"This is an Author's Original Manuscript of an Article whose final and definitive form, the Version of Record, has been published in the Journal of New Music Research, Volume 43, Issue 1, 31 Mar 2014, available online at: http://www.tandfonline.com/doi/full/10.1080/09298215.2013.864681."

Linking Scores and Audio Recordings in Makam Music of Turkey

Sertan Şentürk[a], André Holzapfel[a,b], Xavier Serra[a]

(sertan.senturk, andre.holzapfel, xavier.serra)@upf.edu,

[a] *Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain*
[b] *Boğaziçi University, Istanbul, Turkey.*

Abstract

The most relevant representations of music are notations and audio recordings, each of which emphasizes a particular perspective and promotes different approximations in the analysis and understanding of music. Linking these two representations and analyzing them jointly should help to better study many musical facets by being able to combine complementary analysis methodologies. In order to develop accurate linking methods, we have to take into account the specificities of a given type of music. In this paper, we present a method for linking musically relevant sections in a score of a piece from *makam* music of Turkey (MMT) to the corresponding time intervals of an audio recording of the same piece. The method starts by extracting relevant features from the score and from the audio recording. The features of a given score section are compared with the features of the audio recording to find the candidate links in the audio for that score section. Next, using the sequential section information stored in the score, it selects the most likely links. The method is tested on a dataset consisting of instrumental and vocal compositions of MMT, achieving 92.1% and 96.9% F_1-scores on the instrumental and vocal pieces, respectively. Our results show the importance of culture-specific and knowledge-based approaches in music information processing.

Keywords: Music Information Retrieval, Knowledge-Based Methodologies, Multi-Modality, Culture Specificity, Hough Transform, Directed Acyclic Graphs, Variable-Length Markov Models, Makam Music of Turkey

1. Introduction

Music is a complex phenomenon and there are many types of data sources that can be used to study it, such as audio recordings, scores, videos, lyrics and social tags. At the same time,

for a given piece there might be many versions for each type of data, for example we find cover songs, various orchestrations and diverse lyrics in multiple languages. Each type of data source offers different ways to study, experience and appreciate music. If the different information sources of a given piece are *linked* with each other (Thomas et al., 2012), we can take advantage of their complementary aspects to study musical phenomena that might be hard or impossible to investigate if we have to study the various data sources separately.

The linking of the different information sources can be done at different time spans, e.g. linking entire documents (Ellis and Poliner, 2007; Martin et al., 2009; Serrà et al., 2009), structural elements (Müller and Ewert, 2008), musical phrases (Wang, 2003; Pikrakis et al., 2003), or at note/phoneme level (Niedermayer, 2012; Fujihara and Goto, 2012). Moreover there might be substantial differences between the information sources (even among the ones of the same type) such as the format of the data, level of detail and genre/culture-specific characteristics. Thus, we need content-based (Casey et al., 2008), application-specific and knowledge-driven methodologies to obtain meaningful features and relationships between the information sources. The current state of the art in Music Information Retrieval (MIR) is mainly focussed on Eurogenetic [1] styles of music (Tzanetakis et al., 2007) and we need to develop methodologies that incorporate culture-related knowledge to understand and analyze the characteristics of other musical traditions (Holzapfel, 2010; Şentürk, 2011; Serra, 2011).

In analyzing a music piece, scores provide an easily accessible symbolic description of many relevant musical components. The audio recordings can provide information about the characteristics (e.g. in terms of dynamics or timing) of an interpretation of a particular piece. Parallel information extracted from score and audio recordings may facilitate computational tasks such as version detection (Arzt et al., 2012), source separation (Ewert and Müller, 2012), automatic accompaniment (Cont, 2010) and intonation analysis (Devaney et al., 2012).

In this paper, we focus on marking the time intervals in the audio recording of a piece with the musically relevant structural elements (sections) marked in the score of the same piece (or briefly "section linking"). The proposed method extracts features from the audio recording and the sections in the score. From these features, similarity matrices are computed for each section. The method applies Hough transform (Duda and Hart, 1972) to the similarity matrices in order to detect section candidates. Then, it selects between these candidates by searching through the paths, which reflect the sequence of sections implied by the musical form, in a directed acyclic graph (DAG). We optimize the method for the cultural-specific aspects of makam music of Turkey (MMT). By *linking* score sections with the corresponding fragments in the audio recordings, computational operations that are specific to this type of music, such as *makam* recognition (Gedik and Bozkurt, 2010), tuning analysis (Bozkurt et al., 2009) and rhythm analysis can be done at the section level, providing a deeper insight into the structural, melodic or metrical properties of the music.

[1] We apply this term because we want to avoid the misleading dichotomy of Western and non-Western music.

The remainder of the paper is structured as follows: Section 2 gives an overview of related computational research. Section 3 makes a brief introduction to *makam* music of Turkey. Section 5 makes a formal definition of *section linking* and gives an overview of proposed methodology. Sections 6-8 explains the proposed methodology in detail. Section 4 presents the dataset used to test the methodology. Section 9 presents the experiments carried out to evaluate the method and the results obtained from the experiments. Section 10 gives a discussion on the results, and Section 11 concludes the paper.

Throughout the text, in the data collection and in the supplementary results, we use MusicBrainz Identifier (MBID) as an unique identifier for the compositions and audio recordings. For more information on MBIDs please refer to `http://musicbrainz.org/doc/MusicBrainz_Identifier`.

2. State of the Art

A relevant task to section linking is *audio-score alignment*, i.e. linking score and audio on the note or measure level. Generally, if score and audio recording of a piece are linked on the note or measure level, section borders in the audio can be obtained from the time stamps of the linked notes/measures in the score and audio (Thomas et al., 2012). The current state-of-the-art on audio-score alignment follows two main approaches: hidden Markov models (HMM) (Cont, 2010) and dynamic time warping (DTW) (Niedermayer, 2012) . In general, approaches of audio-score alignment assumes that the score and the target audio recording are structurally identical, i.e. there are no phrase repetitions and omissions in the performance. Fremerey et al. (2010) extended the classical DTW and introduced JumpDTW, which is able to handle such structural non-linearities. However, due to the its level of granularity, audio-score alignment is computationally expensive.

Since section linking is aimed at linking score and audio recordings on the level of structural elements, it is closely related to audio structure analysis (Paulus et al., 2010). The state of the art methods on structure analysis are mostly aimed at segmenting audio recordings of popular Eurogenetic music into repeating and mutually exclusive sections. For such segmentation tasks, self-similarity analysis (Cooper and Foote, 2002; Goto, 2003) is typically employed. These methods first compute a series of frame-based audio features from the signal. Then all mutual similarities between the features are calculated and stored in a so-called self similarity matrix, where each element describes the mutual similarity between the temporal frames. In the resulting square matrix, repetitions cause parallel lines to the diagonal with 45 degrees and rectangular patterns in the similarity matrix. This directional constraint makes it possible to identify the repetitions and 2-D sub-patterns inside the matrix.

When fragments of audio or score are to be linked, the angle of the diagonal lines in the similarity matrix computed are not 45 degrees, unless the tempi of both information sources are exactly the same. This problem also occurs in cover song identification (Ellis and Poliner, 2007;

Serrà et al., 2009) for which a similarity matrix is computed using temporal features obtained from a cover song candidate and the original recording. If the similarity matrix is found to have some strong regularities, they are deemed as two different versions of the same piece of music. A proposed solution is to "squarize" the similarity matrix by computing some hypothesis about the tempo difference (Ellis and Poliner, 2007). However, tempo analysis in makam musics is not a straightforward task (Holzapfel and Stylianou, 2009). The sections may also be found by traversing the similarity matrices using dynamic programming (Serrà et al., 2009). On the other hand, dynamic programming is a computationally demanding task.

Since the sections in a composition follow a certain sequential order, the extracted information can be formulated as a directed acyclic graph (DAG) (Newman, 2010). Paulus and Klapuri (2009) use this concept in self-similarity analysis. They generate a number of border candidates for the sections in the audio recording and create a DAG from all possible border candidates. Then, they use a greedy search algorithm to divide the audio recording into sections.

3. Makam Music of Turkey

The melodic structure of most traditional music repertoires of Turkey is interpreted using the concept of *makam*s. Makams are modal structures, where the melodies typically revolve around a *başlangıç* (starting, initial) tone and a *karar* (ending, final) tone (Ederer, 2011). The pitch intervals cannot be expressed using a 12-TET system (tone equal tempered), and there are a number of different transpositions (*ahenk*) any of which might be favored over others due to instrument/vocal range or aesthetic concerns (Ederer, 2011).

Currently Arel-Ezgi-Uzdilek (AEU) theory is the mainstream theory used to explain makam music of Turkey (MMT) (Özkan, 2006). AEU theory divides a whole tone into 9 equidistant intervals. These intervals can be approximated by 53-TET (tone equal tempered) intervals, each of which is termed as a *Holderian comma* (1 Hc $= \frac{1200}{53} \approx 22.64$ cents) (Ederer, 2011). AEU theory defines the values of intervals based on Holderian commas (Tura, 1988), whereas the performers typically change the intervals from makam to makam and according to personal preferences (Ederer, 2011). Bozkurt et al. (2009) have analyzed selected pieces from renowned musicians to assess the tunings in different makams, and showed that the current music theories are not able to explain these differences well.

For centuries, MMT has been predominantly an oral tradition. In the early 20th century, a score representation extending the traditional Western music notation was proposed and since then it has become a fundamental complement to the oral tradition (Popescu-Judetz, 1996). The extended Western notation typically follows the rules of Arel-Ezgi-Uzdilek theory. The scores tend to notate simple melodic lines but the performers extend them considerably. These deviations include expressive timings, adding note repetitions and non-notated embellishments. The intonation of some intervals in the performance might differ from the notated intervals as much as a

semi-tone (Signell, 1986). The performers (including voice in vocal compositions) usually perform simultaneous variations of the same melody in their own register, a phenomenon commonly referred to as heterophony (Cooke, 2013). These heterophonic interactions are not indicated in the scores. Regarding the structure of pieces, there might be section repetitions or omissions, and taksims (instrumental improvisations) in the performances.

In the paper, we focus on *peşrev*, *saz semaisi* (the two most common instrumental forms) and *şarkı* (the most common vocal form) forms. Peşrev and saz semaisi commonly consists of four distinct *hane*s and a *teslim* section, which typically follow a *verse-refrain*-like structure. Nevertheless, there are peşrevs, which have no teslim, in which case the second half of each hane strongly resembles each other (Karadeniz, 1984). The 4^{th} hane in the saz semaisi form is usually longer, includes rhythmic changes and it might be divided into smaller substructures. Each of these substructures might have a different tempo with respect to the overall tempo of the piece. There is typically no lead instrument in instrumental performances.

A şarkı is typically divided into sections called *aranağme*, *zemin*, *nakarat* and *meyan*. The typical order of the sections is aranağme, zemin, nakarat, meyan and nakarat. Except of the instrumental introduction aranağme, all the sections are vocal and determined by the lines of the lyrics. Each line in the lyrics is usually repeated, but the melody in the repetition might be different. Vocals typically lead the melody; nonetheless heterophony is retained. Some şarkıs have a *gazel* section (vocal improvisation), for which the lyrics are provided in the score, without any melody.

4. Data Collection

For our experiments, we collected 200 audio recordings of 44 instrumental compositions (preşrevs and saz semaisis), and 57 audio recordings of 14 vocal compositions (şarkıs) (i.e. 257 audio recordings of 58 compositions in total). The makam of each composition is included in the metadata.[2] The pieces cover 27 different makams.

The scores are taken from the symbTr database (Karaosmanoğlu, 2012), a database of makam music compositions, given in a specific text format, as well as PDF and as MIDI. The scores in text form are in the machine readable symbTr format (Karaosmanoğlu, 2012), which contains note values on 53-TET resolution and note durations. These symbTr-scores are divided into sections that represent structural elements in makam music (Section 3). The beginning and ending notes of each section are indicated in the instrumental symbTr-scores. In the vocal compositions the sections can be obtained from the lyrics and the melody indicated in the symbTr-score. In this paper we manually label each section in the vocal compositions according to these. The section sequence indicated in the PDF formats is found in the symbTr-scores and MIDI files as well (i.e.

[2]The metadata is stored in MusicBrainz: `http://musicbrainz.org/collection/5bfb724f-7e74-45fe-9beb-3e3bdb1a119e`

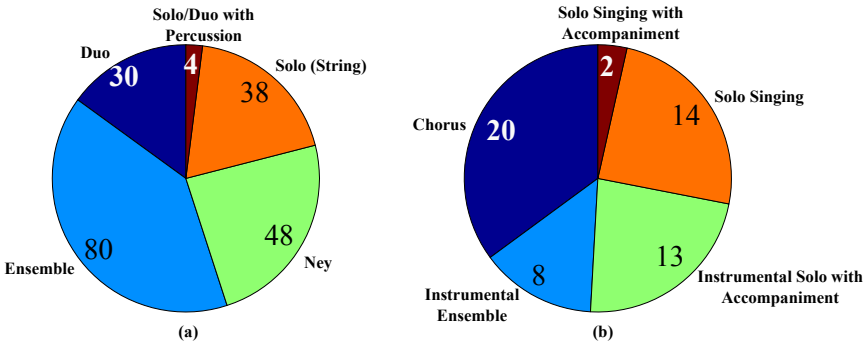

Figure 1: Instrumentation and voicing in the dataset **a)** Instrumentation in the peşrevs and saz semaisis **b)** Voicing in the şarkıs

following the lyric lines, the repetitions, volta brackets, coda signs etc. in the PDF). The duration of the notes in the MIDI and symbTr-score are stored according to the tempo given in the PDF. We divided the MIDI files manually according to the section sequence given in the symbTr-scores. MIDI files include the microtonal information in the form of pitch-bends.

Three peşrevs (associated with 13 recordings) do not have a teslim section in the composition but each section has very similar endings (Section 3). Nine peşrevs (associated with 40 recordings) have less than 4 hanes in the scores. There are notated tempo changes in the 4^{th} hanes of four saz semaisi compositions (in the PDF), and the note durations in the related sections in the symbTr-scores reflect these changes. In most of the şarkıs each line of the lyrics is repeated. Nevertheless, the repetition occasionally comes with a different melody, effectively forming two distinct sections. Two şarkı compositions include gazel sections (vocal improvisations).

The audio recordings are stored in *mp3* format and the sampling rate is 44100 Hz. They are selected from the CompMusic collection,[3] and they are either in public-domain or commercially available. The ground truth is obtained by manually annotating the timings of all sections performed in the audio recordings. There are 1457 and 638 sections performed in the recordings of the instrumental and vocal compositions, respectively (a total of 2095 sections). In all the audio recordings, a section is repeated in succession at most twice. The mean and standard deviation of the duration of each section in the audio recordings are 35.17 and 19.49 seconds for instrumental, and 13.47 and 6.17 seconds for vocal pieces, respectively.

The performances contain tempo changes, varying frequency and kinds of embellishments, and inserted/omitted notes. There are also repeated or omitted phrases inside the sections in the audio recordings. Heterophonic interactions occur between instruments played in different octaves. Figure 1a,b shows the instrumentation and voicing of the audio recordings in the dataset. Among the audio recordings of instrumental compositions, ney recordings are monophonic. They are mostly from the "Instrumental Pieces Played with the Ney" collection (43 recordings),[4] and

[3] `http://compmusic.upf.edu/`.

[4] `http://neyzen.com/ney_den_saz_eserleri.htm`

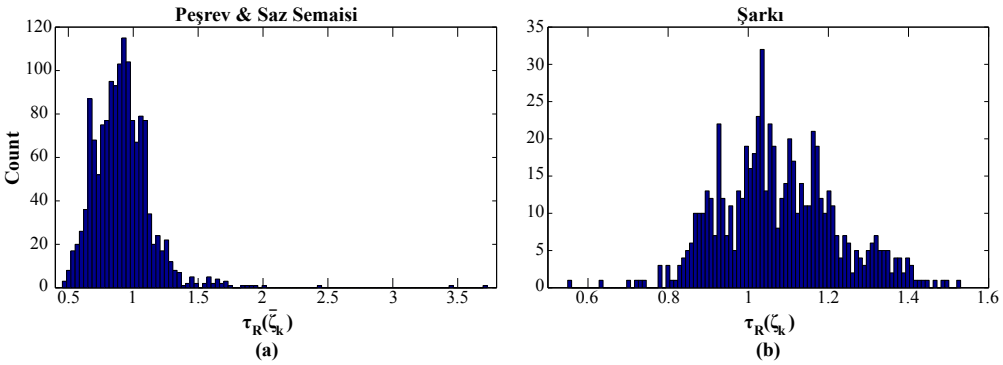

Figure 2: Histograms of relative tempo τ_R in the dataset **a)** Peşrevs and saz semaisis **b)** Şarkıs

performed very similar to the score tempo and without phrase repetitions/omissions. From solo stringed recordings to ensemble recordings the density of heterophony typically increases. All audio recordings of vocal compositions are heterophonic. Hence the dataset represents both the monophonic and the heterophonic expressions in makam music. The ahenk (transposition) varies from recording to recording, which means that the tonic frequency (karar) varies even between interpretations of the same composition. Some of the recordings include material that is not related to any section in the score, such as taksims (non-metered improvisations), applauses, introductory speeches, silence and even other pieces of music. The number of segments labelled as "unrelated" is 220.[5]

We computed the distribution of the relative tempo, which was obtained by dividing the durations of sections in a score by the duration of its occurance in a perfromance. Figure 2 shows all the occured quotients for the annotated sections in the audio recordings in the dataset. The outliers seen in Figure 2a are typically related to performances which omit part of a section, and 4^{th} hanes, which tend to deviate strongly from the annotated tempo. As can be seen from Figure 2, the tempo deviations are roughly Gaussian distributed, with a range of quotients [0.5 1.5] covering almost all observations. This will help us to reduce the search space of our algorithm in Section 7.

5. Problem Definition and Methodology

We define *section linking* as "marking the time intervals in the audio recording at which musically relevant structural elements (sections) given in the score are performed." In this task, we start with a score and an audio recording of a music piece. The score and audio recording are known to be related with the same work (composition) via available metadata, i.e. they are already linked with each other in the document-level.

The score includes the notes, and it is divided into sections, some of which are repeated. These sections are known, and the start and end of each section are provided in the score, including the

[5]The score data, annotations and results are available in `http://compmusic.upf.edu/node/171`.

compositional repetitions. Therefore, we do not need any structural analysis to find the structural elements. From the start and end of each section, the sequence of the sections are known. The tempo and the makam of the piece are also available in the score. The audio recording follows the section sequence given in the score with possible section insertions, omissions, repetitions and substitutions. Moreover the performance might include various expressive decisions such as musical material that are not related to the piece, phrase repetitions/omissions, pitch deviations.

A formal definition of the problem follows:

1. Let $\mathcal{S} = \{\mathcal{S}_s, u\}$ denote the *set of section symbols*. It consists of a set of symbols $\mathcal{S}_s = \{s_1, \ldots, s_N\}$, which represents all the N possible distinct sections in a composition; and an *unrelated section*, u, i.e. a segment with content not related to any structural element of the musical form. The number of unique sections is $|\mathcal{S}| = N + 1$.
2. The sections in the score form the *score section symbol sequence*, $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_M]$, where $\sigma_m \in \mathcal{S}_s$ and $m \in [1 : M]$, with M being the number of sections in a score, repeated sections are counted individually.
3. We define the score section sequence $\bar{\boldsymbol{\sigma}} = [\bar{\sigma}_1, \ldots, \bar{\sigma}_M]$, with each $\bar{\sigma}_m$ consisting of a *section symbol*, σ_m, and a sequence of $\langle$*note-name*, *duration*$\rangle$ tuples, which represents the monophonic melody of the section. The $\langle$*note-name*, *duration*$\rangle$ tuples of the repetitive sections do not have to be identical due to different ending measures, volta brackets etc.
4. For each performance we have the (true) *audio section symbol sequence*, $\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_K]$, where $\zeta_k \in \mathcal{S}$, $k \in [1 : K]$, with K being the number of sections in the performance, including possibly multiple unrelated sections.
5. Analogous, for each performance we have the (true) *audio section sequence*, $\bar{\boldsymbol{\zeta}} = [\bar{\zeta}_1, \ldots, \bar{\zeta}_K]$, $k \in [1 : K]$. Each element of the sequence, $\bar{\zeta}_k$, has the section symbol, ζ_k, and covers a time interval in the audio, $t(\bar{\zeta}_k)$, i.e. $\bar{\zeta}_k = \big\langle \zeta_k,\ t(\bar{\zeta}_k) \big\rangle$. The time interval is given as $t(\bar{\zeta}_k) = \big[t_{ini}(\bar{\zeta}_k)\ t_{end}(\bar{\zeta}_k)\big]$, where $t_{ini}(\bar{\zeta}_1) = 0$ sec; $t_{end}(\bar{\zeta}_k) = t_{ini}(\bar{\zeta}_{k+1}), k \in [1 : K-1]$; and $t_{end}(\bar{\zeta}_K)$ refers to the end of the audio recording.
6. We will apply our method to obtain the (estimated) audio section sequence $\bar{\boldsymbol{\pi}}$ in the audio recording, where each *section link*, $\bar{\pi}_k = \langle \pi_k,\ t(\bar{\pi}_k) \rangle$, in the sequence is paired with a section symbol in the composition $s_n \in \mathcal{S}_s$ or the unrelated section u. Ideally, the *audio section sequence*, $\bar{\boldsymbol{\zeta}}$, and *section link sequence*, $\bar{\boldsymbol{\pi}}$ should be identical.

Given the score representation of a composition and the audio recording of the performance of the same composition, the procedure to link the sections of a score with the corresponding sections in the audio recording is as follows:

1. Features are computed from the audio recording and the musically relevant sections ($\forall s_n \in \mathcal{S}_s$) of the score (Section 6).

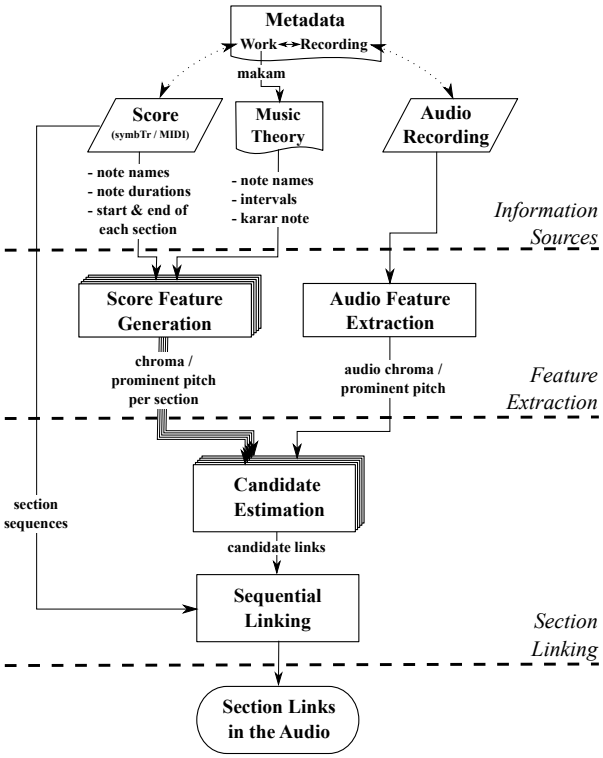

Figure 3: Block Diagram of the Section Linking Methodology

2. A similarity matrix $B(s_n)$ is computed for each section s_n, measuring the similarity between the score features of the particular section and the audio features of the whole recording. By applying Hough transform to the similarity matrices, candidate links $\bar{\pi}_k$, where $\pi_k = s_n \in \mathcal{S}_s$, are estimated in the audio recording for each section given in the score (Section 7).
3. Treating the candidate links as labeled vertices, a directed acyclic graph (DAG) is generated. Using section sequence information ($\bar{\boldsymbol{\sigma}}$) given in the score, all possible paths in the DAG are searched and the most-likely candidates are identified. Then, the non-estimated time intervals are guessed. The final links are marked as section links (Section 8).

From music-theory knowledge, we generate a dictionary consisting $\langle makam, karar \rangle$ pairs, which stores the karar of each makam (e.g. if the makam of the piece is Hicaz, the karar is A4.). The karar note is used as the reference symbol during the generation of score features for each section (Section 6.1). We also apply the theoretical intervals for a makam as defined in AEU theory to generate the score features from the machine-readable score (Section 6.1). By incorporating makam music knowledge, and considering culture-specific aspects of the makam music practice (such as pitch deviations and heterophony), we specialize the section linking methodology to makam music of Turkey.

6. Feature Extraction

Score and audio recording are different ways to represent music. Figure 4a-b shows the score and an audio waveform[6] of the first nakarat section of the composition, *Gel Güzelim*[7]. To compare these information sources, we extract features that capture the melodic content given in each representation. In our methodology, we utilize two types of features: chroma (Gómez, 2006; Müller, 2007) and prominent pitch. Chroma features are the state of the art features used in structure analysis of Eurogenetic musics (Paulus et al., 2010) and also in relevant tasks such as version identification (Serrà et al., 2009) and audio-score alignment (Thomas et al., 2012). We use Harmonic Pitch Class Profiles (HPCPs), which were shown to be robust feature for tonal musics (Gómez, 2006). On the other hand, prominent pitch might be a more accurate feature due to the monophonic nature of melodies given in the score and the heterophonic performance practice (Section 3). In the preliminary experiments (Şentürk et al., 2012), we used YIN (De Cheveigné and Kawahara, 2002) and found that monophonic pitch extractors are not able to provide reliable pitch estimations due to the heterophonic and expressive characteristics of MMT. Instead we use the melody extraction algorithm proposed by Salamon and Gómez (2012), which was shown to outperform other state of the art melody extraction algorithms.

We compare prominent pitches and HPCPs as input features for a section linking operation. There are some differences in the methodology using prominent pitches or HPCPs in the feature computation, which will be described in detail now.

6.1. Score Feature Extraction

To compute the score features, we use a machine readable score, which stores the value and the duration (i.e. the $\langle$*note-name*, *duration*$\rangle$ tuple) of each note. The format of the score is chosen either as a MIDI or a text file according to the feature to be computed (HPCPs or prominent pitches, respectively). Both the symbolic representations contain information about the structure of the composition, i.e. the score section sequence $\bar{\boldsymbol{\sigma}}$, as well. In the text-scores, the indices of the initial and final note are given for each section. In the MIDI-scores, the initial and final time-stamps (in seconds) are given for each section. The note values in the MIDI files also include the microtonal information (see Section 4).

To compute the synthetic prominent pitches per section from the text-score, we select the first occurrence of the section $s_n \in \mathcal{S}_s$, in the score section symbol sequence $\boldsymbol{\sigma}$ and extract the corresponding $\langle$*note-name*, *duration*$\rangle$ tuple sequence from $\bar{\boldsymbol{\sigma}}$. The sum of the durations in the tuples is assigned to the duration of the score section $d(s_n)$. Then we note the makam of the composition, which is given in the score, and obtain the karar-name of the piece by checking the makam in the $\langle makam, karar\rangle$ dictionary. The note-names are mapped to the Hc distances according to AEU

[6]MBID: e7be8c2a-3309-4106-93b7-76cd6102a924
[7]MBID: 9aaf5c0b-4642-40fd-97ba-c861265872ce

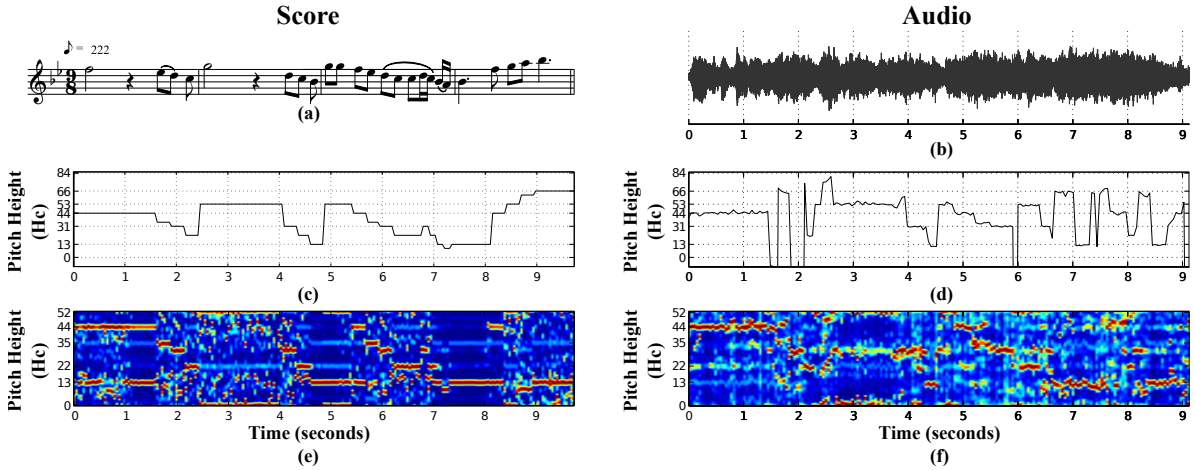

Figure 4: Score and audio representations of the first nakarat section of *Gel Güzelim* and the features computed from these representations. a) Score. b) Annotated section in the audio recording. c) Synthetic prominent pitch computed from the note symbols and durations. d) Prominent pitch computed from the audio recording. The end of the prominent pitch has a considerable number of octave errors. e) HPCPs computed from the synthesized MIDI. f) HPCPs computed from the audio recording.

theory with reference to the karar note. As an example see Figure 4b: here the karar note is G4 (Nihavent makam) and all the notes take on values in relation to that karar, as for instance 13 Hc for the B4♭. In makam music practice, the notes preceding rests may be sustained for the duration of the rest.[8] For this reason, the rests in the score are ignored and their duration is added to the previous note. Finally, a synthetic prominent pitch for each section, $p(s_n)$, $s_n \in \mathcal{S}_s$, is calculated at a frame rate of ∼46 ms, which provides sufficient time resolution to track all changes in pitch in the scores.

To obtain the HPCPs, MIDI-scores are used. First, audio is generated from the MIDI-score.[9] Then, the HPCPs are computed for each section[10] (Figure 4e). We use the default parameters given in (Gómez, 2006). The hop size and the frame size are chosen to be 2048 (e.g. ∼21.5 frames per second) and 4096 samples respectively. The first bin of the HPCPs is assigned to the karar note. For comparison, HPCPs are computed with different number of bins per octave in our experiments (see Section 9). Finally, the HPCP vectors for each section, $h(s_n)$, $s_n \in \mathcal{S}_s$, are extracted by using the start and end time-stamps of each section. Note that the HPCPs contain microtonal information as well, since this information is encoded into the MIDI-scores.

[8] Notice that there are two rests in the score in Figure 4a, but the notes are sustained in the performance as seen in the audio waveform in Figure 4b.

[9] We use TiMidity++ (`http://timidity.sourceforge.net/`) with the default parameters for the audio synthesis. Since there are no standard soundfonts of makam music instruments, we select the default soundfont (grand acoustic piano: `http://freepats.zenvoid.org/sf2/`). Nevertheless the soundfont selection does not affect the HPCP computation greatly since HPCPs were reported to be robust to changes in timbre (Gómez, 2006).

[10] We use Essentia in the computation (Bogdanov et al., 2013).

6.2. Audio Feature Extraction

To obtain the prominent pitch from the audio files, we apply the melody extraction algorithm by Salamon and Gómez (2012) using the default values.[11] The approach computes the melody after separating salient melody candidates from non-salient ones. If there are no salient candidates present for a given interval, that interval is deemed to be unvoiced. However, as MMT is heterophonic (Section 3), unvoiced intervals are very rare. The algorithm using the default parameters treats a substantial amount of melody candidates as non-salient (due to the embellishments and wide dynamic range), and dismisses a significant portion of melodies as unvoiced. Hence, we include all the non-salient candidates to guess prominent pitches. In our experiments, melody extraction is performed using various pitch resolutions (Section 9).

The next step is to convert the obtained frequency values of the melody in Hz to distances in Hc with reference to the karar note. We first identify the frequency of the karar using Makam Toolbox (Gedik and Bozkurt, 2010), using our extracted melodies as input. The pitch resolution of the extracted melody used for karar identification is chosen as 0.44 Hc. The values in Hz are then converted to Hc using the karar frequency as the reference (zero) so that the computed prominent pitches are ahenk (*i.e.* transposition) independent. Finally, we obtain the *audio prominent pitch* $p(a)$, by downsampling the sequence from the default frame rate of ∼344.5 frames per second (hop size of 128 samples) to ∼21.5 frames per second or a period of ∼46 ms (Figure 4d).

The procedure of HPCP computation from the audio recording $h(a)$, is the same as explained in Section 6.1 except that the first bin of the HPCP is assigned to the karar frequency estimated by Makam Toolbox (Figure 4f).

7. Candidate Estimation

To compare the audio recording with each section in the score, we compute a distance matrix between the score feature, $p(s_n)$ or $h(s_n)$, of each section s_n and the audio feature, $p(a)$ or $h(a)$, of the whole recording, for either prominent pitches or HPCP, respectively. Next, the distance matrices are converted to binary similarity matrices (Section 7.1). Applying Hough transform to the similarity matrices, we estimate candidate time intervals in audio for each section given in the score (Section 7.2). In the remainder of the section, we use an audio recording[12] of the composition *Şedaraban Sazsemaisi*[13] for illustration.

7.1. Similarity Matrix Computation

If the prominent pitches are chosen as features, the distance matrix, $D^p(s_n)$, between the audio prominent pitch, $p(a)$, and the synthetic prominent pitch, $p(s_n)$, of a particular section, $s_n \in \mathcal{S}_s$, is

[11] We use the Essentia implementation of the algorithm (Bogdanov et al., 2013).
[12] MBID: efae832f-1b2c-4e3f-b7e6-62e08353b9b4
[13] MBID: 1eb2ca1e-249b-424c-9ff5-0e1561590257

obtained by computing the pairwise Hc distance between each point of the features, i.e. city block (L_1) distance (Krause, 1987), as:

$$D^p_{ij}(s_n) = |p_i(s_n) - p_j(a)|\,, \quad 1 \leq i \leq q \text{ and } 1 \leq j \leq r \tag{1}$$

where $p_i(s_n)$ is the i^{th} point of the synthetic prominent pitch (of length q) of a particular section, and $p_j(a)$ is the j^{th} point of the prominent pitch (of length r) extracted from the audio recording. City block distance gives us a musically relevant basis for comparison by computing how far two pitch values are apart from each othqer in Hc.

The melody extraction algorithm by Salamon and Gómez (2012) is optimized for music that has a clear separation between melody and accompaniment. Since performances of makam music (esp. instrumental) involve musicians playing the same melody in different octaves (Section 3), melody extraction algorithm by Salamon and Gómez (2012) produces a considerable number of octave jumps (Figure 4d). Therefore, the value of each point in the distance matrices, D^p_{ij}, are octave wrapped such that the distances lie between 0 and $\frac{53}{2}$ Hc, with 0 denoting exactly the same pitch class (Figure 5a).

If the HPCPs are chosen as the feature, the distance matrix, $D^h(s_n)$, between the HPCP features h(a) computed from the audio recording, and the HPCP $h(s_n)$, computed for a particular section $s_n \in \mathcal{S}_s$, is obtained by taking cosine distance between each frame. Cosine distance is a common feature used for comparing chroma features (Paulus et al., 2010), computed as:

$$D^h_{ij}(s_n) = 1 - \frac{\sum_{b=1}^{n_{bins}} h_{ib}(s_n)\, h_{jb}(a)}{\sqrt{\left(\sum_{b=1}^{n_{bins}} h^2_{ib}(s_n)\right).\left(\sum_{b=1}^{n_{bins}} h^2_{jb}(a)\right)}}, \quad 1 \leq i \leq m_s \text{ and } 1 \leq j \leq m_a \tag{2}$$

where $h_{ib}(s_n)$ is the b^{th} bin of the i^{th} frame of the HPCPs (of m_s frames) of a given section, $h_{jb}(a)$ is the b^{th} bin of the j^{th} frame of the HPCPs (of m_a frames) extracted from the audio recording and n_{bins} denotes the number of bins chosen for the HPCP computation. The outcome is bounded to the interval [0 1] for non-negative inputs, 0 denoting the "closest," which makes it possible to compare the relative distance between the frames of HPCPs that have unitless values.

In the distance matrices, there are diagonal line segments, which hint the locations of the sections in the audio (Figure 5a). However, the values of the points forming the line segments may be substantially greater than zero in practice, making it harder to distinguish the line segments from the background. Therefore, we apply binary thresholding to the distance matrices to emphasize the diagonal line segments, and obtain a binary similarity matrix $B(s_n)$ as:

$$B_{ij}(s_n) = \begin{cases} 1, & D_{ij} < \beta \\ 0, & D_{ij} \geq \beta \end{cases} \tag{3}$$

where β is the binarization threshold. The binary similarity matrix $B(s_n)$ of a section s_n shows which points between the score feature and the audio feature are similar enough to each other to

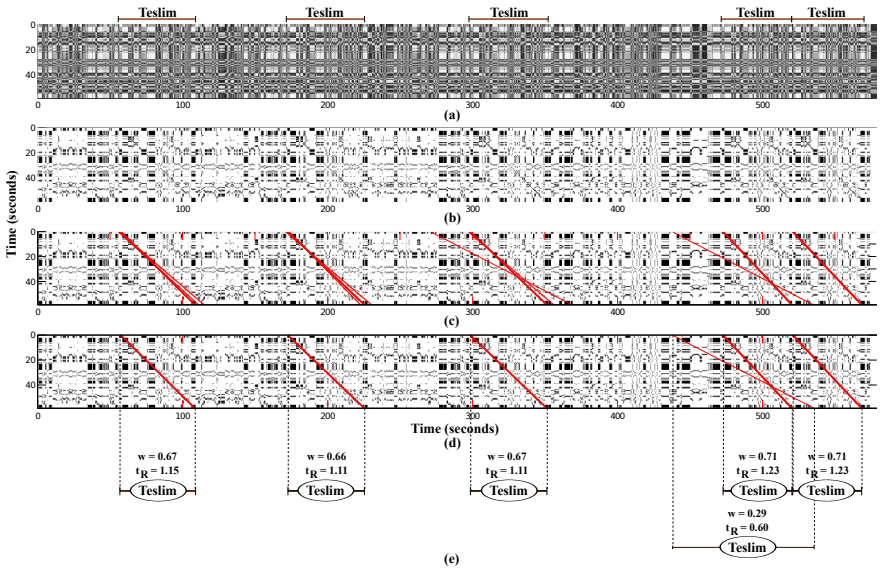

Figure 5: Candidate estimation between the teslim section of the *Şedaraban Sazsemaisi* and an audio recording of the composition shown step by step. a) Annotated teslims and the distance matrix computed from the prominent pitches. White indicates the closest distance (0 Hc). b) Image binarization on distance matrix. White and black represent zero (dissimilar) and one (similar) respectively. c) Line detection using Hough transformation. d) Elimination of duplicates. e) Teslim candidates. The numerical values "w" and "τ_R" indicate the weight and the relative tempo of the candidate respectively.

be deemed as the same note (Figure 5b). For comparison, experiments will be conducted using different binarization threshold values (Section 9).

7.2. Line Detection

After binarization, we apply Hough transform to detect the diagonal line segments (Duda and Hart, 1972). Hough transform is a common line detection algorithm, which has been also used in musical tasks such as locating the formant trajectories of drum beats (Townsend and Sandler, 1993) and detecting repetitive structures in an audio recording for thumbnailing (Aucouturier and Sandler, 2002). The projection of a line segment found by the Hough transform to the time-axis would give an estimated the time-interval $t(\bar{\pi}_k)$ of the candidate section link $\bar{\pi}_k$.

The angle of a diagonal line segment is related to the tempo of the performed section $\tau(\bar{\pi}_k)$ and the tempo of the respective section given in the score $\tau(s_n)$, $\pi_k = s_n$. We define the *relative tempo* for each candidate $\tau_R(\bar{\pi}_k)$ as:

$$\tau_R(\bar{\pi}_k) = \tan(\theta) = \frac{d(s_n)}{|t(\bar{\pi}_k)|} \approx \frac{\tau(\bar{\pi}_k)}{\tau(s_n)}, \quad \pi_k = s_n \tag{4}$$

where $d(s_n)$ is the duration of the section given in the score, $|t(\bar{\pi}_k)|$ is the duration of the candidate section link $\bar{\pi}_k$ and θ is the angle of the line segment associated with the candidate section link. Provided that there are no phrase repetitions, omissions or substantial tempo changes inside the performed section, relative tempo approximately indicates the amount of deviation from the tempo

given in the score. If the tempo of the performance is exactly the same with the tempo, the angle of the diagonal line segment is 45°.

In order to restrict the angles searched in the Hough transform to an interval $[\theta_{min},\ \theta_{max}]$, we computed the relative tempo of all the true section links $\tau_R(\bar{\zeta}_k)$ in the dataset (see Section 4). We constrain the relative tempo $\tau_R(\bar{\pi}_k)$ of a section candidate between 0.5 and 1.5, covering most of the observed tempo distribution. This limits the searched angles in the Hough transform between:

$$[\theta_{min},\ \theta_{max}] = \begin{cases} \theta_{min} = \arctan(0.5) \approx 27^\circ \\ \theta_{max} = \arctan(1.5) \approx 56^\circ \end{cases} \tag{5}$$

The step size of the angles between θ_{min} and θ_{max} is set to 1 degree.

Since some of the sections (such as teslims and nakarats) are repeated throughout the composition (Section 3) and sections may be repeated twice in succession (Section 4), a particular section may be performed at most 8 times throughout a piece. Considering the maximum number of repetitions plus a tolerance of 50%, we pick the highest 12 points in the Hough transform, which show the angle and the distance to the origin of the most prominent line segments. Next, the line segments are computed from this set of points such that the line segment covers the entire duration of the section given in the score (Figure 5c). The number of non-zero pixels forming the line segment is normalized by the length of the line segment, giving the weight $w(\bar{\pi}_k)$ of the segment.

Finally, if two or more line segments have their borders in the same vicinity (±6 seconds), they are treated as duplicates. This occurs frequently because the line segments in the binary matrix are actually blobs. Hence, there might be line segments with slightly different parameters, effectively estimating the same candidate. Among the duplicates, only the one with the highest weight is kept (Figure 5d). The regions covered by the remaining lines are chosen as the candidate time intervals, $t(\bar{\pi}_k) = [t_{ini}(\bar{\pi}_k)\ t_{end}(\bar{\pi}_k)]$ in seconds, for the particular section (Figure 5e).

This operation is done for each section, $s_n \in \mathcal{S}_s$, obtaining candidate section links $\bar{\pi}_k$, $\pi_k = s_n \in \mathcal{S}_s$ (Figure 6b).

8. Sequential Linking

By inspecting Figures 6a and 6b, it can be seen that all ground truth annotations are among the detected candidates, with problems in the alignment of 4^{th} hane. However, as there are also many false positives, we use knowledge about the structure of the composition to improve the candidate selection. Considering the candidate links as vertices in a DAG, we first extract all possible paths from the DAG according to the score section symbol sequence $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_M]$ (Section 8.1). We then decide the most likely paths (Section 8.2). Finally, we attempt to guess non-estimated time intervals in the audio (Section 8.3) and obtain the final section links.

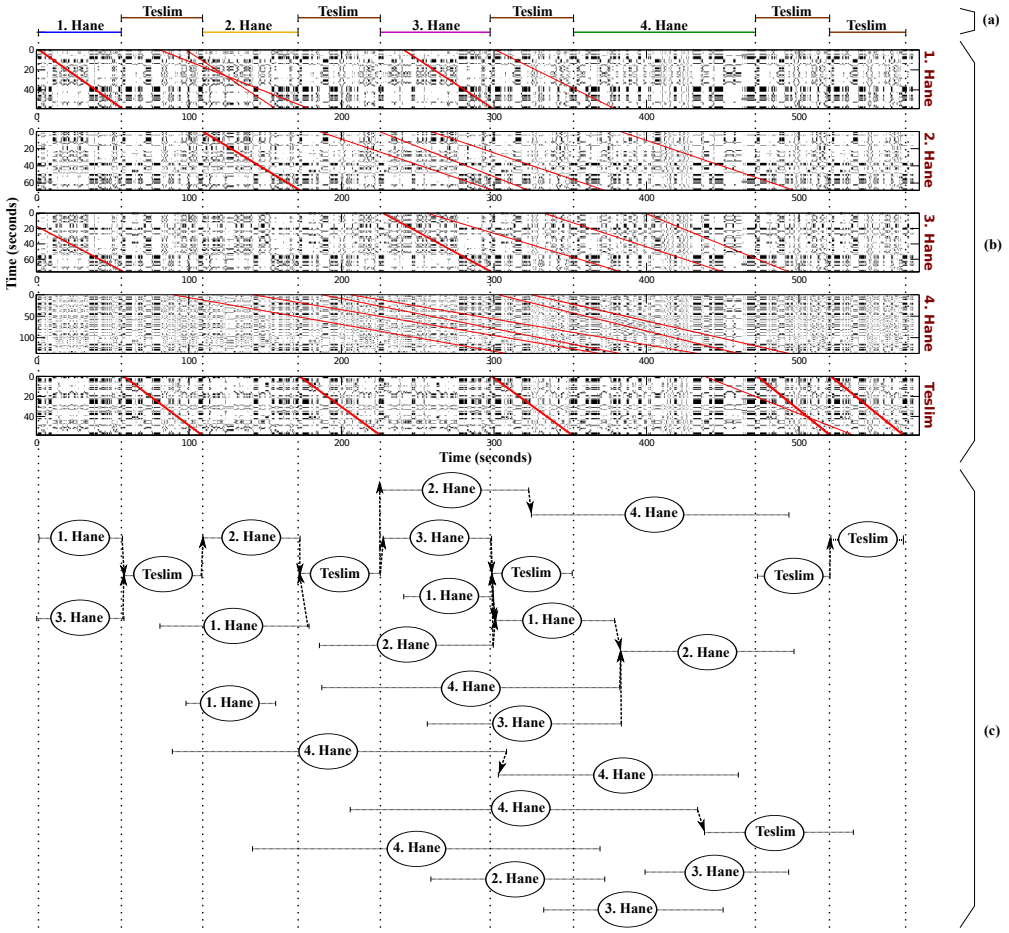

Figure 6: Extraction of all possible paths from the estimated candidates in an audio recording of *Şedaraban Sazsemaisi*. a) Annotated Sections, b) Candidate Estimation, c) The directed acyclic graph formed from the candidate links.

8.1. Path Extraction

Each candidate section link, $\bar{\pi}_k$, may be interpreted as a labeled vertex, which has the following labels:

- Section symbol, $\pi_k \in \mathcal{S}_s$
- Time interval $t(\bar{\pi}_k) = [t_{ini}(\bar{\pi}_k)\ t_{end}(\bar{\pi}_k)]$.
- Weight, $w(\bar{\pi}_k)$, in the interval [0, 1] (see Section 7).
- Relative tempo, $\tau_R(\bar{\pi}_k)$, with its value restricted according to the duration constraint given in Section 7, i.e. to the interval [0.55, 1.5].

If the *final time* of a vertex, $t_{end}(\bar{\pi}_j)$, is close enough to the *initial time* of another vertex, $t_{ini}(\bar{\pi}_k)$, i.e. $|t_{end}(\bar{\pi}_j) - t_{ini}(\bar{\pi}_k)| < \alpha$ (α is chosen as 3 seconds), a directed edge $e_{j\to k} = \langle \bar{\pi}_j,\ \bar{\pi}_k \rangle$ from $\bar{\pi}_j$ to $\bar{\pi}_k$ is formed. The vertices and edges form a directed acyclic graph (DAG), G (Figure 6c).

We define a path p_i as a sequence of vertices $\bar{\boldsymbol{\pi}}_{\boldsymbol{i}} = [\bar{\pi}_{i,1}, \bar{\pi}_{i,2}, \ldots, \bar{\pi}_{i,k}, \ldots, \bar{\pi}_{i,K_i}] \subset \Pi(G)$, where $\Pi(G)$ denotes the vertex set of the graph; and weighted edges $\boldsymbol{e_i} = [e_{i,1}, e_{i,2}, \ldots, e_{i,k}, \ldots, e_{i,K_i-1}] \subset \mathcal{E}(G)$, where $e_{i,k}$ represents the directed edge $e_{i,k\to i,(k+1)} = \left\langle \bar{\pi}_{i,k}, \bar{\pi}_{i,(k+1)} \right\rangle$ and $\mathcal{E}(G)$ denotes the edge set of the graph. The length of the path is $|p_i| = |\boldsymbol{e_i}| = K_i - 1$. We also obtain the section symbol sequence $\boldsymbol{\pi_i} = [\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,k}, \ldots, \pi_{i,K_i}]$, where $k \in [1 : K_i]$ and $\pi_{i,k} \in \mathcal{S}_s$ is the section of the vertex, $\bar{\pi}_{i,k}$.

To track the section sequences in audio with reference to the score section symbol sequence $\boldsymbol{\sigma}$, we construct a variable-length Markov model (VLMM) (Bühlmann and Wyner, 1999). A VLMM is an ensemble of Markov models from an order of 1 to a maximum order of N_{max}. Given a section symbol sequence $\boldsymbol{\pi_i}$, the transition probability $b_{i,k-1}$ of the edge $e_{i,(k-1)}$ is computed as:

$$b_{i,k-1} = P\left(\pi_{i,k} | \pi_{i,(k-1)} \ldots \pi_{i,(k-n)}\right), \quad n = \min\left(N_{max}, k-1\right) \tag{6}$$

In our dataset, the sections are repeated at most twice in succession (Section 4). Hence, the maximum order of the model N_{max} is chosen as 3, which is necessary and sufficient to track the position of the section sequence. VLMMs are trained from the score section symbol sequences, $\boldsymbol{\sigma}$, and audio section symbol sequences, $\boldsymbol{\zeta}$, of other audio recordings whose compositions are built from a common symbol set $\mathcal{S}_s$. If a composition is performed partially in an audio recording, the recording is not used for training.

If a vertex $\bar{\pi}_k$ has outgoing but no incoming edges, it is the starting vertex of a path. A vertex $\bar{\pi}_k$ is connectable to the a path p_i ($|p_i| = K_i - 1$), if the following conditions are satisfied:

i. A directed edge $e_{i,K_i\to k}$ from $\bar{\pi}_{i,K_i}$ to $\bar{\pi}_k$ exists, i.e. $|t_{end}(\bar{\pi}_{i,K_i}) - t_{ini}(\bar{\pi}_k)| < \alpha, \quad \alpha =$ 3 seconds.

ii. The transition probability from $\bar{\pi}_{i,K_i}$ to $\bar{\pi}_k$ is greater than zero, i.e. $P\left(\pi_k|\pi_{i,K_i}\dots\pi_{i,(K_i-n+1)}\right) > 0, \quad n = \min\left(N_{max}, K_i\right)$.

Starting from the vertices with no incoming edges, we iteratively build all paths in the graph by applying the above rules. While traversing the vertices, an additional path is encountered, if:

- A vertex in the path is connectable to more than one vertex. There exists a path for each of these connectable vertices. All these paths share the same starting vertex.
- The transition probability of an edge to the vertex $\bar{\pi}_k$ is zero for the current path p_i, i.e. $|t_{end}(\bar{\pi}_{i,K_i}) - t_{ini}(\bar{\pi}_k)| < \alpha, \quad \alpha = 3$ seconds, and $P\left(\pi_k|\pi_{i,K_i}\dots\pi_{i,(K_i-n+1)}\right) = 0, \quad n = \min\left(N_{max}, K_i\right)$, but the transition probability is greater than zero for a VLMM with order smaller than $0 < n' < n$. In this case, there exists a path that has $\bar{\pi}_{i,(K_i-n'+1)}$ as the starting vertex.

Traversing the vertices and edges, we obtain all possible paths $\mathcal{P}(G) = \{p_1, \dots, p_i, \dots, p_L\}$ from the candidate links, where L is the total number of paths (Figure 7a). The total weight of a path p_i is calculated by adding the weights of the vertices and the transition probabilities of the edges forming the path:

$$w(p_i) = \sum_{k=1}^{K_i} w(\bar{\pi}_{i,k}) + \sum_{k=1}^{K_i-1} b_{i,k} \tag{7}$$

In summary, each path p_i has the following labels:

- A sequence of labeled vertices, $\bar{\boldsymbol{\pi}}_{\boldsymbol{i}} \subset \Pi(G)$, $|\bar{\boldsymbol{\pi}}_{\boldsymbol{i}}| = K_i$.
- Directed, labeled edges connecting the vertices, $\boldsymbol{e_i} \subset \mathcal{E}(G)$, $|\boldsymbol{e_i}| = K_i - 1$.
- Section symbol sequence, $\boldsymbol{\pi_i} = [\pi_{i,1}, \dots, \pi_{i,K_i}]$.
- Time interval $t(p_i) = [t_{ini}(p_i)\ t_{end}(p_i)]$, where $t_{ini}(p_i) = t_{ini}(\bar{\pi}_{i,1})$ denotes the initial time and $t_{end}(p_i) = t_{end}(\bar{\pi}_{i,K_i})$ denotes the final time of the path.
- Total weight, $w(p_i)$.

8.2. Elimination of Improbable Candidates

Correct paths usually have a greater number of vertices (and edges) as depicted in Figure 7a. Moreover, the correct vertices typically have a higher weight than the others. Therefore, the correct paths have a higher total weight than other paths within their duration. Assuming p^* is the path with the highest total weight, we remove all other vertices within the duration of the path $[t_{ini}(p^*)\ t_{end}(p^*)]$ (Algorithm 1, Figure 7b,d). Notice that p^* can remove one or more vertices

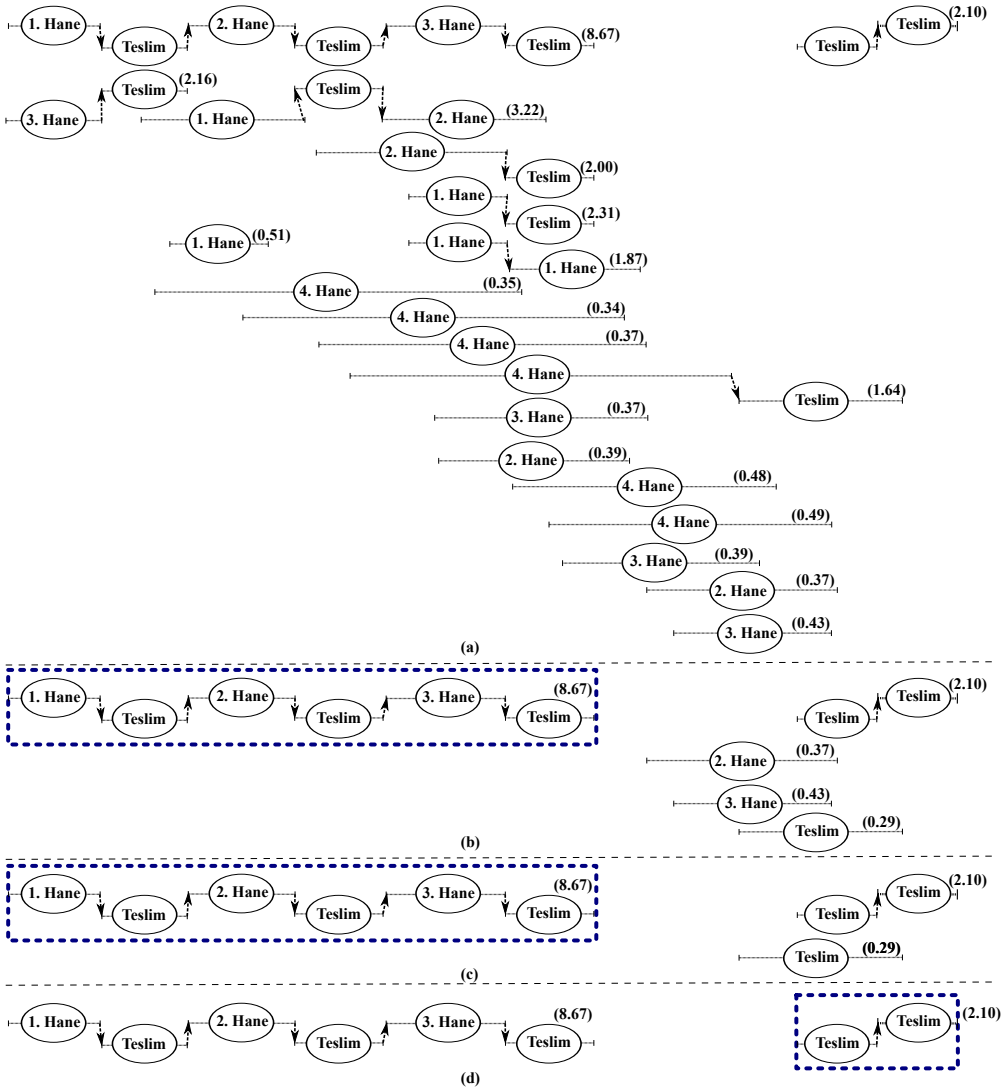

Figure 7: Graphical example for the sequential linking for the *Şedaraban Sazsemaisi*. a) All possible paths extracted from the graph. The number in parenthesis in the right side of each path indicates the total weight of the path. b) Overlapping vertices with respect to the path with the highest weight are removed (see Alg. 1). c) Inconsequent vertices with respect to the path with the highest weight are removed (see Alg. 2). d) Overlapping vertex with respect to the path with the second highest weight is removed.

Algorithm 1 Remove overlapping vertices

function REMOVE_OVERLAP($\Pi(G)$, p^*)
  $\Pi_{chk} \Leftarrow \Pi(G) - \bar{\boldsymbol{\pi}}^*$;
  for $\bar{\pi}_k \in \Pi_{chk}$ **do**
    if $[t_{ini}(p^*)\ t_{end}(p^*)] \cap [t_{ini}(\bar{\pi}_k)\ t_{end}(\bar{\pi}_k)] > 3$ seconds **then**
      $\Pi(G) \Leftarrow \Pi(G) - \bar{\pi}_k$;
  return $\Pi(G)$

from the "middle" of another path, which has a longer time duration than p^*; effectively removing edges, splitting the path into two, and hence creating two separate paths.

After removing the vertices within the time interval covered by the path p^*, the related section sequence $\bar{\boldsymbol{\pi}}^*$ ($|\bar{\boldsymbol{\pi}}^*| = K^*$) becomes unique within this time interval, and are therefore considered final section links. The section symbol sequence of the path $\boldsymbol{\pi}^*$ follows a *score section symbol subsequence* $\boldsymbol{\sigma}^* = [\sigma_j, \ldots, \sigma_k]$ of the score section symbol sequence $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_j, \ldots, \sigma_k, \ldots, \sigma_M]$, $1 \leq j \leq k \leq M$. Next, we remove inconsequent vertices occurring before and after the audio section sequence, p_i with respect to $\boldsymbol{\sigma}^*$ (see Algorithm 2).

We define two score section symbol subsequences $\boldsymbol{\sigma}^-$ and $\boldsymbol{\sigma}^+$, which occur before and after $\boldsymbol{\sigma}^*$, respectively. Since the sections may be repeated twice in succession within a performance (Section 4), they depend on the first two section symbols, $\{\pi_1^*, \pi_2^*\}$, and the last two section symbols, $\{\pi_{K^*-1}^*, \pi_{K^*}^*\}$, of the section symbol sequence $\boldsymbol{\pi}^*$ of the path p^*:

$$\boldsymbol{\sigma}^- = \begin{cases} \emptyset, & \pi_1^* = \pi_2^* = \sigma_1 \\ [\sigma_1, \ldots, \sigma_{j-1}], & \pi_1^* = \pi_2^* \neq \sigma_1 \\ [\sigma_1, \ldots, \sigma_j], & \pi_1^* \neq \pi_2^* \end{cases}, \quad \boldsymbol{\sigma}^+ = \begin{cases} \emptyset, & \pi_{K^*-1}^* = \pi_{K^*}^* = \sigma_M \\ [\sigma_{k+1}, \ldots, \sigma_M], & \pi_{K^*-1}^* = \pi_{K^*}^* \neq \sigma_M \\ [\sigma_k, \ldots, \sigma_M], & \pi_{K^*-1}^* \neq \pi_{K^*}^* \end{cases} \tag{8}$$

Since sections given in the $\boldsymbol{\sigma}^-$ and $\boldsymbol{\sigma}^+$ have to be played in the audio before and after $\boldsymbol{\pi}^*$ respectively, we may remove all the vertices occurring before and after p^*, which do not follow these score section symbol subsequences (Algorithm 2, Figure 7c).

Algorithm 2 Remove inconsequent vertices

function REMOVE_INCONSEQUENT($\Pi(G)$, p^*)
  $\Pi_{chk} \Leftarrow \Pi(G) - \bar{\boldsymbol{\pi}}^*$;
  $\boldsymbol{\sigma}^-, \boldsymbol{\sigma}^+ \Leftarrow$ get_prevNext_sectionSubsequences($\boldsymbol{\pi}^*, \boldsymbol{\sigma}^*$) ▷ Equation 8
  for $\bar{\pi}_k \in \Pi_{chk}$ **do**
    if $t_{ini}(\bar{\pi}_k) < t_{ini}(p^*)$ & $\pi_k \notin \boldsymbol{\sigma}^-$ **then**
      $\Pi(G) \Leftarrow \Pi(G) - \bar{\pi}_k$;
    else if $t_{end}(\bar{\pi}_k) > t_{end}(p^*)$ & $\pi_k \notin \boldsymbol{\sigma}^+$ **then**
      $\Pi(G) \Leftarrow \Pi(G) - \bar{\pi}_k$;
  return $\Pi(G)$

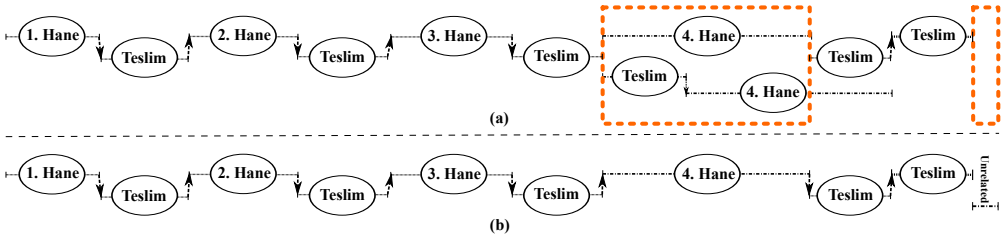

Figure 8: Guessing non-estimated time intervals shown on an audio recording of *Şedaraban Sazsemaisi* a) Possible paths computed with respect to the median of the relative tempos of all vertices. b) Final links

In order to obtain the optimal (estimated) audio section sequence $\bar{\boldsymbol{\pi}}$, we iterate through the paths ordered by weight w_i and remove improbable vertices according to this path by using Algorithms 1 and 2. Note that the final sequence might be fragmented into several disconnected paths, as shown *e.g.* in Figure 7d. The final step of our algorithm attempts to fill these gaps based solely on information about the compositional structure.

8.3. Guessing non-linked time intervals

After we obtained a list of links based on audio and structural information, there might be some time intervals where there are no sections linked (Figure 7d).

Assume that the time interval $t^* = [t^*_{ini}\ t^*_{end}]$ is not linked and it lies between two paths, $\{p^-, p^+\}$, before and after the non-linked interval. Note that the path p^- or p^+ can be empty, if the time interval is in the start or the end of the audio recordings, respectively. These paths would follow the score section symbol subsequences $\boldsymbol{\sigma}^-$ and $\boldsymbol{\sigma}^+$, respectively, and there will be a score section symbol subsequence $\boldsymbol{\sigma}^* = [\sigma^*_1, \ldots, \sigma^*_{M^*}]$, lying between $\boldsymbol{\sigma}^-$ and $\boldsymbol{\sigma}^+$. This score symbol subsequence can be covered in the time interval t^*. Since the sections may be repeated twice in succession within a performance (Section 4), the first and the last symbol of $\boldsymbol{\sigma}^*$ depend on the last two section symbols of $\boldsymbol{\pi}^-$ and the first two section symbols of $\boldsymbol{\pi}^+$ (similar to Equation 8).

From the VLMMs, we compute all possible section symbol sequences, $\{\boldsymbol{\pi}^*_{\mathbf{1}}, \ldots, \boldsymbol{\pi}^*_{\boldsymbol{R}}\}$, that obey the subsequence $\boldsymbol{\sigma}^*$, where R is the total number of computed sequences. From the possible section symbol sequences, we generate each path $\mathcal{P}^* = \{p^*_1, \ldots, p^*_r, \ldots, p^*_R\}$, $r \in [1:R]$. The relative tempo of each vertex in the possible paths is set to the median of the relative tempo of all previously linked vertices, i.e. $\tau_R(\bar{\pi}^*_{r,k}) = \text{median}(\tau_R(\bar{\pi}_k), \forall \bar{\boldsymbol{\pi}}_{\boldsymbol{k}} \in \Pi(G))$, where $\bar{\pi}^*_{r,k} \in \bar{\boldsymbol{\pi}}_{\boldsymbol{r}}$ (Figure 8a). Therefore the duration of the vertices in the path becomes $|t(\bar{\pi}^*_{r,k})| = d(s_n)/\tau_R(\bar{\pi}^*_{r,k}), \forall \bar{\pi}^*_{r,k} \in \bar{\boldsymbol{\pi}}_{\boldsymbol{r}}$ and $\pi^*_{r,k} = s_n$.

We then compare the duration of each path and the interval, $|t^* - t^*(p^*_r)|$. We pick p^*_r, such that $r = \arg\min_r(|t^* - t^*(p^*_r)|)$ with the constraint $|t^* - t^*(p^*_r)| < 3$ seconds. If no path is found, the interval is labeled as "unrelated" to composition, i.e. $\pi_k = u$ (Figure 8b). Finally, all the links $\bar{\pi}_k$ are marked as section links.

9. Experiments

We link the sections given in the score with segments in the audio recordings using the approach described in Sections 6-8. The signal features are either chosen as HPCPs or prominent pitches (Section 6). For comparison, the features are computed with 12, 24, 48 and 120 bins per octave (4.42, 2.21, 1.10, 0.44 Hc resolution). The binarization threshold β range from 0.5 to 9 Hc (*i.e.* a whole tone) for distance matrices calculated from prominent pitches, and from 0.20 to 0.50 for distance matrices computed from HPCPs.

Besides section linking, we extract pitch features from annotated audio segments and link them to the audio recording itself, i.e. "self-linking" the section annotations to the audio. This operation represents an upper limit for the possible results achievable by section linking (Figure 9). For repeated sections, the first annotation of any repeated section in the audio recording is selected for self-linking. Self-linking should ideally be able to link all the sections in the audio recording except the repeated sections with phrase omissions, repetitions and tempo changes.

We compare the initial and final times, t_{ini} and t_{end}, of the section links after sequential linking and self linking with the manually annotated time intervals separately. A link is marked as a true positive, if an annotation in the audio recording and the link has the same section label, and the link is aligned with the annotation, allowing a tolerance of ± 3 seconds. All links that do not satisfy these two conditions are considered as false positives. If a section annotation does not have any links in the vicinity of $\pm$ 3 seconds, it is marked as false negative. If an unrelated link is aligned with an unrelated annotation, allowing a tolerance of ± 3 seconds, the unrelated link is a true negative. From these quantities we compute specificity, recall, precision, and F_1-scores as:

$$P = \frac{t_p}{t_p + f_p}, \quad R = \frac{t_p}{t_p + f_n}, \quad S = \frac{t_n}{t_n + f_p}, \quad F_1 = 2\,\frac{P\,R}{P + R} \tag{9}$$

t_p, t_n, f_n, f_p, P, R, S and F_1 stand for number of true positives, number of true negatives, number of false negatives, number of false positives, precision, recall, specificity and F_1-score, respectively. For all results below, the term "significant" has the following meaning: the claim is statistically significant at the $p = 0.01$ level as determined by a multiple comparison test using the Tukey-Kramer statistic.

9.1. Results

To find the optimal parameters for section linking, the experiments are done over a range of pitch feature precisions and binarization thresholds (β) using annotated karars (Figure 9). The HPCPs with 4.42 Hc pitch precision (12 bins per octave) perform better with a binarization threshold at around 0.3. For pitch precisions higher than 4.42 Hc, the optimal results are obtained for a binarization threshold between 0.30 and 0.45 (Figure 9a). Increasing the precision produces slightly better but insignificant results ($p = 0.85$) for HPCPs. The optimal range of binarization threshold for prominent pitch is observed between 1.5 and 4 Hc (Figure 9b). The F_1-scores are

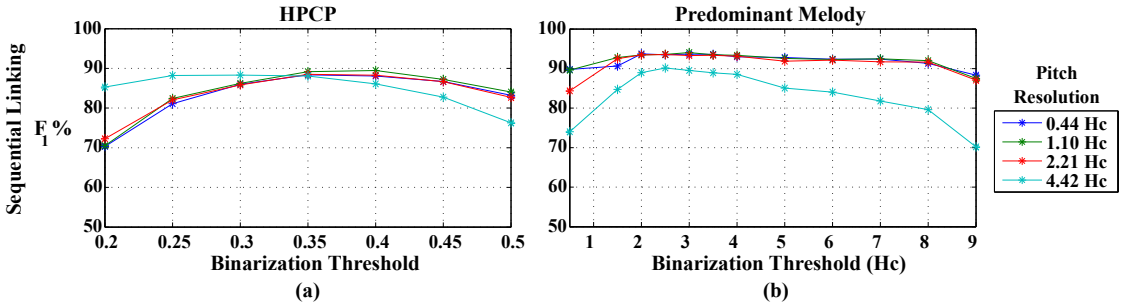

Figure 9: F_1 scores obtained for different pitch feature precision and binarization threshold. **a)** Sequential linking results using HPCPs and annotated karars, **b)** Sequential linking results using prominent pitch and annotated karars

Table 1: Section linking results obtained for candidate estimation and sequential linking with annotated karar, and for self-linking. Optimal parameters are used for computation. The results are given for the instrumental pieces and the vocal pieces separately.

		HPCPs			**Prominent pitch**		
		Cand. Est.	**Seq. Link**	**Self Link**	**Cand. Est.**	**Seq. Link**	**Self Link**
Instr.	P%	28.9	89.8	97.2	32.9	92.6	96.5
	R%	88.5	86.3	96.8	93.7	91.6	96.6
	S%	0	46.8	77.7	0	55.3	73.2
	F_1%	**43.6**	**88.1**	**97.0**	**48.7**	**92.1**	**96.6**
Vocal	P%	43.2	92.3	97.5	54.1	96.9	96.4
	R%	85.0	87.1	97.3	95.6	97.0	96.6
	S%	0	39.5	69.2	0	64.3	58.9
	F_1%	**57.3**	**89.6**	**97.4**	**69.2**	**96.9**	**96.5**

similar for HPCPs (88.3%) and prominent pitch (90.2%) with semi tone (4.42 Hc) pitch precision and optimal binarization thresholds at this precision ($\beta = 0.3$ for HPCPs and $\beta = 2.5$ Hc for prominent pitches). Since increasing the pitch precision more than quarter tone (< 2.21 Hc) does not have any significant effect, we select 2.21 Hc pitch precision as the optimal pitch precision. In the remainder of this section, we are reporting the detailed results using the optimal parameters (2.21 Hc pitch precision for both features; $\beta = 0.35$ for HPCPs and $\beta = 2.5$ Hc for prominent pitches), unless stated otherwise. The optimal parameters will be further discussed in Section 10.

The F_1-score obtained from the entire dataset using the optimal parameters and annotated karar is 88.5% (vs. 97.1% from self-linking) for HPCPs and 93.6% (vs. 96.5% from self-linking) for prominent pitch, with the differences between the features being statistically significant. The average distance between each boundary of a true positive and the corresponding annotation is 0.36 and 0.44 seconds with a standard deviation of 0.41 and 0.49 seconds for links found from HPCPs and prominent pitches using optimal parameters, respectively. There is no significant change in F_1-scores with respect to the instrumentation of the instrumental pieces or voicing of the vocal pieces.

Table 1 gives the results for instrumental pieces and vocal pieces, using the annotated karar

with optimal parameters. Apart from the results for the complete algorithm (Seq. Link), and the self-linking (Self Link), we also report the results obtained from the candidate estimation (Cand. Est.), *i.e.* without applying any candidate selection. While the precision rate is low for candidate estimation, sequential linking greatly increases the precision rate by effectively removing improbable candidates. In the meantime, the recall rate slightly drops for instrumental pieces, and a considerable number of non-linked intervals (44 tp for HPCPs and 20 tp for prominent pitches) are guessed correctly for vocal compositions, effectively increasing the recall rate. The specificity of candidate estimation is always 0, since unrelated time-intervals are marked in sequential linking (Section 8.3). Moreover, 155 (45.2% specificity) and 167 (57.0% specificity) unrelated annotations (out of 220 unrelated annotations) are correctly marked after guessing un-linked time intervals, using HPCPs and prominent pitches, respectively. While there is no significant difference between self-linking results for HPCPs and prominent pitches (97.1% F_1-score for HPCPs and 96.5% F_1-score for prominent pitches using optimal parameters), prominent pitches significantly outperform HPCPs in recall, precision and F_1-scores in both candidate estimation and sequential linking.

Using the prominent pitch with optimal parameters and annotated karar, most of the errors are due to structural changes within sections in a performance (29 out of 129 false positives and 39 out of 146 false negatives in the whole dataset) and substantial tempo changes within the sections (23 out of 129 false positives and 30 out of 146 false negatives in the whole dataset) that Hough transform cannot handle. Most of these errors in the dataset arise from the 4^{th} hane of the instrumental compositions, with 42 out of 129 false positives and 56 out of 146 false negatives in the instrumental compositions being related to the 4^{th} hane. Hough transformation is not able to find the lines associated with 21 (out of 53 annotations) of the 4^{th} hane annotations, due to their tempo deviating more than the allowed ratio with respect to the tempos given in the score. Remaining 4^{th} hane annotations errors are due to omissions, repetitions and tempo changes observed in the performances. One audio recording is too slow [14] and two are too fast[15] such that the relative tempo of the sections are beyond the allowed interval. In these recordings Hough transform fails to detect the appropriate lines (2 true positives out of 24 section annotations, 14 false positives and 20 false negatives).

The results of section linking using automatic karar identification using optimal parameters are reported in Table 2. Compared to section linking using annotated karar (Table 1), there is a considerable drop in all of the recall, precision, and hence F_1-scores. This decrease is due to karar identification, which fails (i.e. the octave-wrapped distance between the annotated karar and identified karar are more than 2.5 Hc, i.e. the optimal binarization threshold) for 62 pieces (53 instrumental pieces and 9 şarkıs) out of 257 recordings (24.1% error). The karar identification typically fails in pieces with more complex makams such as *Ferahfeza* (10 out of 10 recordings),

[14] MBID: 812828e6-3cb6-49c5-93fe-bf649c3096ae
[15] MBIDs: 031a6e72-903a-479a-9a4c-e2a3335e4a0a, 35d127d1-39e1-49d9-ab81-f8180b32590c

Table 2: Results obtained from automatic karar identification using optimal parameters for the instrumental pieces and the vocal pieces. The results are given for the instrumental pieces and the vocal pieces separately.

		HPCPs		**Prominent pitch**	
		Cand. Est.	**Seq. Link**	**Cand. Est.**	**Seq. Link**
Instr.	P%	21.5	76.3	23.3	77.9
	R%	68.1	65.9	72.3	69.8
	S%	0	25.0	0	27.0
	F_1%	**32.7**	**70.7**	**35.2**	**73.6**
Vocal	P%	35.3	84.5	45.4	87.9
	R%	74.2	75.4	85.6	84.4
	S%	0	23.0	0	31.4
	F_1%	**47.9**	**79.7**	**59.3**	**86.1**

Hicazkar (8 out of 9 recordings), *Kürdilihicazkar* (9 out of 19 recordings), *Hüzzam* (5 out of 8 recordings), *Segah* (7 out of 14 recordings) and *Acemaşiran* (3 out of 7 recordings). For those makams, often more emphasis is put on notes different from the karar, which usually leads to the assignment of the karar to one of these notes. The F_1-score obtained from the entire dataset using the optimal parameters and automatic karar identification is 73.5% for HPCPs and 77.5% for prominent pitch.

We also computed the elapsed time for section linking excluding feature computation and karar identification. On average, our implementation in MATLAB takes 3% of the duration of the audio recording (with a standard deviation of 1%) to link the sections of the particular audio recording with a 64 bit Ubuntu machine with 13.5 GB RAM and 3.33 GHz processor.

10. Discussion

The results show that our method is effective in linking sections given in the score to their corresponding time intervals in audio recordings, given a wide variety of instrumental and vocal timbres. The method is able to achieve good results by using different pitch features with fast computation time and accurate link boundaries.

To find optimal parameters for binarization threshold and pitch precision, we examined the results of section linking using different binarization threshold and pitch precision (Section 9.1). The optimal range for binarization threshold for prominent pitch is between 1.5 and 4 Hc. Other than for HPCP, the threshold range of prominent pitch has a musical interpretation since the performed notes might deviate from the theoretical frequency of a note by as much as a semi-tone, as explained in Section 3. This makes prominent pitches more intuitive to apply to MMT, and possibly to other musics with a clear emphasis on melody.

For both HPCPs and prominent pitches, semitone pitch precision (4.42 Hc) performs worse than higher pitch precisions. This shows that pitch precisions higher than semitone are necessary to capture the melodic characteristics of makam music of Turkey. Nevertheless, increasing the pitch

precision more than quarter tone (< 2.21 Hc) does not lead to further increase in the f1-score. Therefore, both precision and binarization threshold lie in the vicinity of 2 Hc. While deviations between theory and practice were observed in single cases to reach a semi-tone, the usual deviation can be assumed to lie close to that value of 2 Hc. This proves the necessity of resolutions higher than the semi-tone resolution when attempting even such a high level task as we do in this paper. Even though increasing the pitch precision beyond 2.21 Hc (24 bins per octave) does not change the F_1-score practically, the default pitch precision can be increased further to use the same pitch tracks for more precision-demanding tasks such as karar identification, audio-score alignment or intonation analysis.

When comparing the results obtained from self-linking and sequential linking using prominent pitch and annotated karar (Table 1), it is evident that the method is able to achieve practically the maximum possible F_1-score for vocal pieces (96.9% vs. 96.5% from self linking[16]) and a very high F_1-score for instrumental pieces (92.1% vs. 96.6% from self linking). The drop in the F_1-score for the instrumental pieces is mostly due to the errors related to specific performance characteristics (internal repetitions, omissions and tempo changes) which Hough transformation cannot handle effectively. Moreover, most of these errors are related to the 4^{th} hanes. In fact, resolving all the errors related to 4^{th} hanes would increase the F_1-score to approximately 95.7% (vs. 96.6% from self linking). In the saz semaisi form, dividing the 4^{th} hane further into its substructures (Section 3) might help to handle these problems. More generally, such performance features could be better handled by aligning audio and the score at the note level.

Statistical significance tests show that our feature and similarity matrix computation is resilient to changes in timbre and density of heterophony. Moreover, recall rates obtained in candidate estimation step (Table 1) show that Hough transform is able to give reliable estimations for section links. It only fails in three audio recordings (out of 257) in which the performance is beyond the allowed tempo ratios. To remove the angle constraints from the line detection step, we need to estimate an average tempo of performance. Increasing the range of searched angles in Hough transform and then deducing the average tempo ratio of the performance from candidates with high weights might be sufficient for tempo estimation.

Comparing the recall rates of candidate estimation and sequential linking in Table 1, it can be seen that guessing non-linked intervals improves the section linking in vocal pieces. However it does not make any improvement in instrumental pieces, mostly due to the non-linearities in the performances of 4^{th} hanes. Guessing non-linked intervals is dependent on the median of the relative tempo of the performed sections (see Section 8.3). In the case, where the 4^{th} hane does not follow the tempo indicated in the score or there are structural deviations inside the performance of the section, the duration of the guessed paths do not match the duration of the non-linked audio

[16]In both cases the number of true positives are the same (616 true positives), however section linking using annotated karar produces slightly less errors than self linking (20 vs. 23 false positives and 19 vs. 22 false negatives). This difference is insignificant.

segment.

The self-linking results (Table 1) imply that both chroma features and prominent pitch can ideally perform equally well. However, candidate estimation using HPCPs misses more true positives than prominent pitch, and sequential linking is not able to reduce the gap between the F_1-scores (Tables 1 and 2). This indicates that the prominent pitch is a more adequate representation, when aiming at a comparison between score and audio in this musical context.

Our method can not search "unrelated" annotations directly and the errors within the time interval can only be removed by Algorithm 2, for false positives that do not obey the section sequences. Moreover the unrelated region can not be marked correctly if the time interval of an unrelated region is estimated poorly due to a neighboring section with tempo changes, phrase repetition/omissions. Detecting "non-musical" events such as applause and silence can help to distinguish the unrelated regions and eliminate errors due to tempo changes typically occurring in the end of the recordings.

The bottleneck of the system is the automatic karar identification. If the karar of the piece is recognized incorrectly, no true lines will be present in the binarized similarity matrices obtained from either of the two feature types. While the results with automatic karar identification using Makam Toolbox are still good, the errors becomes a noticeable drawback especially for pieces composed in complex makams. Nevertheless, by using the melodic information in the scores we can greatly increase the accuracy of the karar identification accuracy. Recently, in Şentürk et al. (2013), we extracted the stable pitches from the audio recording (i.e. the peaks of the pitch distribution computed from the prominent melody) and attempted to link the repetitive section in the score using the candidate estimation method explained in Section 7, assuming each stable pitch as the karar. Using the same data collection explained in Section 4 we achieved an accuracy of 99.6% (1 fail out of 257) effectively solving the karar identification problem for pieces with an available score. The F_1-scores obtained from section linking using repetitive section linking for karar identification is 88.6% (vs. 88.5% using annotated karar) for HPCPs and 93.2%[17] (vs. 93.6% using annotated karar) for prominent pitch for the whole dataset.

11. Conclusion

We presented a novel methodology to link musically relevant sections in a score with corresponding time intervals in an audio recording of the same piece. We tested our approach using HPCPs, a pitch feature previously applied to music with strong emphasis on harmony, and prominent

[17] Using prominent pitch with optimal parameters, all of the sections are correctly linked (i.e. 100% recall) in the audio recording with failed karar identification (MBID: f5a89c06-d9bc-4425-a8e6-0f44f7c108ef). In this recording the distance between the estimated and actual karar is slightly higher than 2.5 Hc, i.e. the optimal binarization threshold. Our section linking methodology is still able to link the sections even though the resultant weights are low.

pitches, a melodic pitch feature. We demonstrated that prominent pitches capture the heterophonic characteristics of MMT better than HPCPs. Since scales in MMT need resolutions higher than a semitone, we also tried section-linking over a range of pitch precisions and binarization thresholds. It was observed that the pitch precision has to be higher than semitone to represent the melodic granularity of MMT. Unlike HPCPs, the optimal range of binarization threshold for prominent pitches was musically interpretable. Therefore using prominent pitches is more intuitive for music, where there is a clear melody, and concepts like functional harmony do not exist. Our results show the importance of culture-aware and knowledge-based systems. Nevertheless, we have also achieved remarkable results using HPCPs. It may be argued that that the methodology can be easily adapted to Eurogenetic musics, which can be typically conceptualized with the help of harmony.

Our approach is fast and accurate in matching both the section labels and their corresponding time intervals. Section links can be used as a complementary information in computational tasks such as form analysis and audio score alignment. Moreover, the computational steps in our approach can be modified to be used in similar research problems such as pattern matching and version detection. In Şentürk et al. (2013), we used the candidate estimation methodology (Section 7) to identify the performed karar of the audio recording. Our results indicate that score information greatly simplifies the karar identification task. Parallel to findings of Aucouturier and Sandler (2002), the self-linking results (Figure 9c,f) imply that repetitive section linking can be effectively used for audio thumbnailing.

Nevertheless, there is still room for improvement for a more reliable automatic system. As a next step in this research we want to increase the granularity of the linking between audio and score in order to provide more insights on the dynamics and the intonation of makam music performances. This should solve section linking problems due to performance particularities such as omissions, repetitions and tempo changes. We plan to use JumpDTW proposed by Fremerey et al. (2010), which allows jumps between the measures indicated in the score. Since we know the score section sequence, we can further modify JumpDTW to allow jumps between the sections and between the measures within each section. Section linking prior to audio-score alignment might increase the f1-score and reduce computational time.

Content-based creation, analysis and discovery of multimodal and inter-linked music collections is becoming an active research area in the last few years, thanks to the advances in information technology and the emergence of vast numbers of available multi-modal information sources such as audio, MIDI, sheet music, video and editorial metadata (Cornelis et al., 2010; Thomas et al., 2012). Under the CompMusic Project we are developing *Dunya*, a system to browse and interact with music collections in a culturally informed way (Sordo et al., 2012; Porter et al., 2013). We will integrate our *section linking* methodology to the system and use it as a culture-specific tool for navigation and discovery of *makam music of Turkey*. We hope that our approach will contribute to the information technologies aimed at preserving, discovering and appreciating musical cultures.

12. Acknowledgements

We are thankful for Ajay Srinivasamurthy, Dr. Mohamed Sordo and Dr. Barış Bozkurt for their valuable suggestions and comments about the experiments and preparation of the the paper. This work is partly supported by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583).

References

Arzt, A., Böck, S., and Widmer, G. (2012). Fast identification of piece and score position via symbolic fingerprinting. In *Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 433–438.

Aucouturier, J.-J. and Sandler, M. (2002). Finding repeating patterns in acoustic musical signals: Applications for audio thumbnailing. In *Proceedings of 22nd International Audio Engineering Society Conference: Virtual, Synthetic, and Entertainment Audio*.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In *Proceedings of 14th International Society for Music Information Retrieval Conference (ISMIR)*.

Bozkurt, B., Yarman, O., Karaosmanoğlu, M. K., and Akkoç, C. (2009). Weighing diverse theoretical models on Turkish maqam music against pitch measurements: A comparison of peaks automatically derived from frequency histograms with proposed scale tones. *Journal of New Music Research*, 38(1):45–70.

Bühlmann, P. and Wyner, A. J. (1999). Variable length Markov chains. *The Annals of Statistics*, 27(2):480–513.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696.

Cont, A. (2010). A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):974–987.

Cooke, P. (accessed April 5, 2013). Heterophony. Grove Music Online. `http://www.oxfordmusiconline.com/subscriber/article/grove/music/12945`.

Cooper, M. and Foote, J. (2002). Automatic music summarization via similarity analysis. In *Proceedings of 3rd International Society for Music Information Retrieval Conference (ISMIR)*, pages 81–85.

Cornelis, O., Lesaffre, M., Moelants, D., and Leman, M. (2010). Access to ethnic music: Advances and perspectives in content-based music information retrieval. *Signal Processing*, 90(4):1008–1031.

De Cheveigné, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of Acoustical Society of America*, 111(4):1917–1930.

Devaney, J., Mandel, M., and Fujinaga, I. (2012). A study of intonation in three-part singing using the automatic music performance analysis and comparison toolkit (AMPACT). In *Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 511–516.

Duda, R. O. and Hart, P. E. (1972). Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15.

Ederer, E. B. (2011). *The Theory and Praxis of Makam in Classical Turkish Music 1910-2010*. PhD thesis, University of California, Santa Barbara.

Ellis, D. P. and Poliner, G. E. (2007). Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 1429–1432.

Ewert, S. and Müller, M. (2012). Score-informed source separation for music signals. In Müller, M., Goto, M., and Schedl, M., editors, *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 73–94. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany.

Fremerey, C., Müller, M., and Clausen, M. (2010). Handling repeats and jumps in score-performance synchronization. In *Proceedings of 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 243–248.

Fujihara, H. and Goto, M. (2012). Lyrics-to-audio alignment and its application. In Müller, M., Goto, M., and Schedl, M., editors, *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 23–36. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany.

Gedik, A. C. and Bozkurt, B. (2010). Pitch-frequency histogram-based music information retrieval for Turkish music. *Signal Processing*, 90(4):1049–1063.

Gómez, E. (2006). *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra.

Goto, M. (2003). A chorus-section detecting method for musical audio signals. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 437–40.

Holzapfel, A. (2010). *Similarity methods for computational ethnomusicology*. PhD thesis, University of Crete.

Holzapfel, A. and Stylianou, Y. (2009). Rhythmic similarity in traditional Turkish music. In *Proceedings of 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 99–104.

Karadeniz, M. E. (1984). *Türk Musıkisinin Nazariye ve Esasları*, page 159. İş Bankası Yayınları (in Turkish).

Karaosmanoğlu, K. (2012). A Turkish makam music symbolic database for music information retrieval: SymbTr. In *Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 223–228.

Krause, E. F. (1987). *Taxicab geometry: An adventure in non-Euclidean geometry*. Dover Publications.

Martin, B., Robine, M., and Hanna, P. (2009). Musical structure retrieval by aligning self-similarity matrices. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 483–488.

Müller, M. (2007). *Information retrieval for music and motion*, volume 6. Springer Heidelberg.

Müller, M. and Ewert, S. (2008). Joint structure analysis with applications to music annotation and synchronization. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 389–394, Philadelphia, Pennsylvania, USA.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

Niedermayer, B. (2012). *Accurate Audio-to-Score Alignment - Data Acquisition in the Context of Computational Musicology*. PhD thesis, Johannes Kepler Universität, Linz.

Özkan, I. H. (2006). *Türk mûsikısi nazariyatı ve usûlleri: Kudüm velveleleri*. Ötüken Neşriyat (in Turkish).

Paulus, J. and Klapuri, A. (2009). Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1159–1170.

Paulus, J., Müller, M., and Klapuri, A. (2010). State of the art report: Audio-based music structure analysis. In *Proceedings of 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 625–636.

Pikrakis, A., Theodoridis, S., and Kamarotos, D. (2003). Recognition of isolated musical patterns using context dependent dynamic time warping. *IEEE Transactions on Speech and Audio Processing*, 11(3):175–183.

Popescu-Judetz, E. (1996). *Meanings in Turkish Musical Culture*. Pan Yayıncılık, Istanbul.

Porter, A., Sordo, M., and Serra, X. (2013). Dunya: A system for browsing audio music collections exploiting cultural context. In *Proceedings of 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil.

Salamon, J. and Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770.

Şentürk, S. (2011). Computational modeling of improvisation in Turkish folk music using variable-length Markov models. Master's thesis, Georgia Institute of Technology.

Şentürk, S., Gulati, S., and Serra, X. (2013). Score informed tonic identification for makam music of Turkey. In *Proceedings of 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 175–180, Curitiba, Brazil.

Şentürk, S., Holzapfel, A., and Serra, X. (2012). An approach for linking score and audio recordings in makam music of Turkey. In *2nd CompMusic Workshop*, pages 95–106, Istanbul, Turkey.

Serrà, J., Serra, X., and Andrzejak, R. G. (2009). Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9).

Serra, X. (2011). A multicultural approach in music information research. In *Proceedings of 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 151–156.

Signell, K. L. (1986). *Makam: Modal practice in Turkish art music*. Da Capo Press.

Sordo, M., Koduri, G. K., Şentürk, S., Gulati, S., and Serra, X. (2012). A musically aware system for browsing and interacting with audio music collections. In *Proceedings of 2nd CompMusic Workshop*, pages 20–24, Istanbul, Turkey. Universitat Pompeu Fabra.

Thomas, V., Fremerey, C., Müller, M., and Clausen, M. (2012). Linking sheet music and audio - Challenges and new approaches. In Müller, M., Goto, M., and Schedl, M., editors, *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 1–22. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany.

Townsend, M. and Sandler, M. B. (1993). Pattern recognition for formant trajectories using the Hough transform. In *14. Colloque sur le traitement du signal et des images, FRA, 1993*, pages 1355–1358. Groupe d'Etudes du Traitement du Signal et des Images (GRETSI).

Tura, Y. (1988). *Türk Musıkisinin Meseleleri*. Pan Yayıncılık, Istanbul (in Turkish).

Tzanetakis, G., Kapur, A., Schloss, W. A., and Wright, M. (2007). Computational ethnomusicology. *Journal of interdisciplinary music studies*, 1(2):1–24.

Wang, A. L.-C. (2003). An industrial strength audio search algorithm. In *Proceedings of 4th International Society for Music Information Retrieval Conference (ISMIR)*, pages 713–718.