COMPOSITION IDENTIFICATION IN OTTOMAN-TURKISH MAKAM MUSIC USING TRANSPOSITION-INVARIANT PARTIAL AUDIO-SCORE ALIGNMENT

Sertan Şentürk, Xavier Serra
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{sertan.senturk, xavier.serra}@upf.edu

ABSTRACT

The composition information of audio recordings is highly valuable for many tasks such as automatic music description and music discovery. Given a music collection, two typical scenarios are retrieving the composition(s) performed in an audio recording and retrieving the audio recording(s), where a composition is performed. We present a composition identification methodology for these two tasks, which makes use of music scores. Our methodology first attempts to align a fragment of the music score of a composition with an audio recording. Next, it computes a similarity from the best obtained alignment. True audio-score pair emits a high similarity value. We repeat this procedure between all audio recordings and music scores, and filter the true pairs by a simple approach using logistic regression. The methodology is specialized according to the cultural-specific aspects of Ottoman-Turkish makam music (OTMM), achieving 0.96 and 0.95 mean average precision (MAP) for composition retrieval and performance retrieval tasks, respectively. We hope that our method would be useful in creating semantically linked music corpora for cultural heritage and preservation, semantic web applications and musicological studies.

1. INTRODUCTION

Version identification is an important task in music information retrieval which aims to find the versions of a music piece from a collection of audio recordings automatically [1, 2]. For popular music such as rap, pop and rock, the task aims to identify the covers of an original audio recording. For classical music traditions a more relevant task is associating compositions with the audio performances. The composition information is highly useful in many other computational tasks such as automatic content description and music discovery (e.g. searching the performances of a composition in a music collection).

For classical music cultures, music collections consisting of music scores and audio recordings along with editorial metadata are desirable in many applications involving cultural heritage archival, music preservation and musicological studies. Composition identification is a crucial step linking performances and compositions during the creation of such music corpus from unlabeled musical data [3].

Composition information can be used to generate and improve linked musical data, enhance the music content description and facilitate navigation in semantic web applications. Consider a scenario, where a musician uploads his interpretation of a composition to a platform such as SoundCloud, YouTube etc. The performed compositions can be automatically identified and labeled semantically using an ontology, e.g. [4]. Next the performance can be linked with related concepts (e.g. form, composer, score) available in other sources such as biographies of the performing artist, the music score of the composition or the musical and editorial metadata stored in open encyclopedias such as MusicBrainz and Wikipedia. Such a scheme would facilitate searching, accessing and navigating relevant music content in a more informed manner. Likewise, tasks such as enhanced listening and music recommendation may also benefit from the musical data linked via automatic composition identification.

Due to inherent characteristics of the oral tradition and the practice of Ottoman-Turkish makam music (OTMM), performances of the same piece may be substantially different from one another. This aspect brings certain computational challenges for the computational analysis and retrieval of OTMM (Section 2). In this paper, we propose a composition identification methodology, which makes use of the available music scores of the relevant compositions using partial audio-score alignment. The methodology is designed to address the culture-specific challenges brought by OTMM. To the best of our knowledge, our methodology is the first automatic composition identification proposed for OTMM. We consider two composition identification scenarios, **1)** identifying the compositions performed in an audio recording, **2)** identifying the audio recordings in which a composition is performed. Note that there might not be any relevant audio recordings for some compositions, and vice versa. Our methodology also aims to identify such cases. Our contributions can be summarized as:

1. The first composition identification methodology applied to Ottoman-Turkish makam music
2. An open and editorially complete dataset for composition identification in OTMM (Section 5.1)
3. Extending the state of the art in transposition-invariant partial audio-score alignment for OTMM by in-

troducing subsequence dynamic time warping (Section 4.2.2)

4. Simplifications and generalizations of the fragment selection and the fragment duration steps used in the score-informed tonic identification method proposed by [5] and verification of this method on a larger dataset as a side product of the composition identification experiments (Table 1)

For reproducibility purposes, relevant materials such as musical examples, data and results are open and publicly available via the Compmusic Website. [1]

The rest of the paper is structured as follows: Section 2 provides an introduction to Ottoman-Turkish makam music. Section 3 gives a definition of the composition identification tasks we are dealing with. Section 4 explains the methodology applied to both composition identification scenarios explained above. Section 5 presents the experimental setup, the test dataset and the results. Section 6 discusses the obtained results. Section 7 wraps up the paper with a brief conclusion.

2. OTTOMAN-TURKISH MAKAM MUSIC

The melodic structure of most of the traditional music repertoires of Turkey follow the concept of *makam*s [6]. Currently, Arel-Ezgi-Uzdilek (AEU) theory is the mainstream theory for OTMM [6]. AEU theory argues that there are 24 equal intervals and that a whole tone is divided into 9 equidistant intervals. These intervals can be approximated from 53-TET (tone equal tempered) intervals, each of which is termed as a *Holdrian comma* (Hc) [6].

For centuries, OMMT has been predominantly an oral tradition. Since the start of the 20th century, a notational representation extending standard Western music notation has been used in OTMM complementary to the oral practice [7]. This notation typically follows the rules of AEU theory.

Below we list some of the characteristics of OTMM, which pose challenges for composition identification:

- There is no definite tonic frequency (e.g. $A4 = 440$Hz) in the performances. The performed tonic is occasionally transposed due to instrument/vocal range or aesthetic reasons [6]. This necessitates automatic tonic identification for any fully-automatic alignment method (Section 4.2).
- The performances of OTMM occasionally include improvisations played before, after or even within a composition. It is also common to repeat, insert or omit sections of a composition.
- Until the 20^{th} century, most of these music has been strictly transmitted from a master to the students within the oral tradition. This resulted in the musical material propagating differently in different "schools." Therefore, performances of the same composition may differ from each other substantially.
- OTMM is a heterophonic music tradition. Musicians simultaneously perform the same "melodic idea;" Yet they are supposed to show their virtuosity by changing the tuning and the intonation of some intervals, adding embellishments and/or inserting, repeating and omitting notes and phrases during the performance. Melody extraction algorithms might not perform well in recordings with substantial heterophonic interactions [8].
- Most of the scores of OTMM are descriptive and they are transcribed sometimes centuries later. The scores typically notate basic, monophonic melodic lines. They do not usually indicate the heterophony, intonation deviations and other expressive elements observed in the performances.

In the experiments, we focus on *peşrev* and *saz semaisi*, which are the two most common instrumental forms of the classical repertoire. Both *peşrev* and *saz semaisi* typically consist of four non-repeating sections called *hane* and a repetitive section called *teslim* performed between these *hane*s.

3. PROBLEM DEFINITION

Given a specific music collection, two basic composition identification scenarios are:

1. **Composition retrieval:** Identification of the compositions which are performed in an audio recording.
2. **Performance retrieval:** Identification of the audio recordings in which a composition is performed.

These scenarios are ranked retrieval problems where the query is an audio recording and the retrieved documents are the compositions in the composition identification task, and vice versa. In both cases, the common step is to estimate whether a composition and an audio recording are *relevant* to one another. The relevances in the composition identification problems are binary, i.e. 1 if the composition and the audio recordings are paired and 0 otherwise.

The results in both cases can be aggregated by applying this step to multiple documents and queries. Nevertheless, there might be situations where it may be impossible or impractical to retrieve the whole collection, for example restricted access to copyrighted music material or the lack of computational resources in fast-query applications (e.g. real-time composition identification in mobile applications). Moreover, both scenarios might require different constraints to obtain better results and/or process more efficiently. For example, a good performance retrieval method should find multiple relevant audio recordings for a composition; on the other hand only the top ranked documents are important in composition retrieval when more than a single composition is rarely performed in the queried audio recordings (Section 5.1). In this paper, we deal with these two tasks separately and leave the joint retrieval task as a future direction to explore.

4. METHODOLOGY

We assume that the scores of the compositions are available and estimate the relevance by partially aligning the

[1] http://compmusic.upf.edu/node/306

score of a composition (n) with the audio recording of a performance (m). The alignment step in our methodology is based on the score-informed tonic identification procedure described in [5], which we use to obtain the best possible alignment between a score and an audio recording in a manner invariant of the transposition of the performance (Section 4.2). Next, we compute a similarity value $\in [0, 1]$ between the composition and the performance from the best alignment path. We observe a high similarity value, if the composition (n) is indeed performed in the audio recording (m) (Section 4.3). The block diagram of transposition invariant partial-audio score alignment is given in Figure 1. The alignment process is repeated between each audio recording and music score, and a similarity value is obtained for each composition and performance pair in our collection. Finally, the performance-composition pairs with low similarity values are discarded using outlier detection (Section 4.4) and the relevant pairs are obtained.

4.1 Feature Extraction

In audio-score alignment of Eurogenetic music, features which can capture the harmony such as chroma features [9, 10] are typically used. In [8], it is shown that predominant melody performs better for OTMM due the melodic nature of the music tradition. In our method we follow the melodic features proposed for audio-score alignment of OTMM in [5] and [8].

From the audio recording (m), we extract a predominant melody using a version of the methodology proposed in [11], optimized for makam music [12]. [2] The pitch precision of the predominant melody is taken as 7.5 cents ($\approx 1/3$ Hc), which is a suitable value for tracking pitch deviations in makam music [13]. The frame rate of the extracted predominant melody is downsampled from ~ 2.9 ms to ~ 46 ms, which is shown to be sufficient for audio-score alignment in OTMM [8]. We denote the predominant melody extracted from the audio recording (m) as $X^{(m)} = \left(x_1^{(m)}, \ldots, x_{I^{(m)}}^{(m)}\right)$, $m \in [1 : M]$, where M is number of audio recordings in the collection and $I^{(m)}$ is the number of samples in the audio predominant melody (Figure 1d).

From the machine readable score of the composition (n), we first pick a short fragment (either from the start of the score or from the repetition) indicated in the score. We also try different fragment durations in Section 5. Then we sample the note symbols in the note sequence of the selected fragment according to their durations in nominal tempo indicated in the score [8]. In practice, the previous note is commonly sustained in the place of a rest, so we omit the rests in the score and add their duration to the previous note [8]. The sampled symbols are mapped to the theoretical scale-degrees in cents according to the AEU theory such that the tonic symbol is assigned to 0 cents (Figure 1b). The generated synthetic pitch track has a sampling rate of ~ 46 ms, equal to the frame rate of the predominant melody. We denote the synthetic melody computed from the score of the composition (n) as $Y^{(n)} = \left(y_1^{(n)}, \ldots, y_{J^{(n)}}^{(n)}\right)$, $n \in [1 : N]$, where N is number of compositions with scores in the collection and $J^{(n)}$ is the number of samples in the synthetic melody (Figure 1b).

Notice that the unit of the pitch values in the audio predominant melody $X^{(m)}$ is Hertz, whereas the unit of the pitch values in the synthetic melody $Y^{(n)}$ is cents. For proper alignment of $Y^{(n)}$ within $X^{(m)}$ (provided that they are related with each other), $X^{(m)}$ has to be normalized with respect to the tonic frequency.

To identify the tonic, we first compute a pitch class distribution from the audio predominant melody [5]. We use kernel-density estimation to obtain a smooth pitch class distribution without spurious peaks [5]. We select the bin width of the distribution as 7.5 cents (the same as the pitch precision of the audio predominant melody) and use a Gaussian kernel with a standard deviation of 15 cents ($\approx 2/3$ Hc) so that a pitch value contributes in an interval slightly smaller than a semitone, which is reported as optimal for this task [5]. The width of the kernel is selected as 75 cents center to tail (i.e. 5 times the standard deviation) as the contribution to the samples beyond this width are redundant.

Finally we pick the peaks of the distribution as the tonic candidates [5, 13] (Figure 1e). We denote the tonic candidates for the audio recording (m) as $C^{(m)} = \{c_1^{(m)}, \ldots, C_{K^{(m)}}^{(m)}\}$, where $K^{(m)}$ is the number of peaks in the pitch class distribution (Figure 1e). Notable is that the candidates correspond to (stable) pitch classes instead of frequencies. This choice reduces the computational complexity as we will compute the "octave-wrapped" pitch distances in the alignment step (Section 4.2, Equation 2).

4.2 Transposition-Independent Partial Alignment

Assuming a candidate $c_k^{(m)}$ obtained from the pitch class distribution as the tonic frequency, we normalize each pitch sample in the audio predominant melody to cent scale by:

$$\hat{x}_i^{(m,k)} = 1200 \log_2\left(x_i^{(m)} / c_k^{(m)}\right) \quad (1)$$

Note that there are 1200 cents in an octave. We denote the predominant melody normalized with respect to the tonic candidate $c_k^{(m)}$ as $\hat{X}^{(m,k)}$. Next, we attempt to align the score fragment to the corresponding location in the audio recording by searching the synthetic melody $Y^{(m,k)}$, computed from the selected score fragment in the normalized audio predominant melody $\hat{X}^{(m,k)}$. We compare two methods for partial alignment: 1) Hough transform, and 2) Subsequence DTW.

4.2.1 Hough Transform

The Hough transform is a simple and yet effective parametric line detection method [14]. It is previously used in section-level audio-score alignment [8], tonic identification [5] and tempo estimation [15] in OTMM and found to produce comparable results to methodologies using complex models such as hierarchical hidden Markov models [15]. Nevertheless, it cannot handle extensive tempo deviations or insertions, repetitions and omission in a musical phrase since it is a linear operation.

[2] The implementation is available in `https://github.com/sertansenturk/predominantmelodymakam`

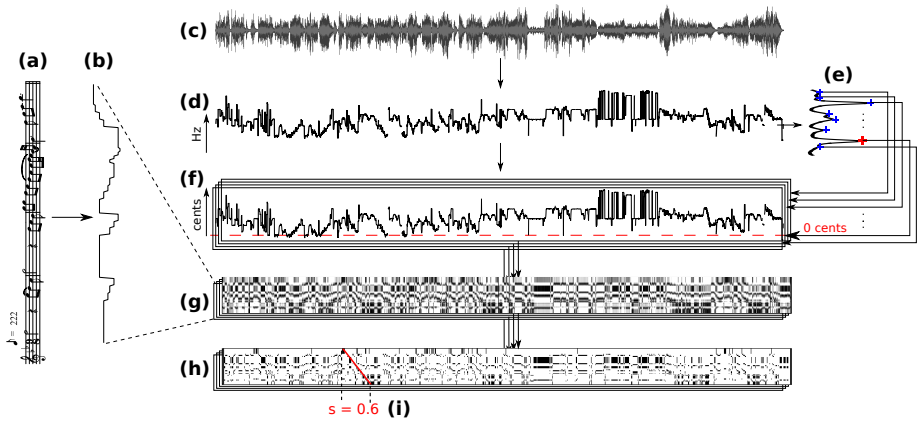

Figure 1. Block diagram of the transposition invariant partial audio-score alignment using the Hough transform **a**) A short fragment selected from the score, **b)** The synthetic pitch computed from the score fragment, **c)** The audio recording, **d)** The predominant melody extracted from the audio recording, **e)** The pitch class distribution computed from the audio predominant melody and its detected peaks, **f)** The set of predominant melodies normalized with respect to the detected peaks, **g)** The set of distance matrices between the synthetic melody and the normalized predominant melodies, **h)** The set of binary similarity matrices computed from the distance matrices. A linear alignment path obtained using the Hough transform is displayed on top of one of the binary similarity matrices along with **i)** the similarity value computed for the path. All the blocks except **g** and **h** are the same for partial alignment using SDTW.

If the Hough transform is selected for partial alignment, a distance metric is computed between the synthetic pitch track $Y^{(n)}$ and the normalized audio predominant melody $\hat{X}^{(m,k)}$. Each element $D(i,j)$ in the distance matrix D is computed as:

$$D\left(i,j\right) = min\Big(\big(|\hat{x}_i^{(m,k)} - y_j^{(n)}|\ \bmod\ 1200\big), \\ 1200 - \big(|\hat{x}_i^{(m,k)} - y_j^{(n)}|\ \bmod\ 1200\big)\Big) \quad (2)$$

where $\hat{x}_i^{(m,k)}$ denotes the i^{th} sample of the normalized audio predominant melody $\hat{X}^{(m,k)}$ and $y_j^{(n)}$ denotes the j^{th} sample of the synthetic pitch $Y^{(n)}$, respectively. This distance may be interpreted as the shortest distance in cents between two pitch classes. It is not affected by octave-errors in the predominant melody or the tonic.

If the selected score fragment is performed within the audio recording and the predominant melody is normalized with the correct tonic frequency, the distance matrix will show blob(s) in a diagonal trajectory formed by low distance values. The projection of the blob to the audio-axis indicates the time-interval in the audio recording where the score fragment is performed. To make the line segment more prominent, we binarize the distance matrix and obtain a binary similarity matrix B (Figure 1h). We use the binarization criteria proposed in [8] and compute each element $B(i,j)$ in the binary similarity matrix as:

$$B(i,j) = \begin{cases} 1, & D\big(i,j\big) \leq \alpha \\ 0, & D\big(i,j\big) > \alpha \end{cases} \quad (3)$$

Here two pitch values are considered to belong to the same note if the distance (in cents) is less than the given binarization threshold, α. We take $\alpha = 50$ cents, which is reported as an optimum of this value for makam music [8].

As can be seen in Figure 1g, these blobs can be approximated as line segments. To detect the line segments, we apply the Hough transform to the binary similarity matrix (Figure 1h). We restrict the searched angles between $-26.57°$ and $-63.43°$, which allows the alignment to have a tempo deviation between 0.5 and 2 times the nominal tempo indicated in the score. From the obtained transformation matrix, we select the highest peak, which indicates the most prominent line segment [14]. The linear path $p^{(m,n,k)}$, which the line segment follows, is simply the sequence of the points that has accumulated this peak in the transformation matrix. An example alignment found by the Hough transform can be seen in Figure 1h.

4.2.2 Subsequence DTW

Dynamic programming and more specifically dynamic time warping (DTW) are the state-of-the-art methodologies for many relevant tasks such as cover song identification [1,2] and audio score alignment [16, 17]. Unlike the Hough transform, DTW is robust to changes in tempo and musical insertions, deletions and repetitions. However, it can be prone to pathological warpings.

We use subsequence DTW (SDTW), which is a typical variant used when one of the time series is a subsequence of the other [18,19]. In this variant the paths are allowed to start/end within target. We refer the readers to [19, Chapter 4] for a thorough explanation of DTW and SDTW.

Using SDTW, we compute an element $A(i,j)$ in the accumulated cost matrix A recursively as:

$$A\left(i,j\right) = \begin{cases} 0, & i = 0 \\ +\infty, & j = 0 \\ D\left(i,j\right) + min \begin{cases} A\left(i-1,j-1\right) & \\ A\left(i-2,j-1\right), & i > 1 \\ A\left(i-1,j-2\right), & j > 1 \end{cases} & i,j \neq 0 \end{cases} \quad (4)$$

As seen above, we select the step size condition as $\{(2,1),$

$(1,1),(1,2)\}$. Analogous to the angle restriction in the Hough transform (Section 4.2.1), this step size ensures that the intra-tempo variations in any path will stay between half and double the nominal tempo indicated in the score. Moreover, we use Equation 2 as the local distance measure to calculate the accumulated cost matrix. Also, notice that the accumulated cost matrix is extended with a zeroth row and column, initialized to enable subsequence matching. Finally we back-track the path $p^{(m,n,k)}$ ending at $\arg\min_{(i)} A(i, J^{(m)})$ (remember that $J^{(m)}$ is the length of the synthetic melody), which emits the lowest accumulated cost [19, Chapter 4].

4.3 Similarity Computation

Using either the Hough transform or SDTW, we obtain a path $p^{(m,n,k)} = \left(p_1^{(m,n,k)} \dots p_{L^{(m,n,k)}}^{(m,n,k)}\right)$ between the audio recording of the performance (m) and the score of the composition (n) using the tonic candidate c_k with $p_l^{(m,n,k)} = \left(r_l^{(m,n,k)}, q_l^{(m,n,k)}\right), r_l^{(m,n,k)} \in [1, I^m], q_l^{(m,n,k)} \in [1, J^m]$ and $l \in \left[1 : L^{(m,n,k)}\right]$, where $L^{(m,n,k)}$ is the length of the path $p^{(m,n,k)}$. We compute a similarity, $s^{(m,n,k)} \in [0:1]$, between the score fragment and the audio recording for the tonic candidate $c_k^{(m)}$ by:

$$s^{(m,n,k)} = \frac{\sum_l B(r_l^{(m,n,k)}, q_l^{(m,n,k)})}{L^{(m,n,k)}} \quad (5)$$

$s^{(m,n,k)}$ gives us a measure of how closely the score fragment is followed by the corresponding time-interval in the audio recording indicated by the path. For example if the difference between the matched values of the audio predominant melody and the synthetic predominant melody are always below 50 cents, the similarity is 1.

For the partial alignment between the score of the composition (n) and the audio recording (m), we obtain a set of alignment similarities as $S^{(m,n)} = \left\{s^{(m,n,1)}, \dots, s^{(m,n,K^{(m)})}\right\}$, where $s^{(m,n,k)}$ is the alignment similarity between the composition (m) and audio recording (n) for the tonic candidate $c_k^{(m)}, k \in \left[1 : K^{(m)}\right]$.

The similarity between the composition (n) and the audio recording (m) is simply taken as the maximum alignment similarity value, i.e: $s^{(m,n)} = max(S^{(m,n)})$.

Figure 2 show the similarities computed between the performances in our audio collection (Section 5.1) and the composition, "Acemaşiran Peşrevi." [3] In this example the similarity of the relevant audio recordings are much higher compared to the non-relevant ones.

Note that finding a true pair also implies correctly identifying the tonic pitch class [5], i.e. $\zeta^{(m,n)} = \arg\max_{c_k^{(m)}} (S^{(m,n)})$, where $\zeta^{(m,n)}$ is the estimated tonic pitch class of the performance of the composition in the audio recording.

4.4 Irrelevant Document Rejection

In many common retrieval scenarios, including composition identification, the users are only interested in checking the top documents [20]. After applying partial audio-score alignment between the query and each document, we rank the documents with respect to the similarities obtained. We then reject documents with low similarities according to an automatically learned threshold.

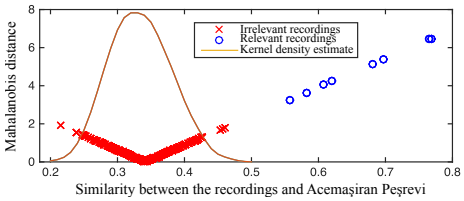

Figure 2. Similarity vs Mahalanobis distance between the composition "Acemaşiran Peşrevi" and the audio recordings in the dataset, and the kernel density-estimate computed from the similarity values between the audio recordings and the composition.

As seen in Figure 2, the relevant documents stand as "outliers" among the irrelevant documents with respect to the similarities they emit. To fetch the relevant documents per query, one can apply "outlier detection" using similarities between each document and query. Outlier detection is a common problem, which has many applications such as fraud detection and server malfunction detection [21].

Upon inspecting the similarity values emitted by irrelevant documents, we have noticed that the values roughly follow a Normal distribution (Figure 2). However, the distributions observed for each query have a different mean and variance. This is expected since the similarity computation could be affected by several factors such as the melodic complexities of the score fragment and the audio performance, as well as the quality of the extracted audio predominant melody. To deal with this variability, we compute the Mahalanobis distance of each similarity value to the distribution represented by the other similarity values (Figure 2). [4] Mahalanobis distance is a unitless and scale-invariant distance metric, which outputs the distance between a point and a distribution in standard deviations.

To reject irrelevant documents we apply a simple method where all documents below a certain threshold are rejected. To learn the decision boundary for thresholding, we apply logistic regression [20], a simple binary classification model, to the similarity values and the Mahalanobis distances on labeled data (Section 5.1). The training step will be explained in Section 5 in more detail.

After eliminating the documents according to the learned decision boundary, we add a last document called *none* to the end of the list. This document indicates that the query might not have any relevant document in the collection if all of the documents above are irrelevant.

5. EXPERIMENTS

In the experiments, we compare two alignment methods (Hough vs. SDTW). We try to align either the repetition in

[3] http://musicbrainz.org/work/01412a5d-1858-43b3-b5b0-78f383675e9b

[4] Note that the Mahalanobis distances shown in Figure 2 are less than what a "real" Normal distribution would produce. This is because of the contribution by the true pairs to the distribution.

the score as done in [5] or the start in the score as a simpler alternative and for the case when the structure information is not available in the score. We search the optimal fragment duration between 4 and 24 seconds.

As mentioned in Section 3, we evaluate the performance retrieval and the composition retrieval tasks separately. To test the document rejection step, we use 10-fold cross validation. We run the transposition-invariant partial audio score alignment between each score fragment and audio recording (Section 4.2) and then compute the similarity value for each performance-composition pair in the training set (Section 4.3). We also compute the Mahalanobis distance for each query (performance in composition retrieval task and vice versa). We apply logistic regression to the similarity values and the Mahalonobis distances computed for each annotated audio-score pair (with the binary relevances 0 or 1), and learn a decision boundary between the relevant and irrelevant documents. Then given a query (a composition in the performance retrieval task, and vice versa) from the testing set, we carry out all the steps explained in Section 4 and reject all the documents (performances in the performance retrieval task, and vice versa) "below" the decision boundary.

We use mean average precision (MAP) [20] to evaluate the methodology. MAP can be considered as a summary of how a method performs for different queries and the number of documents retrieved per query. For the document rejection step, we report the average MAP obtained from the MAPs of each testing set. We also conduct 3-way ANOVA tests on the MAPs obtained from each testing set to find if there are significant differences between the alignment methods, fragment locations and fragment durations. 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.

5.1 Dataset

For our experiments, we gathered a collection of 743 audio recordings and 146 music scores of different *peşrev* and *saz semaisi* compositions. The audio recordings are selected from the CompMusic corpus [22]. These recordings are either in public-domain or commercially available. The scores are selected from the SymbTr score collection [23]. SymbTr-scores are given in a machine readable format, which stores the duration and symbol of each note. The structural divisions in the compositions (i.e. the start and end note of each section) and the nominal tempo are also indicated in the scores.

We manually labeled the compositions performed in each audio recording. In the dataset there are 360 recordings associated with 87 music scores, forming 362 audio-score pairs. This information along with other relevant metadata such as the releases, performers and composers are stored in MusicBrainz. [5] Figure 3 shows the histogram of the number of relevant compositions per audio recording and the number of relevant audio recordings per composition. The number of recordings for a particular composition in our collection may be as many as 11. On the other hand, the releases of OTMM are typically organized such that there is a single composition performed in each track. For this reason, we were only able to obtain two audio recordings in which there are two compositions performed. Note that the tonic frequency changes in the performances of each composition in these two recordings.

[5] http://musicbrainz.org/

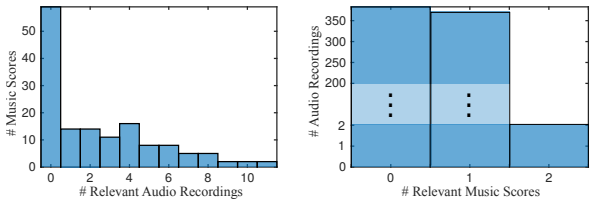

Figure 3. The number of relevant documents for the queries a) Histogram of the number of relevant audio recordings per score, b) Histogram of the number of relevant scores per audio recording

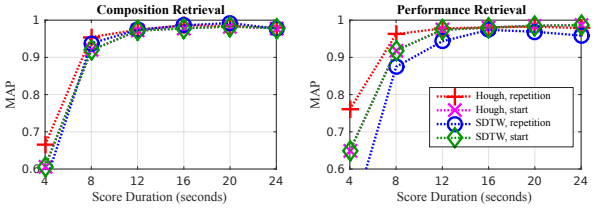

Figure 4. MAP for composition and performance retrieval task before document rejection, across different methods, fragment locations and durations. Only the queries with at least one relevant document are considered.

The average cardinalities of the compositions per audio recording and audio recordings per composition are 0.49 and 2.48, respectively. Notice that we have also included some compositions in our data collection, which do not have any relevant performances, and vice versa (Figure 3). Our methodology also aims to identify such queries without relevant documents. If we consider this case as an additional, special "document" called *none*, the average cardinality for compositions per audio recording and audio recordings per composition is 1.00 and 2.88, respectively.

5.2 Results

Before document rejection, the MAP is around 0.47 for both composition retrieval and performance retrieval tasks

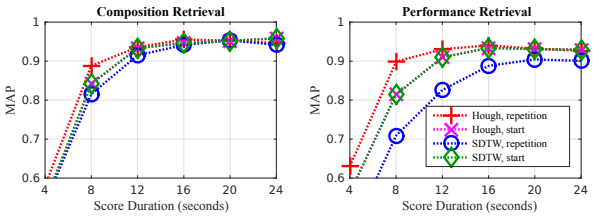

Figure 5. MAP for composition and performance retrieval task after document rejection, across different methods, fragment locations and durations. All queries are considered.

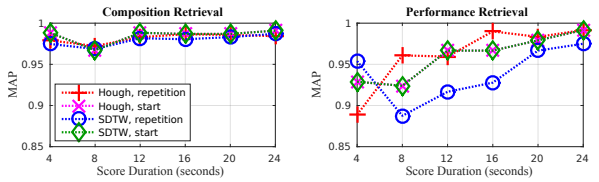

Figure 6. MAP for composition and performance retrieval task after document rejection, across different methods, fragment locations and durations. Only the queries with no relevant documents are considered.

Methods	Locations	Durations (sec.)					
		4	8	12	16	20	24
Hough	**Start**	30	15	2	3	2	2
	Repetition	14	5	0	0	0	0
SDTW	**Start**	32	6	3	3	3	3
	Repetition	24	3	1	2	3	3

Table 1. Number of errors in tonic identification

using either of the alignment methods, fragment locations and fragment durations longer than 8 seconds. The MAP is low before document rejection since the queries without relevant documents will practically have 0 average precision. Figure 4 shows the composition retrieval and performance retrieval results before document rejection only for the queries with relevant documents. The retrieval results before document rejection show that most of the audio-score pairs may be found by partial audio-score alignment by using a score fragment of at least 12 seconds. Although Hough transform performs slightly better than SDTW, these increases are not significant for fragment durations longer than 8 seconds.

Figure 5 shows the average MAPs from all queries obtained using different fragment durations, fragment locations and partial alignment methods in a 10-fold cross validation scheme. The best average MAP is 0.96 for composition retrieval using either the Hough transform or SDTW and aligning 24 seconds from the start. For performance retrieval the best average MAP of 0.95 is achieved using the Hough transform and aligning 16 seconds from the start. When we inspect average MAPs obtained from the queries without any relevant documents (Figure 6), we observe that the document rejection step always achieves an average MAP higher than 0.95 for all the parameter combinations in the composition retrieval task and an average MAP closer to or higher than 0.9 for all the parameter combinations in the performance retrieval task, respectively.

When we inspect the alignment results, we find that the score fragments were aligned properly for most of the cases. Moreover the tonic is identified almost perfectly for all the audio recordings by aligning the relevant scores (Table 1), and we achieved 100% accuracy out of the 362 audio-score pairs by aligning at least 12 seconds from the repetition using the Hough transform.

6. DISCUSSION

The results show that even aligning an 8 second fragment is highly effective, nevertheless, the optimal value of fragment duration for composition identification is around 16 seconds. Using a fragment duration longer than 16 seconds is not necessary since it increases the computation time without any significant benefit on identification performance. The results further show that aligning the start is sufficient, and there is no need to exploit the structure information to select a fragment from the repetition as in [5].

If a fragment of 16 seconds from the start of the score is selected, the Hough transform and SDTW produces the same results in both composition retrieval and performance retrieval tasks. One surprising case is the lower MAP's obtained in the performance retrieval task using SDTW to align the repetition. Although the drop is not significant for fragment durations longer than 12 seconds, we observed that SDTW tends to align irrelevant subsequences in the performances with the score fragments, which have similar note-symbol sequences but different durations.

Both the Hough transform and SDTW have a complexity of $O(I^{(m)}J^{(n)})$, where $I^{(m)}$ is the length of the predominant melody extracted from the audio recording (m) and $J^{(n)}$ is the length of the synthetic melody generated from the score of the composition (n). Nonetheless, the Hough transform is applied to a sparse, binary similarity matrix, hence it can operate faster than SDTW. Moreover, the Hough transform is a simpler algorithm. These properties make the Hough transform an alternative to more complex alignment algorithms, when precision in intra-alignment (e.g. note-level) is not necessary. Given these observations, we select alignment of the first 16 seconds of the score using the Hough transform as the optimal setting.

For the score fragments longer than 8 seconds, the tonic identification errors always occur in two historical recordings, where the recording speed (hence the pitch) is not stable and another recording where the musicians sometimes play the repetition by transposing the melodic intervals by a fifth. Even though the tonic identification has failed in these cases, the fragments are correctly aligned to the score. For such recordings, the stability of the tonic frequency can be assessed and the tonic frequency can be refined locally by referring to aligned tonic notes in the alignment path computed using SDTW.

From Figure 5, we can observe that by using a simple outlier detection step based on logistic regression, we were able to reject most of the irrelevant documents in both composition retrieval and performance retrieval scenarios. By comparing Figure 4 with Figure 5, we can also conclude that this step does not remove many relevant documents, providing reliable performance and composition matches. The usefulness of this step is more evident when the results for the queries with no relevant documents are checked (Figure 6). For such queries, since all the documents typically have a low, comparable similarity, our methodology is able to reject almost all the irrelevant documents. From Figure 6, we can also observe that the document rejection step is robust to changes in the fragment duration, the fragment location and the alignment method.

7. CONCLUSION

In this paper, we presented a methodology to identify the relevant compositions and performances in a collection consisting of audio recordings and music scores, using transposition invariant partial audio-score alignment. To the best of our knowledge, our methodology is the first automatic composition identification proposed for OTMM. The methodology is highly successful, achieving 0.95 MAP in retrieving the compositions performed in a recording and 0.96 MAP in retrieving the audio recordings where a composition is performed. What is more, our methodology is not only reliable in identifying relevant compositions and audio recordings but also identifying the cases when there are no relevant documents for a given query. Our algorithm additionally identifies the tonic frequency of the performance of each composition in the audio recording almost perfectly, as a result of partial audio-score alignment.

Our results indicate that the Hough transform can be a cheaper and effective alternative to alignment methods with more temporal flexibility such as SDTW in finding musically relevant patterns. As the next step we would like to evaluate our method on more forms, possibly with shorter structural elements such as the vocal form, *şarkı*. We would also like to investigate network analysis methods to identify the relevant performances and compositions jointly.

Our method can easily be adapted to neighboring music cultures such as Greek, Armenian, Azerbaijani, Arabic and Persian music, which share similar melodic characteristics. We hope that our method would be a starting point for future studies in automatic composition identification, and facilitate future research and applications on linked data, automatic music description, discovery and archival.

Acknowledgments

We are thankful to Dr. Joan Serrà for his suggestions on this work. 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).

8. REFERENCES

[1] D. P. Ellis and G. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 2007, pp. 1429–1432.

[2] J. Serrà, X. Serra, and R. G. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, no. 9, 2009.

[3] V. Thomas, C. Fremerey, M. Müller, and M. Clausen, "Linking sheet music and audio - Challenges and new approaches," in *Multimodal Music Processing*, ser. Dagstuhl Follow-Ups. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2012, vol. 3, pp. 1–22.

[4] Y. Raimond, S. A. Abdallah, M. Sandler, and F. Giasson, "The music ontology," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 417–422.

[5] S. Şentürk, S. Gulati, and X. Serra, "Score informed tonic identification for makam music of Turkey," in *Proceedings of 14th International Society for Music Information Retrieval Conference*. Curitiba, Brazil: Pontifícia Universidade Católica do Paraná, 2013, pp. 175–180.

[6] 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.

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

[8] S. Şentürk, A. Holzapfel, and X. Serra, "Linking scores and audio recordings in makam music of Turkey," *Journal of New Music Research*, vol. 43, no. 1, pp. 34–52, 3 2014.

[9] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, Universitat Pompeu Fabra, 2006.

[10] M. Müller, F. Kurth, and M. Clausen, "Audio matching via chroma-based statistical features." in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, 2005, p. 6th.

[11] 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.

[12] H. S. Atlı, B. Uyar, S. Şentürk, B. Bozkurt, and X. Serra, "Audio feature extraction for exploring Turkish makam music," in *3rd International Conference on Audio Technologies for Music and Media*, Bilkent University. Ankara, Turkey: Bilkent University, 2014.

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

[14] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, 1972.

[15] A. Holzapfel, U. Şimşekli, S. Şentürk, and A. T. Cemgil, "Section-level modeling of musical audio for linking performances to scores in Turkish makam music," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brisbane, Australia: IEEE, 2015, pp. 141–145.

[16] M. Müller and D. Appelt, "Path-constrained partial music synchronization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 65–68.

[17] B. Niedermayer, "Accurate audio-to-score alignment - data acquisition in the context of computational musicology," Ph.D. dissertation, Johannes Kepler Universität, Linz, February 2012.

[18] X. Anguera and M. Ferrarons, "Memory efficient subsequence DTW for Query-by-Example spoken term detection," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2013, pp. 1–6.

[19] M. Müller, *Information retrieval for music and motion*. Springer Heidelberg, 2007, vol. 6.

[20] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008, vol. 1.

[21] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.

[22] B. Uyar, H. S. Atlı, S. Şentürk, B. Bozkurt, and X. Serra, "A corpus for computational research of Turkish makam music," in *1st International Digital Libraries for Musicology Workshop*, London, United Kingdom, 2014, pp. 57–63.

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