{tol} = 1s$ illustrate this behavior, with the performance differences being statistically significant.

A more detailed illustration of the temporal accuracy of the two methods can be obtained from Figure 3. Using the section boundary detections from the experiments with $T_{tol} = 3s$ we determine how many of those detections would still be correct at a smaller tolerance value. It can be seen from Figure 3 that when demanding a lower tolerance, *i.e.* a decreasing misplacement between estimation and true section onset, the HOUGH method (red dashed line) is outperformed by the HHMM method (black bold line). This difference is most likely to be caused by the capability of the HHMM system to adapt to local tempo changes, compared to the HOUGH method that imposes a stable tempo throughout a section.

An apparent advantage of the HOUGH method is the faster execution time. In order to compare for this, the runtimes were recorded and the real time factors as the quotient of the execution time and the duration of the audio file were computed. The mean values of the individual real-time factors are listed in Table 2, where it is apparent

Table 1: Performance with $T_{tol} = 3s$ and $T_{tol} = 1s$

T_{tol}	Algorithm	Precision	Recall	F-measure
3s	HHMM	95.6	93.7	94.6
3s	HOUGH	94.5	92.0	93.2
1s	HHMM	85.2	84.1	84.6
1s	HOUGH	80.7	78.6	79.7

Table 2: Real-time factors

Algorithm	HHMM	HOUGH	HHMM (downs.)
Real-time factor	0.254	0.030	0.018

that the HHMM in its described parametrization is almost an order slower then the HOUGH system. It should be pointed out , however, that we did not attempt any pruning steps as proposed by [6], which would significantly speed up the inference. Instead, we experimented with a straight-forward way to reduce the size of our state-space, which is by downsampling the input data. We increased the sampling period of the data by factor 3 from $46.6ms$ to $139.8ms$ by a simple median filtering followed by a selection of every third data sample. This helps to reduce the size of the state-space since it is determined for each piece by the product $S \times R \times D \times C$[1]. This downsampling leads to a dramatic decrease of the real-time factor, as shown in the fourth column of Table 2. As the dotted black line in Figure 3 shows, this downsampling leads to a significant decrease of performance only when a tolerance of less than $300ms$ is demanded. Since the evaluation of note-to-note alignments is often performed using values around $300ms$ it is apparent that such an accuracy is sufficient for our task.

5. CONCLUSION

In this paper, we proposed a score-informed hierarchical Hidden Markov Model for modeling musical audio signals from a coarser temporal level, where the main idea is to explicitly model the long range and hierarchical structure of music signals. We address the section linking problem in Turkish makam music, which is a challenging task due to the substantial differences between the performances and the reference score. Our model enables for rapid inference while maintaining the advantages of flexibility to tempo changes and comprehensibility of the model structure. The comprehensibility of the model makes its adaptation to different repertoires a straight-forward task. Furthermore, phrasing the problem in such a probabilistic framework also enables for automatic adaptation of model parameters to new datasets.

We compared the proposed model with a rule-based approach [2] that was tailored in order to cope well with the idiosyncrasies of Turkish makam music such as micro-tonality and heterophony. Our experiments indicate that the HHMM provides a higher temporal accuracy than the rule-based model, while the inference can be sped up significantly by simple downsampling.

We plan to include further features into the proposed model, such as the consideration of rhythmical properties of the piece. Furthermore, the occurrence of improvised parts in a performance pose a problem that the HHMM in its current structure cannot deal with. Ways to cope with such conditions will be the next steps to improve the performance of the model in a general context.

[1] For our dataset, the largest value of C decreases from 3975 to 1326. A typical value for S is 8.

6. REFERENCES

[1] Meinard Müller, Daniel P. W. Ellis, Anssi Klapuri, and Gaël Richard, "Signal processing for music analysis," *J. Sel. Topics Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[2] Sertan Şentürk, Andre Holzapfel, and Xavier Serra, "Linking scores and audio recordings in makam music of Turkey," *Journal for New Music Research*, vol. 43, no. 1, pp. 34–52, 2014.

[3] Christian Fremerey, Meinard Mller, and Michael Clausen, "Handling repeats and jumps in score-performance synchronization," in *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2010, pp. 243–248.

[4] Paul H. Peeling, Ali Taylan Cemgil, and Simon J. Godsill, "A probabilistic framework for matching music representations.," in *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2007, pp. 267–272.

[5] Arshia Cont, "A coupled duration-focused architecture for real-time music-to-score alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 974–987, 2010.

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

[7] Kemal Karaosmanoğlu, "A Turkish makam music symbolic database for Music Information Retrieval: Symbtr," in *Proc. of ISMIR - International Conference on Music Information Retrieval*, Porto, Portugal, 2012.

[8] Barıs Bozkurt, Ruhi Ayangil, and Andre Holzapfel, "Computational analysis of makam music in Turkey: review of state-of-the-art and challenges," *Journal for New Music Research*, vol. 43, no. 1, pp. 3–23, 2014.

[9] Justin Salamon, Emilia Gómez, Daniel P. W. Ellis, and Gaël Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 2, pp. 118–134, 2014.

[10] Sertan Şentürk, Sankalp Gulati, and Xavier Serra, "Score informed tonic identification for makam music of Turkey," in *Proc. of ISMIR - International Conference on Music Information Retrieval*, Curitiba, Brazil, 2013, pp. 175–180.