AN APPROACH FOR LINKING SCORE AND AUDIO RECORDINGS IN MAKAM MUSIC OF TURKEY

Sertan Şentürk[1], André Holzapfel[1,2], Xavier Serra[1]
[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain.
[2] Bahçeşehir University, Istanbul, Turkey.
sertan.senturk@upf.edu, andre.holzapfel@upf.edu, xavier.serra@upf.edu

ABSTRACT

The main information sources to study a particular piece of music are symbolic scores and audio recordings. These are complementary representations of the piece and it is very useful to have a proper linking between the two of the musically meaningful events. For the case of *makam* music of Turkey, linking the available scores with the corresponding audio recordings requires taking the specificities of this music into account, such as the particular tunings, the extensive usage of non-notated expressive elements, and the way in which the performer repeats fragments of the score. Moreover, for most of the pieces of the classical repertoire, there is no score written by the original composer. In this paper, we propose a methodology to pair sections of a score to the corresponding fragments of audio recording performances. The pitch information obtained from both sources is used as the common representation to be paired. From an audio recording, fundamental frequency estimation and tuning analysis is done to compute a pitch contour. From the corresponding score, symbolic note names and durations are converted to a synthetic pitch contour. Then, a linking operation is performed between these pitch contours in order to find the best correspondences. The method is tested on a dataset of 11 compositions spanning 44 audio recordings, which are mostly monophonic. An F_3-score of 82% and 89% are obtained with automatic and semi-automatic *karar* detection respectively, showing that the methodology may give us a needed tool for further computational tasks such as form analysis, audio-score alignment and *makam* recognition.

1. INTRODUCTION

In analyzing a music piece, the score, when available, is a highly valuable source to study since it provides an easily accessible symbolic description of many relevant musical components. The audio recordings of a performance of the same piece are another powerful source of information, since they can provide information about the characteristics of the interpretation *e.g.* in terms of dynamics or timing. If these information sources can be connected together by time-aligning fragments from each source (or in other words linking the score excerpts with the corresponding regions in the audio recordings), we can take profit of their complementary aspects to study the music piece. Parallel information extracted from the scores and audio recordings will facilitate computational tasks such as version detection, *makam* recognition [1], tuning analysis [2], intonation analysis, form analysis, melodic modeling [3], musical similarity [4], and expressive performance modeling. Furthermore, in previous work [3], it was discussed that parallel to information retrieval from scores, audio analysis is integral to study the unique characteristics of *makam* music of Turkey.

The current state of the art in music information retrieval involving scores and audio recordings is mainly aimed at Western musics (Section 2.3). In these cases, typically the scores and audio are both polyphonic. However, all *makam* music scores are monophonic and the performances done from them (esp. ensemble performances) are typically heterophonic (Section 2.1). Thus, the methodologies used for Western musics cannot be applied to *makam* music of Turkey and we have to develop approaches aware of the properties of *makam* music (Section 3.2).

To match fragments of symbolic data to fragments of audio recordings, we can link the melodic score excerpts (or motifs) with the pitch information obtained from audio recording or match metric templates of the scores with the onset values extracted from audio recordings. In this paper, we focus on linking score sections with the corresponding fragments in the audio recordings, i.e. finding the time interval in the audio recording of a piece, where a particular section indicated in the score of the same piece is performed. From this linking, computational operations such as *makam* recognition, *usul* detection or audio-score alignment can be done at the section level, providing a deeper insight on structural, melodic or metric properties of the music.

The remainder of the paper is structured as follows: Section 2 gives a brief introduction to *makam* music of Turkey, properties of *makam* music notation and related computational research. Section 3 explains the proposed methodology. Section 4 presents the experiments carried to evaluate the method and the results obtained from the experiments. Section 5 gives a discussion on the results, and Section 6 ends the paper with a brief conclusion.

2. BACKGROUND

2.1 Makam Music of Turkey

The melodic structure of most classical and folk repertoires of Turkey is explainable by makams. Makams are modal structures, where the melodies typically revolve around a *başlangıç* (starting, initial) tone and a karar (ending, final) tone [5]. The octave is divided into at least 17 intervals [5], the intervals are not equally tampered, and there is no single fixed tuning. There are a number of different tunings (*ahenk*) any of which might be favored over others due to instrument and/or vocal range or aesthetic concerns [5].

The metric structure of *makam* music is explained by *usul*. The term *usul* can be roughly translated to cyclic meter. Nevertheless, *usul* is a wider concept, which is not limited to metric implications, since a change in *usul* can disrupt the melodic progressions (*seyir*) and even change the perception of the *makam* [6].

For centuries, *makam* music has predominantly been an oral tradition. In the early 20th century, a score representation based on extending Western music notation started to be used, and it has become a fundamental complement to the oral tradition [7]. The music written in scores are typically monophonic; nevertheless performances (esp. ensemble performances) involve various heterophonic peculiarities.

Currently Arel-Ezgi-Uzdilek theory is the mainstream theory used to explain *makam* music [8]. Arel-Ezgi-Uzdilek theory argues that there are 24 intervals in an octave, a subset of the steps obtained by dividing each tempered whole tone into 9 equidistant intervals [8] [1] . The extended Western notation typically follows the constraints of Arel-Ezgi-Uzdilek theory. Nevertheless, the theory is controversial due to some critical differences with the practice [5, 6].

In the experiments (Section 4), we focused on the two most common instrumental forms in classical *makam* tradition, namely the *saz semaisi* and *peşrev* forms. These two forms commonly consists of four distinct *hane*s and a *teslim* section between the *hane*s. These sections can be roughly considered as analogous to *verse* and *chorus*. Nevertheless, there are *peşrev*s, which has no *teslim*, yet the second half of each *hane* strongly resembles each other [9]. The fourth *hane* is typically longer and have a change in the *makam* and *usul*. Also, the last measures of each *teslim* may differ with respect to the *hane* it is being connected.

2.2 Presciptive vs. Descriptive Notation

The intent of musical notations can be either (1) prescriptive notations, used as a means to explain the performers how to perform a musical piece, or (2) descriptive notations, which narrate how the music is performed by musicians [10]. In this context, the majority of compositions in Western classical music would use prescriptive notations and the transcriptions done from a performance would be considered descriptive.

The available *makam* music scores are guidelines for the performers [11], even tough a considerable number of compositions (esp. the ones composed before 20^{th} century) are actually transcriptions of performances. The performers not only deviate considerably from the score, but they normally play differently every time; showing their musicality and virtuosity by using expressive timings, adding note repetitions and non-notated embellishments. Moreover, the intonation of some intervals might change, or even a neighboring tone might be played instead of the one written in the score [12]. As a last remark, the complex heterophonic interactions in the ensemble performances are not indicated in the scores. Therefore, the scores of *makam* music can be considered both prescriptive and descriptive.

2.3 Related Computational Research

There is very little work done on the automatic segmentation of *makam* musics. The only published experiment was conducted by Lartillot and Ayari [13]. They used computational models with low-level and high-level heuristics to make structural segmentations of modal ney improvisations in Tunisian maqam music. They compare their automatic results with segmentations performed by human subjects with different cultural and musical backgrounds.

The current state-of-the-art systems on section analysis are mostly aimed at dividing audio recordings of Western popular music into repeated and mutually exclusive sections. For these segmentations, typically self-similarity analysis [14, 15] is employed [2] , in which a similarity matrix is computed by taking the distance of temporal features obtained from the audio recording by itself. Since the resultant matrix is square, the repetitions may only occur in the direction of the diagonal ($\pm$45 degrees, depending on the orientation). This directional constraint makes it possible to identify repetitions, 2D sub-patterns inside the matrix. However, as explained in Section 2.1, there are some special cases in *makam* music, where there are no repeated sections. In such cases self-similarity may not only be useless but it may also give false results.

Due to inherent characteristics of the oral tradition and the practice of *makam* music of Turkey, performances of the same piece may be substantially different from each other. A similar situation occurs in cover song identification [17, 18] for which a similarity matrix is computed from the temporal descriptors obtained from a cover song candidate and the original recording. If the similarity matrix is found to have some strong regularities (i.e. several prominent paths with minimal costs), they are deemed as two different versions of the same piece of music. In this case, the similarity matrix is non-square unless the audio recordings have exactly the same duration. A proposed solution is to "squarize" the similarity matrix is by computing some hypothesis about the tempo difference [17]. However, *usul* analysis in *makam* musics is not a straightforward task [19]. The sections may also be found by traversing the similarity matrices using dynamic programming [18]. On the other hand, dynamic programming is a computationally demanding task, and the approach may only link a single section at a time, i.e. the algorithm needs

[1] An interval equal to $1/9^{th}$ of a whole tone is also termed as *Holderian Comma* (Hc) and they divide an octave into 53 equal intervals.

[2] For an overview of section analysis (and structural analysis in general) tasks and relevant approaches, the readers can refer to [16].

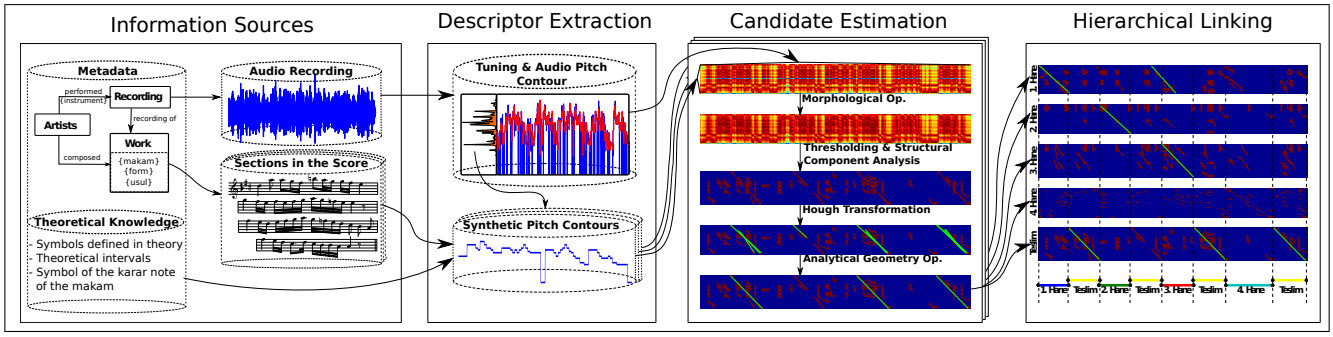

Figure 1: Block diagram of the section linking methodology between a score of a piece and an audio recording of the same piece.

to run multiple times to locate any repeated sections in an audio recording.

When the score is available, incorporating information extracted from it might be more insightful for structural analysis than solely relying on the audio recordings. Martin et al. have proposed a methodology to structurally align symbolic structural queries and audio recordings by making 2D comparisons of self-similarity matrices calculated from the symbolic queries and self-similarity matrices calculated from the recordings [20]. However, the method heavily relies on the timings of the annotated queries and it is not impervious to changes at the excerpt boundaries. Nevertheless, the system is better aimed at retrieving a previously annotated audio recording inside a large audio database than at locating the sections in an arbitrary audio recording of a music piece.

Structure analysis is related to some research in image processing, since the similarity matrices computed may be interpreted as topology maps, and the problem may be regarded as finding regularities inside these maps. To find these regularities, structure analysis may utilize image processing solutions such as morphological operations [21], Hough transform [22] or geodesics [23].

3. METHODOLOGY

Here, we explain the proposed methodology for linking selected score sections of a music composition with the corresponding audio recordings of performances. The method uses a machine readable version of the score of a composition and an audio recording as the inputs along with some complementary metadata about these information sources and some concepts from *makam* music theory (Section 3.1). From the audio recording, the fundamental frequency, f0, is estimated and processed to obtain an audio pitch contour. The f0 estimation is also used to calculate a pitch histogram in order to identify the tuning and the note intervals (Section 3.2.1). From the score information, we read the note symbols, the sections and the *makam* of the piece, and generate a synthetic pitch contour (Section 3.2.2). In order to estimate the candidate locations of the sections in the audio, the method compares these relevant pitch representations (Section 3.3). In the final step, the candidates are hierarchically checked to link the sections of the score to the corresponding parts in the audio (Section 3.4). The block diagram of the methodology is given in Figure 1.

3.1 Information Sources

To link the identified score sections with their performances we use machine-readable scores and audio recordings. These information sources are already associated with each other through complementary metadata available, so that there is no need to apply version detection prior to section linking operations. The scores are encoded as symbTr files [24], a Humdrum-like machine readable format. The starting and ending of the sections are explicitly marked in the scores. We also use some theoretical knowledge, namely the letter symbols of the notes, the letter symbol of the *karar* note of the *makam* of the piece and melodic intervals, to process the audio recordings and the symbolic scores, which will be explained in Sections 3.2.1 and 3.2.2.

3.2 Descriptor Extraction

Since the scores are not strictly followed by a performer, conversion to a more flexible representation is needed. The data should make it possible to make one-to-one mappings in subsequences where both sources could fit relatively well into each other; however they should also provide a level of fuzziness to avoid confusions in substantially dissimilar regions. To achieve a robustness in linking the score and the audio, we use post-processed pitch tracks extracted from the audio recordings and score, which we name "pitch contours."

3.2.1 Pitch Tracking and Tuning Analysis on the Audio

To obtain the audio pitch contour, f0 estimations from the audio recordings are extracted using the *Makam* Toolbox [25]. *Makam* Toolbox uses YIN [26], which has been shown to be highly reliable to estimate the fundamental frequency over time in monophonic music [3]. The hop size is 10ms for pitch tracks. *Makam* Toolbox also post-processes the YIN output to fix the octave errors. Additionally, it has an additional option to quantize the pitch tracks into stable notes. The advantage of the quantized f0 estimation is that

[3] *Makam* Toolbox can also process f0 estimations from other pitch tracking algorithms. However we started the initial experiments with monophonic ney recordings (Section 4.1) and empirically observed reliable estimations. Adaptation of other melodic descriptors is discussed in Section 5.

it takes out minor pitch variations such as vibratos. Afterwards, we further apply a median filter with a window length of 41 frames (410ms) to fix short drops in the f0 estimation.

Together with to the pitch contour calculation, an histogram analysis is done on the raw f0 estimations using the *Makam* Toolbox to identify the *karar* tone and the intervals played [1]. The bin width of the histogram is taken as $1/3$ Holderian commas (Hc)[4]. The intervals played in the performance are obtained by picking the peak values in the histogram. To neutralize the differences in pitch height due to different *ahenk*s, the values of pitch contours are converted to Hc and normalized by subtracting the Hc value of the *karar* tone from each. In other words, the pitch contour shows the floating scale degree of the progression in the audio in Hc, where the *karar* note is assigned 0 Hc. Then, all pitch values are folded to the pitch range given in the score with a tolerance of 14 Hc (approx. 1.5 semitones below and above the theoretical frequencies of the lowest and highest pitched notes given in the score). This threshold allows some space for embellishments in the highest and lowest registers.

In order to find the rests in the audio recording, the audio file is divided into 50% overlapping frames with 100ms length. The average power in each frame is calculated, and normalized with respect to the overall average power of the audio recording. A dynamic threshold is computed by applying a median filter with a length of 100 frames (10 seconds) to the logarithm of the average power values per frame. The silent regions in the audio are detected by picking the frames which have a lower average power than the dynamic threshold. Then, a *pseudo*-value is assigned as the pitch value of the rest (34 Hc below the lowest register) to avoid immense penalties in case a rest is not present in the score and vice versa. These the pitch contour is downsampled by 10 (i.e. 10 samples per second) to emphasize the structural changes and for computational concerns. The peaks detected in the histogram and the rest value are also noted to be used later in the synthetic pitch contour generation (Section 3.2.2).

3.2.2 Score Data Extraction and Synthetic Pitch Contour Generation

From the score, we read the *makam* of the piece, the starting event numbers of the sections, the note names and their durations. If the *teslim*s have different endings, only the note sequence of the first *teslim* is considered. The symbolic format is mapped to theoretical pitches with respect to the theoretical information given, such that the *karar* note is assigned to 0 Hc and all note symbols are converted to their respective theoretical scale degree values (i.e. the symbol B4♭2 is converted to 7Hc, where the *karar* note of a piece is A4 = 0Hc). Then each value obtained from the theoretical intervals is interchanged with the scale degrees in the performance obtained through histogram analysis (Section 3.2.1), provided that there is a single prominent peak observed in the pitch histogram in the vicinity of the theoretical value (a maximum distance of 1 Hc). The rests in the score are assigned the same *pseudo*-value, which was noted in audio pitch contour generation (Section 3.2.1). Then, the note and time sequences are divided into sections by using the event number of the start of each section. Finally a synthetic pitch contour of each section is generated from the durations and the Hc values of the note sequences in the segments with a sampling period of 10ms to match the hop size of the pitch contour. Like for the audio pitch contour (Section 3.2.1), the synthetic pitch contours are downsampled by 10.

[4] Holderian commas are picked due to their common usage in scholarly articles about *makam* music of Turkey.

Table 1: Structural element defined for the dilation operation

1	1	1	.	.	.	.	.	.
1	1	1	.	.	.	.	.	.
1	1	1	1	.	.	.	.	.
.	.	1	1	.	.	.	.	.
.	.	.	.	1	.	.	.	.
.	.	.	.	.	1	1	.	.
.	.	.	.	.	1	1	1	1
.	.	.	.	.	.	1	1	1
.	.	.	.	.	.	1	1	1

Table 2: Structural element defined for the erosion and opening operations

1	1	.	.	.
1	1	.	.	.
.	.	1	.	.
.	.	.	1	1
.	.	.	1	1

3.3 Candidate Estimation

After computing the pitch contours the method tries the estimate candidate time locations that can form the links between each section of the score and the audio recordings. Similarity matrices are calculated by taking the City Block (L1) distance [27] between each point of the synthetic pitch contour associated with the section and the audio pitch contour. The similarity matrices are normalized so that the distances stay between 0 and 1.

In the normalized similarity matrices long, diagonal "valleys" are observed, which identify the regions where a section in the score might have been performed and are present in the audio recording. In order to detect these diagonal shapes, we first emphasize them by utilizing a number of structural morphological operations [21, 22]. To properly apply morphological operations, the similarity matrix is first subtracted from 1 such that the "valleys" become "hills." Then, the image is dilated. The structural element is picked as a binary diagonal beam lying in the 2^{nd} and 4^{th} quadrants with the focus at the origin (Table 1). Next, the similarity matrix is eroded twice. The structural element is a similar beam to the beam defined for dilation, but smaller (Table 2). Later, to remove noises, the similarity matrix is opened with the same structuring element

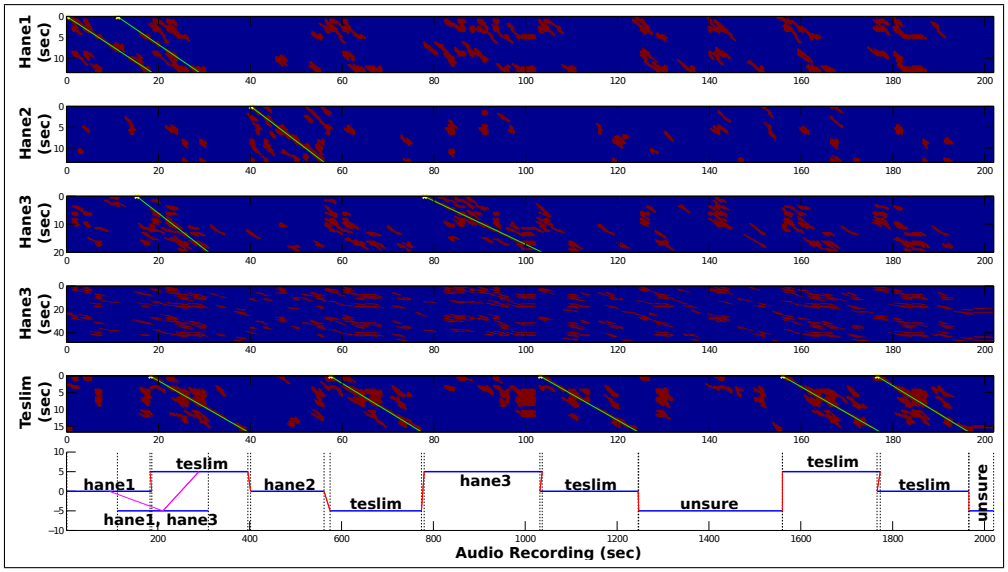

Figure 2: Section candidates shown on top of the processed similarity matrices, estimated for an audio recording of Muhayyer Saz Semâi (recording #29 in Table 3) and groups connected prior to hierarchical linking. Horizontal blue lines show the group borders, red lines indicate connections of preceding and following groups and pink links mark overlapping regions.

used in the erosions. Next, the similarity matrices are converted into binary images by applying thresholding, such that all values higher than 0.96 are given the value one and all other values are assigned to zero. Structural component analysis is done on the binary image to find the blobs. All blobs that are not in the desired diagonal orientation (i.e. lying between 0 and -90 degrees) are removed. From the remaining blobs only the biggest 20% are picked. As a last step in pre-processing the similarity matrix, the image is dilated by a 3x3 square structuring element to slightly widen the diagonals.

After pre-processing the similarity matrices, Hough transform [22] is applied on each similarity matrix to detect the prominent lines. The peaks between -25 and -65 degrees are detected in the transformation matrix, and the peaks which have accumulated a value higher than .3 are picked. The detected peaks are then used to extract line segments: in this process only the lines which are longer than 150 pixels are selected. Since the diagonals are actually blobs, there are a number of lines in the same region with minimal variances in locations and angles: all of these lines are removed except the longest one. Moreover, some prominent diagonals might have discontinuities resulting in more than one line segment on different parts of a diagonal. These lines are connected with each other provided their combined projection to the score (i.e. the range in the corresponding y axis) covers more than 60% of the score. Finally, all line segments covering more than 70% of the score are extrapolated to the edges and all other lines are removed. By combining the parallel results, candidate locations for all sections are obtained.

3.4 Hierarchical Linking

Through inspecting the candidates obtained from the estimations of each section, most of the sections may be linked with their corresponding regions in the audio recording. Nevertheless, there might be some erroneous candidates in several locations apart from the true location. Since the candidate estimations for each section are temporally independent from each other, such erroneous links might overlap or enclose other candidates, and produce conceptually problematic outcomes. Moreover, there might also be some unsure regions where no candidate was estimated.

Nevertheless, since the sequence of the sections in the score is known, an additional step making use of the sequence of the sections given in the composition might be introduced. This step would be hierarchically able to eliminate any erroneous candidates and guess unsure regions, and therefore increase the overall accuracy of the method.

First, the candidates are gathered such that when the borders of a candidate is inside the borders of another (i.e. one candidate is enclosing another), they are grouped together. Since there is always a chance for the shorter candidate to be exceeding a border of the longer candidate by a very small duration, an expansion outside the border of the longer candidate by less than 10% of the duration of the longest candidate is tolerated. Next, regions, where candidate estimation did not predict any candidates, are labeled as "unsure." Afterwards, these groups are connected together so that any preceding, following and overlapping groups may be traversed (Figure 2).

After the enclosing groups are formed, linking is commenced iteratively. First, any non-overlapping groups having a single candidate are temporarily linked. Next, each *hane* candidate is checked whether its location is impossible with respect to already linked candidates. For example, if a 2^{nd} *hane* is linked and there are other 2^{nd} *hane* candidates occurring later in the audio recording, which are not directly connected to the link (i.e. a sequence of $\{2^{nd}$ *hane,* 2^{nd} *hane*$\}$ is not observed) or through an unsure re-

Table 3: The dataset used in the experimentation. h_n, t and u stand for the n^{th} hane, teslim and unrelated region respectively. $t*$ indicates ends of the teslims vary in the composition.

Rec. #	Composition	Composer	Instrumentation	Dur.	Annotations	Remarks on the Recording
1	Acemaşiran Peşrev	Neyzen Salih Dede	Ney	4:19	h_1, h_2, h_3, h_4	Kız Ahenk
2			Ney	4:22	h_1, h_2, h_3, h_4	Kız Ahenk
3			Ney	4:22	h_1, h_2, h_3, h_4	Mansur Ahenk
4	Hicaz Saz Semâî	Muhittin Erev	Ney	4:00	h_1, t, h_2, t, h_3, t, h_4, t	Kız Ahenk
5			Ney	4:00	h_1, t, h_2, t, h_3, t, h_4, t	Mansur Ahenk
6	Hüseyni Peşrev	Kul Mehmet	Ney	5:21	h_1, h_2, h_3, h_4	Kız Ahenk
7			Ney	5:22	h_1, h_2, h_3, h_4	Mansur Ahenk
8	Hüseyni Peşrev	Lavtacı Andon	Ensemble	5:17	h_1, t*, h_2, t*, h_3, t*, h_4, t*, u	Silence in the End
9			Ensemble	5:15	h_1, t*, h_2, t*, h_3, t*, h_4, t*, u	Silence in the End
10	Hüseyni Saz Semâî	Lavtacı Andon	Ney	4:48	h_1, t, h_2, t, h_3, t, h_4, t	Kız Ahenk
11			Ney	4:48	h_1, t, h_2, t, h_3, t, h_4, t	Mansur Ahenk
12	Hüseyni Saz Semâî	Tatyos Efendi	Ensemble	3:01	h_1, t, h_2, t, h_3, h_3, t, h_4, t, t, u	Silence in the End
13			Ensemble	5:38	h_1, t, t, h_2, h_2, t, t, h_3, h_3, t, t, h_4, t, t, u	Silence in the End
14			Tanbur, Kemençe	3:21	h_1, t, t, h_2, t, h_3, h_3, t, t, h_4, t, t, u	Repetitions in Hane 4 Omitted Silence in the End
15			Ud	7:31	u, h_1, h_1, t, t, h_2, h_2, t, t, h_3, h_3, t, t, h_4, t, t, u	Speech and Taksim in the Start Taksim and Silence in the End
16	Kürdilihicazkar Peşrev	Vasilaki	Ensemble	1:10	h_1, t*	Partial Performance
17			Ensemble	1:11	h_1, t*	Partial Performance
18			Tanbur	4:05	h_1, t*, h_2, t*, h_3, t*, h_4, t*, u	Denoised Recording of Below Silence in the End
19			Tanbur	4:07	h_1, t*, h_2, t*, h_3, t*, h_4, t*, u	Noisy Recording Silence in the End
20			Ud	4:19	h_1, t*, h_2, t*, h_3, t*, h_4, t*, u	Silence in the End
21			Ensemble	5:48	h_1, t*, h_2, t*, h_3, t*, h_4, t*, u	Silence in the End
22			Ensemble	2:07	h_1, t*, h_2, t*	Partial Performance
23	Muhayyer Saz Semâî	Tanburi Cemil Bey	Ud	6:32	u, h_1, t, t, h_2, t, t, h_3, t, t, h_4, t, t, u	Silence in the Start and the End
24			Ud	4:08	h_1, t, t, h_2, t, t, h_3, t, t, h_4, t, t, u	Silence in the End
25			Ud	4:16	h_1, t, t, h_2, t, t, h_3, t, t, h_4, t, t, u	Silence in the End
26			Ensemble	5:33	h_1, t, t, h_2, t, t, h_3, t, t, h_4, t, t, u	Silence in the End
27			Ney	4:20	h_1, t, h_2, t, h_3, t, h_4, t	Kız Ahenk
28			Ney	4:20	h_1, t, h_2, t, h_3, t, h_4, t	Mansur Ahenk
29			Ensemble	3:22	h_1, t, h_2, t, h_3, t, h_4, t, t, u	Silence in the End
30	Rast Peşrev	Osman Bey	Ney	4:10	h_1, t, h_2, t, h_3, t, h_4, t	Kız Ahenk
31			Ney	4:09	h_1, t, h_2, t, h_3, t, h_4, t	Mansur Ahenk
32	Segah Saz Semâî	Yusuf Paşa	Ensemble	2:36	h_1, t*	Partial Performance
33			Violin	7:35	u, h_1, t*, h_2, t*, h_3, t*, h_4, t*, u	Silence in the Start and the End
34			Ney, Percussion	3:27	h_1, t*, h_2, t*	Percussion is Recorded Loud
35			Cello, Viola	14:03	h_1, t*, h_2, t*, h_3, t*, h_4, t*, u	Group Taksim, Suzidil Saz Semaisi and Silence in the End
36			Ney, Kanun	6:39	h_1, t*, h_2, t*, h_3, t*, h_4, t*	
37	Uşşak Saz Semâî	Salih Dede	Tanbur	6:45	h_1, t, t, h_2, t, t, h_3, t, t, h_4, h_4, t, t	
38			Tanbur, Kemençe	4:16	h_1, t, h_2, t, h_3, t, h_4, t, u	Silence in the End
39			Ud	5:53	h_1, t, t, h_2, t, t, h_3, t, t, h_4, t, t	
40			Tanbur	5:44	h_1, t, t, h_2, t, t, h_3, t, t, h_4, t, t, u	Silence in the End
41			Kemençe	5:20	h_1, t, h_2, t, h_3, t, t, h_4, u, h_4, t, t, u	Taksim in the Middle Silence in the End
42			Ney	5:56	h_1, t, h_2, t, h_3, t, h_4, t	Kız Ahenk
43			Ney	5:56	h_1, t, h_2, t, h_3, t, h_4, t	Mansur Ahenk
44			Ney	7:16	h_1, t, t, h_2, t, t, h_3, t, t, h_4, t, t	Müstahsen Ahenk

gion (i.e. a sequence of $\{2^{nd}$ *hane, unsure,* 2^{nd} *hane*$\}$ is not observed), these future candidates are removed even if they are already linked. Moreover any earlier candidates which should not occur before a hane link (i.e. 3^{rd} *hane* and 4^{th} *hane* candidates occurring before a 2^{nd} hane link) or should not occur after a hane link (i.e. 1^{st} *hane* candidates occurring after a 2^{nd} *hane* link) are removed. This way, most of the false positives occurring before and after the true *hane* link may be taken care of, while linking *hane* repetitions and expressive elements not related to the composition (i.e. *taksim* etc.) between two *hane*s of the same label are still allowed.

After this step, the indices of links (i.e. order of the section given in the score) are noted, where possible. Since each *hane* has an unique index in the score, our starting point is to note the indices of the linked *hane*s. For example, if the score is in the form $\{1^{st}$ *hane, teslim,* 2^{nd} *hane,* ..., 4^{th} *hane, teslim*$\}$, the index of a 2^{nd} hane link will be 3. If a *teslim* or a *teslim* repetition is found, the index will be the index of the previous neighboring *hane* plus one or the index of the next neighboring *hane* minus one, provided either one is known. If the indices of both the previous and the next neighboring *hane* link is known, they must be consecutive (i.e. $\{1^{st}hane, teslim(s), 2^{nd}hane\}$), or the indices for the *teslim* will be left indeterminate. The indices of the links are used to estimate the unsure groups and groups with mulitple candidates, which will be explained later.

Through inspecting the enclosing groups, it was seen that if a group is overlapping with at least two other groups, the candidates inside the group are almost never true positives. All such overlapping groups are removed to increase precision in exchange with a minimal-to-zero decrease in recall.

After each step, if all the candidates of an enclosing group is removed, the group is assigned "unsure." Moreover, if an unsure group is followed by another, both groups are merged into one. Unsure groups are also not allowed to overlap with other groups. If such a case occurs the interval overlapping with the other groups is trimmed from the unsure group.

The final confusion arises when a group does not have any candidates (unsure group) or there are at least two candidates that are both linkable. To guess an unsure group, both of the immediate neighbor groups must be already linked [5]. If the neighbors are consecutive *hane*s, the algorithm predicts a *teslim* for the unsure group. If both of the neighbors are *teslim*s, the algorithm predicts a *hane* in between, provided that at least one of the composition index of the (*teslim*) neighbors are previously noted. If both indices are known, they must be even consecutive [6] so that there can only be a single *hane* nominee. If these conditions are not met and only one of the neighbors is a *teslim*, the algorithm predicts a *teslim* repetition. Otherwise, the group is left as unsure. For groups, which multiple candidate are possible, the same operation is done. Nevertheless, a multiple-candidate group only requires a single neighbor to be linked before. Moreover, if the unlinked neighbor has more than one candidate (i.e. it is also a multi-candidate group), all candidates in this neighboring group are considered one-by-one to link the multi-candidate group.

The iterative process is finished if no border changes or linking is done in a cycle. Afterwards the gaps between each neighboring link are closed provided there is one. The first and the final links are also widened to the start and the end of the audio recording provided the are not further from the start/end more than 10% of the duration of the longest candidate. Finally, all of the remaining unsure regions are converted to links indicating regions which indicate unrelated parts in the performance with respect to the given composition.

4. EXPERIMENTS

To test the methodology, we have gathered scores of instrumental pieces and the corresponding audio recordings (Section 4.1). The method is applied to each audio recording, linking the sections marked in the score with the corresponding audio fragments. The links found between the audio recordings and scores are then compared with manually linked regions (Section 4.2).

4.1 Data

For the experiments we have used a set of 44 audio recordings associated with 11 scores of different compositions (Table 3) [7]. The scores and parallel audio recordings come from the *CompMusic* database, the *SymbTr* database [24] and the *Instrumental Pieces Played with the Ney* collection [8].

All the scores follow the Arel-Ezgi-Uzdilek theory. In the experiments, we are using a single score per composition, which is either obtained from the *SymbTr* database or obtained by encoding the scores as the symbTr files [24] by referring to the version given in the *Instrumental Pieces Played with the Ney*. As score fragments, we use the actual sections of the pieces, a total of 53 fragments. All of the audio recordings are in *wav* format and either public-domain or commercially available. The recordings encompass a wide variety of instrumentation (Table 3) such as solo ney recordings, which are monophonic; solo stringed instruments, which involve heterophonic peculiarities; duo, trio and ensembles, which are heterophonic. The recordings also cover a substantial amount of expressive decisions such as changes in performance speed, different density of embellishments, note suspension and repetitions, melodic excerpts played in different octaves and various *ahenk*s. Some of the recordings include some material that is not related to the scores such as *taksim*s (non-metered improvisations), applauses, introductory speeches, silences and even other pieces of music. These audio materials are not manually removed.

[5] With the exception of the first and the last groups since they are in the start and end of the recording respectively. For the first and the last groups respectively, only the next and previous groups are needed to be linked before.

[6] Since both *saz semaisi* and *peşrev* forms start with 1^{st} *hane* , *teslim*s always occupy even indices.

[7] The data will be available in `http://compmusic.upf.edu/`

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

Table 5: The results of the section linking experiment including all audio recordings. *K-*, *K+*, *H-* and *H+* indicate results obtained from fully-automatic karar recognition, semi-automatic karar recognition, candidate estimation and hierarchical linking respectively.

	K-H-	K+H-	K-H+	K+H+
Accuracy	65.17%	69.83%	73.45%	80.45%
Specificity	0%	0%	13.33%	13.04%
Recall	72.38%	79.28%	81.01%	89.11%
Precision	86.75%	85.42%	88.15%	88.86%
F_1 score	78.92%	82.23%	84.43%	88.98%
F_3 score	73.60%	79.86%	81.67%	89.08%

Almost all the pieces in the *Instrumental Pieces Played with the Ney* collection include both the audio recording and the score used by the musician to play from. The procedure of adding a piece to the collection is as follows: **1.** The musician looks a few scores of the same composition, picks the one she/he prefers; **2.** The musician makes corrections to the score if necessary; **3.** The musician performs the piece while referring to the score.

4.2 Results and Evaluation

To evaluate the method, we built the ground truth by manually identifying the particular fragment of the score section by labeling the time boundaries in the audio recordings. A composition-related link is deemed as true positive, if and only if it is coinciding with an annotation of the same section, and the average distance between the borders of the annotation and the link does not exceed 10% of the duration of the annotation. Links, which do not meet these constraints are treated as false positives. If a composition related annotation does not coincide with any link with the distance constraint given above, it is labeled as a false negative.

Since the system is not meant to identify what a non-related region actually is, the boundaries of the links labeled as "unrelated" do not have to coincide with the borders of an unrelated annotation. Therefore, any consecutive unrelated regions (i.e. introductory speech followed by a *taksim*) are combined into a single one, and evaluation is done on the links which are enclosed by a non-compositional region. Links enclosed by a non-compositional region are obtained by the enclosing operation explained in Section 3.4. All links labeled as "unrelated" enclosed by a non-compositional annotation are labeled as true negative. All other enclosed links are treated as false positives. Any unguessed parts in these annotations are neither awarded or penalized.

We have computed accuracy, specificity, recall, precision, F_1-score and F_3-score from the true positives, true negatives, false positives and false negatives. These results are reported for both candidate estimation and hierarchical linking. The automatic *karar* recognition obtained via *Makam Toolbox* has failed in 7 pieces (recordings #1, #2, #3, #6, #7, #8 and #22, indicated as bold in Table 4), which are corrected via the graphical interface of the *Makam* Toolbox. The true positive, true negative, false positive, false negative scores calculated per experiment is given in Table 4. The global accuracy, specificity, recall, precision, F_1 score and F_3 score obtained from the candidate estimation and hierarchical linking with automatic and semi-automatic *karar* recognition are given in Table 5.

In order to assess the effectiveness of pitch contours proposed, it is necessary to check the results obtained from the candidate estimation with respect to the density of heterophonic and expressive elements. However, it is not straightforward to directly measure the level of heterophony and expressivity of an audio recording. On the other hand, since these elements are related to instrumentation, the results obtained from candidate estimation are grouped and compared with respect to different types instrumentation (Table 6).

The time elapsed per experiment are also recorded. The timings are then normalized with respect to the duration of the audio recordings with the given formula:

$$t_{Ni} = \frac{t_i}{dur_i} * \frac{\sum_i^n dur_i}{n} \quad (1)$$

where t_i is the time elapsed during the section linking, dur_i is the duration of the i^{th}audio recording and n is the number of the recordings (Table 5). It takes an average of 42 seconds with a standard deviation of 15 seconds to link the sections of a audio recording approximately 275 seconds long (i.e. the average duration of an audio recording in the dataset), when the implementation is run on computer with a 4GB RAM and 2.26 GHz processor.

5. DISCUSSION

The results in Table 5 points that the methodology is quite successful in linking sections given in the scores with the corresponding audio recordings. The method is able to deal with a wide number of situations such as compositions without any section repetitions, various *ahenk*s, partial performances, *hane* or *teslim* repetitions and recordings with unrelated parts. Table 5 also shows that hierarchical linking has a clear success over candidate estimation, even when failed *karar* detections are not altered.

The advantage of the hierarchical linking is more evident, when results per piece (Table 4) are inspected. Except the 14^{th} experiment, where candidate estimation produced one erroneous link enclosing a true link and hierarchical linking preferred the erroneous one, hierarchical linking emits more true positives and less false negatives. Moreover, there is no increase in the number of false positives obtained through all experiments, thus hierarchical linking presents much better precision, recall and f-scores over evaluation on raw links provided by the section estimation.

The results also show that the pitch contours successfully allow a flexible means of section linking specific to *makam* music of Turkey. Nevertheless, in Table 6, it can be seen that as the instrumentation of a recording gets more complex, i.e. the tendency of observing heterophonic and expressive elements in an audio recording increases, the accuracy and the F_1-score decreases almost monotonically.

Table 4: The results per piece. t and t_N indicate the time and normalized time elapsed per experiment with semi-automatic *karar* recognition. *K-*, *K+*, *H-* and *H+* indicate results obtained from fully-automatic karar recognition, semi-automatic karar recognition, candidate estimation and hierarchical linking respectively.

	#Sections /	t / t_N	True Positive				True Negative				False Negative				False Positive			
Rec. #	#Unrelated	(sec)	K-H-	K+H-	K-H+	K+H+	K-H-	K+H-	K-H+	K+H+	K-H-	K+H-	K-H+	K+H+	K-H-	K+H-	K-H+	K+H+
1	4	32 / 34	**0**	4	**0**	4	0	0	0	0	**4**	0	**4**	0	0	2	0	0
2	4	26 / 27	**0**	4	**0**	4	0	0	0	0	**4**	0	**4**	0	0	3	0	0
3	4	26 / 28	**0**	4	**0**	4	0	0	0	0	**4**	0	**4**	0	0	3	0	0
4	8	28 / 32	7	7	8	8	0	0	0	0	1	1	0	0	0	0	0	0
5	8	33 / 37	7	7	8	8	0	0	0	0	1	1	0	0	1	1	0	0
6	4	39 / 33	**0**	4	**0**	4	0	0	0	0	**4**	0	**4**	0	0	1	0	0
7	4	39 / 33	**0**	4	**0**	4	0	0	0	0	**4**	0	**4**	0	0	0	0	0
8	8 / 1	30 / 26	**0**	3	**0**	5	0	0	0	0	**8**	5	**8**	3	0	0	0	1
9	8 / 1	32 / 28	7	7	8	8	0	0	0	0	1	1	0	0	0	0	0	0
10	8	27 / 32	7	7	8	8	0	0	0	0	1	1	0	0	1	1	0	0
11	8	27 / 32	8	8	8	8	0	0	0	0	0	0	0	0	3	3	0	0
12	10 / 1	28 / 42	4	4	5	5	0	0	1	1	6	6	5	5	1	1	2	2
13	14 / 1	67 / 66	10	10	12	12	0	0	0	0	4	4	2	2	0	0	2	2
14	12 / 1	46 / 62	11	11	10	10	0	0	0	0	1	1	2	2	2	2	2	2
15	15 / 2	126 / 89	13	13	14	14	0	0	2	2	2	2	0	0	1	1	3	3
16	2	13 / 53	2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0
17	2	14 / 52	1	1	2	2	0	0	0	0	1	1	0	0	0	0	0	0
18	8 / 1	30 / 34	7	7	7	7	0	0	0	0	1	1	1	1	0	0	1	1
19	8 / 1	28 / 31	5	5	6	6	0	0	0	0	3	3	2	2	0	0	0	0
20	8 / 1	29 / 30	5	5	6	6	0	0	0	0	3	3	2	2	0	0	1	1
21	8 / 1	32 / 30	4	4	8	8	0	0	0	0	4	4	0	0	0	0	0	0
22	4	17 / 37	**0**	2	**0**	4	0	0	0	0	**4**	2	**4**	0	0	0	0	0
23	12 / 2	40 / 33	5	5	7	7	0	0	1	1	7	7	5	5	0	0	2	2
24	12 / 1	32 / 36	4	4	7	7	0	0	0	0	8	8	5	5	1	1	1	1
25	12 / 1	59 / 63	7	7	8	8	0	0	0	0	5	5	3	3	1	1	5	5
26	12 / 1	50 / 50	7	7	10	10	0	0	0	0	5	5	2	2	0	0	2	2
27	8	28 / 29	7	7	8	8	0	0	0	0	1	1	0	0	0	0	0	0
28	8	31 / 33	7	7	7	7	0	0	0	0	1	1	1	1	1	1	2	2
29	9 / 1	40 / 54	8	8	9	9	0	0	0	0	1	1	0	0	2	2	1	1
30	8	33 / 36	8	8	8	8	0	0	0	0	0	0	0	0	3	3	0	0
31	8	36 / 39	8	8	8	8	0	0	0	0	0	0	0	0	5	5	0	0
32	2	15 / 44	2	2	2	2	0	0	0	0	0	0	0	0	1	1	0	0
33	8 / 2	45 / 32	7	7	8	8	0	0	1	1	1	1	0	0	4	4	0	0
34	4	23 / 31	4	4	4	4	0	0	0	0	0	0	0	0	0	0	0	0
35	8 / 1	101 / 33	5	5	5	5	0	0	0	0	3	3	2	2	2	2	4	4
36	8	44 / 36	8	8	8	8	0	0	0	0	0	0	0	0	3	3	1	1
37	13	76 / 61	12	12	12	12	0	0	0	0	1	1	0	0	0	0	2	2
38	8 / 1	32 / 34	7	7	7	7	0	0	0	0	1	1	1	1	2	2	1	1
39	12	61 / 57	11	11	12	12	0	0	0	0	1	1	0	0	0	0	1	1
40	12 / 1	93 / 90	10	10	10	10	0	0	1	1	2	2	2	2	1	1	1	1
41	11 / 2	63 / 54	9	9	10	10	0	0	0	0	2	2	1	1	0	0	2	2
42	8	39 / 37	8	8	8	8	0	0	0	0	0	0	0	0	0	0	1	1
43	8	41 / 38	8	8	8	8	0	0	0	0	0	0	0	0	2	2	1	1
44	12	69 / 44	12	12	12	12	0	0	0	0	0	0	0	0	3	3	1	1
Total	364 / 24	1817 / 1831 (**Av:** 41 / 42)	262	287	290	319	0	0	6	6	100	75	68	39	40	49	39	40

Table 6: The results obtained from the candidate estimation with semi-automatic *karar* detection. The results are grouped per instrumentation. *#Rec., #Sec., #Un., tp, fn, fp, Accur., Precis.,* F_1, F_3 stand for number of recordings, number of sections, number of unrelated regions, number of true positives, number of false negatives, number of false positives, accuracy, precision, F_1-score and F_3-score respectively.

	#Rec.	#Sec.	#Un.	tp	fn	fp	Accur.	Recall	Precis.	F_1	F_3
Solo Ney	17	116	0	111	5	28	77.08%	95.69%	79.86%	87.06%	93.83%
Solo Stringed	12	131	14	95	36	8	68.35%	72.52%	92.23%	81.20%	74.10%
Duo / Trio	4	36	3	31	5	9	68.89%	86.11%	77.50%	81.58%	85.16%
Ensemble	11	79	7	50	29	4	60.24%	63.29%	92.59%	75.19%	65.36%
All	44	362	24	287	75	49	69.83%	79.28%	85.42%	82.23%	79.86%

This suggests that an improvement in the extraction of audio pitch contour is necessary. Through inspecting errors in the audio recording level, it is seen that the current bottleneck of the system is the pitch estimation. Since YIN is designed for monophonic sounds, lots of confusions arise in the fundamental frequency estimations due to the heterophonic nature of *makam* music, especially in ensemble performances. Moreover, YIN is found to lose its robustness, where there are substantial usage of expressive elements such as legatos, slides and tremolos. This problem should be tackled by using multi-pitch extraction and prominent melody detection [28].

A second problem occurs when the performers substantially deviate from the score i.e. a performer suspends the note while the rest of the performers continue playing, some notes in an melodic excerpt is played an octave up/down. In these situations, Hough transformation detects either a short, single line segment or several line segments in the region, where a section is being performed. However, as explained in (Section 3.3), the synthetic pitch contour do not link to its corresponding location in the performance under these circumstances, unless 70% of the section is covered by the line segments. To handle these problems, a metric, which compensates for octave differences might be devised, analogous to octave-resilient methods used for Western music [29]. Moreover, arithmetic geometry operations might be made more flexible by removing the 70% coverage constraint and using the ratio of the coverage as a confidence measure for hierarchical linking. This way, the method will be allowed to link partial similarity between the pitch contours.

It is also observed that hierarchical linking predicts a considerable amount of regions which candidate estimate do not (100 vs. 68 false negatives with automatic *karar* recognition and 75 vs. 39 false negatives with semi-automatic *karar* recognition). Most of the remaining false negatives (30 false negatives out of 39, and 11 related false positives out of 40 with semi-automatic *karar* detection) after hierarchical linking are due to Hough transformation not able to yield any links for regions encompassing at least two consequent composition related annotations in the previous step. These regions might be linked to multiple sections by allowing hierarchical linking make multiple decisions based on the duration of the particular region with respect to the previously linked sections. Nevertheless, the core reason of this type of confusion is due to the partial differences in the pitch contours explained above. We predict that by implementing the relevant measures proposed above, this type of confusions will diminish without rendering the hierarchical linking step much more complex.

Another drawback of the method is the detection of the unrelated regions in hierarchical linking [9]. In this step, unrelated links are currently found indirectly by locating related sections. Even if there are no estimations given for a unrelated region after candidate estimation, hierarchical linking typically predicts an erroneous link in these regions (16 false positives out of 40 with semi-automatic *karar* detection), resulting in a low specificity. To increase the detection of true negatives, some direct means of linking the audio signal with some types of unrelated events, i.e. through silence and speech detection, may be useful.

Currently hierarchical linking does not have any restrictions on the duration of a candidate link. By adding some constraints in the duration of links (i.e. comparison of the performance speed of a candidate in the audio recording by the speed of its synthetic pitch contour and the speeds of the pitch contours of other sections already linked), an ample amount of erroneous links to silent regions and regions spanning to multiple annotations may also be avoided. Moreover, since the current approach for hierarchical linking is completely rule-based, every single special case should be considered explicitly, which makes the implementation hard to maintain and prone to errors. This type of situation is highly suitable for applying principles of fuzzy logic [30]. Fuzzy logic might also lower the complexity of the code and increase human readability.

6. CONCLUSION AND FUTURE WORK

We have proposed a method to link sections of a musical score of a composition with the corresponding regions in an audio recording of the performance of the same composition. We have tested the method with 11 instrumental compositions of *makam* music of Turkey associated with 44 audio recordings, obtaining remarkable performance in a fast operation time.

Since a score section is basically a sequence of note events, the candidate estimation step might be generalized to link any type of melodic fragments with an audio recording. A generalized fragment linking methodology might be helpful in computational tasks such as audio-score alignment, embellishment detection, tonic analysis, tuning detection, intonation analysis and version detection. Conversely, the candidate estimation methodology might require specific adjustments for each task. Comparative candidate estimation experiments should be carried using other techniques such as general Hough transform [22], SAX [31], dynamic programming [18], minimal geodesics [23].

Currently, candidate estimation uses similarity matrices computed from descriptors which are specifically designed for *makam* music. Similarly, the method can be adapted to other musical cultures by computing descriptors, which are musically relevant to the culture being studied. As an example, semi-improvised jazz music performances, where musicians build variations of predefined melodies through improvisation, share a similar basis with *makam* music. Instead of generation of a monophonic pitch contour from the score based on the properties of *makam* music, generation of a harmonic contour from the initial melody based on jazz harmony might be useful to traverse the variations through out a performance. Also, candidate estimation and hierarchical linking might be adapted to structure analysis in Western music by replacing the pitch contours with some harmonic descriptors and using a multi-dimensional distance metric to calculate a similarity matrix.

[9] Note that candidate estimation does not currently produce any unrelated links since it conceptually only tries to link patterns it is provided, and leaves the time-related decisions to the hierarchical linking step.

Acknowledgments

We would like to thank Barış Bozkurt and Kemal Karaosmanoğlu for providing us data and Marcelo Bertalmío for the insightful discussions. We would also like to thank Mehmet Yücel for the *Instrumental Pieces Played with the Ney* dataset and the all musicians whose recordings made this project possible. This research was funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement 267583 (CompMusic Project).

7. REFERENCES

[1] A. Gedik and B. Bozkurt, "Pitch-frequency histogram-based music information retrieval for Turkish music," *Signal Processing*, vol. 90, no. 4, pp. 1049–1063, 2010.

[2] B. Bozkurt, O. Yarman, M. K. Karaosmanoğlu, and C. Akkoç, "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*, vol. 38, no. 1, pp. 45–70, Mar. 2009.

[3] S. Şentürk, *Computational modeling of improvisation in Turkish folk music using variable-length markov models*, Master's thesis, Georgia Institute of Technology, 2011.

[4] A. Holzapfel, *Similarity methods for computational ethnomusicology*, Ph.D. dissertation, University of Crete, 2010.

[5] E. B. Ederer, *The theory and praxis of makam in classical Turkish music 1910-2010*, Ph.D. dissertation, University of California, Santa Barbara, September 2011.

[6] Y. Tura, *Türk Musıkisinin Meseleleri*. Pan Yayıncılık, 1988.

[7] E. Popescu-Judetz, *Meanings in Turkish Musical Culture*. Pan Yayıncılık, 1996.

[8] I. Özkan, *Türk mûsikısi nazariyatı ve usûlleri: Kudüm velveleleri*. Ötüken Neşriyat, 2006.

[9] M. E. Karadeniz, *Türk Musıkisinin Nazariye ve Esasları*. İş Bankası Yayınları, 1984, p. 159.

[10] H. Myers, *Ethnomusicology: an Introduction*. WW Norton, 1992, pp. 110–164.

[11] F. W. Stubbs, *The art and science of taksim: an emprical analysis of traditional improvisation from 20th century Istanbul*, PhD dissertation, Wesleyan University, 1994.

[12] K. Signell, *Makam: Modal practice in Turkish art music*. Da Capo Press, 1986.

[13] O. Lartillot and M. Ayari, "Cultural impact in listeners' structural understanding of a Tunisian traditional modal improvisation, studied with the help of computational models," *Journal of interdisciplinary music studies*, vol. 5, no. 1, pp. 85–100, 2011.

[14] M. Cooper and J. Foote, "Automatic music summarization via similarity analysis," in *Proceedings of ISMIR 2002*, 2002, pp. 81–85.

[15] M. Goto, "A chorus-section detecting method for musical audio signals," *Proceedings of ICASSP 2003*, vol. 5, 2003, pp. 437–440.

[16] J. Paulus, M. Müller, and A. Klapuri, "State of the art report: Audio-based music structure analysis," *Proceedings of ISMIR 2010*, 2010, pp. 625–636.

[17] D. Ellis and G. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," *Proceedings of ICASSP 2007*, vol. 4, 2007, pp. 1429–1432.

[18] J. Serrà, X. Serra, and R. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, 093017, 2009.

[19] A. Holzapfel and Y. Stylianou, "Rhythmic similarity in traditional Turkish music," *Proceedings of ISMIR 2009*, 2009, pp. 99–104.

[20] B. Martin, M. Robine, P. Hanna *et al.*, "Musical structure retrieval by aligning self-similarity matrices," *Proceedings of ISMIR 2009*, 2009, pp. 483–488.

[21] J. Serra, *Image analysis and mathematical morphology*. Academic Press, 1982.

[22] D. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[23] R. Kimmel and J. Sethian, "Computing geodesic paths on manifolds," *Proceedings of the National Academy of Sciences*, vol. 95, no. 15, p. 8431, 1998.

[24] K. Karaosmanoğlu, "A Turkish makam music symbolic database for music information retrieval: Symbtr," *Proceedings of ISMIR 2012*, 2012.

[25] B. Bozkurt, "An automatic pitch analysis method for Turkish maqam music," *Journal of New Music Research*, vol. 37, no. 1, pp. 1–13, 2008.

[26] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal of Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[27] E. Krause, *Taxicab geometry: An adventure in non-Euclidean geometry*. Dover Publications, 1987.

[28] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[29] M. Muller, S. Ewert, and S. Kreuzer, "Making chroma features more robust to timbre changes," *Proceedings of ICASSP 2009*, 2009, pp. 1877–1880.

[30] G. Klir and B. Yuan, *Fuzzy sets and fuzzy logic*. Prentice Hall, 1995.

[31] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2003, pp. 2–11.