TOWARDS ALIGNMENT OF SCORE AND AUDIO RECORDINGS OF OTTOMAN-TURKISH MAKAM MUSIC

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

1. ABSTRACT

Audio-score alignment is a multi-modal task, which facilitates many related tasks such as intonation analysis, structure analysis and automatic accompaniment. In this paper, we present a audio-score alignment methodology for the classical Ottoman-Turkish music tradition. Given a music score of a composition with structure (section) information and an audio performance of the same composition, our method first extracts a synthetic prominent pitch per section from the note values and durations in the score and a audio prominent pitch from the audio recording. Then it identifies the performed tonic frequency by using melodic information in the repetitive section in the score. Next it links each section with the time intervals where each section performed in the audio recording (i.e. structure level alignment) by comparing the extracted pitch features. Finally the score and the audio recordings are aligned in the note-level. For the initial experiments we chose DTW, a standard technique used in audio-score alignment, to show how well the state-of-the-art performs in makam musics. The results show that our method is able to handle the tonic transpositions and the structural differences with ease, however improvements, which address the characteristics of the music scores and the performances of makam musics, are needed in our note-level alignment methodology. To the best knowledge this paper presents the first audio-score alignment method proposed for makam musics.

2. INTRODUCTION

Audio recordings and scores are the two most relevant representations of music, both of which provide invaluable information about the melodic, metrical, expressive and cultural characteristics of the music. Audio-score alignment is a multimodal music information retrieval task, which aims to synchronise the musical events in an audio performance of a music piece with corresponding events in the score of the same piece. The aligned information from both sources may be used to enhance or facilitate tasks such as automatic accompaniment, audio-lyrics alignment, source separation, structure analysis, music discovery, tuning and intonation analysis, rhythmic analysis (Müller, 2007).

The current state-of-the-art focuses on aligning scores and audio of Eurogenetic musics. Nevertheless, incorporating knowledge specific to different music traditions in computational tasks might produce more accurate results (Şentürk et al., 2014). In this paper, we propose a audio-score alignment method which addresses some of specificities of Ottoman-Turkish *makam* music. To the best of our knowledge, this is the first audio-score alignment method applied to *makam* musics.

This remainder of the paper is structured as follows: Section 3 introduces the basic concepts of Ottoman-Turkish *makam* music, Section 4 explains the proposed methodology. Section 5 presents the data collection. Section 6 presents the experiments. Section 7 presents the results and the discussions and Section 8 give a brief conclusion.

3. MAKAM MUSIC

*Makam*s are modal structures, which are commonly used in Turkey, Middle East, North Africa, Greece and Balkans. Makams typically involve intervals smaller than a semitone. Moreover the notes are not equal tempered and the tuning/intonation of the intervals might differ with respect to the region, *makam* and performer. Each performance involves expressive usage of note repetitions, omissions, embellishments and timings (Ederer, 2011). Moreover, the musicians may decide to repeat, insert or omit musical phrases or even entire sections. It is also common to transpose the tonic of a performance due to instrument/vocal range or aesthetic concerns (Ederer, 2011). Makam music is also rich with heterophony, i.e. simultaneous variations of the same melody performed in the register of the instrument(s) or the voice(s) in the vocal pieces (Cooke, 2013). Typically the degree of heterophony increases with the number of instruments/voices, i.e. a solo ney recording is monophonic, while an ensemble/chorus recordings typically consist of complex heterophonic interactions.

In this paper, we focus on the classical Ottoman-Turkish tradition. Arel-Ezgi-Uzdilek (AEU) theory is the mainstream music theory used to explain the classical Ottoman-Turkish music. Note that the tuning and/or the intonation of a performed note might be substantially different from the theoretical interval (Bozkurt et al., 2009). We test our methodology on a collection of audio recordings and scores of *peşrev* form, which is one of the most common instrumental forms of classical tradition of the Ottoman-Turkish music. *Peşrev*s commonly consist of four distinct hanes and a teslim section, which typically follow a verse-refrain-like structure.

While *makam* musics are predominantly oral, in classical Ottoman-Turkish music a score representation extending the Western music notation is typically used to complement the practice (Popescu-Judetz, 1996). The scores do not show any embellishments, expressive decisions, heterophonic interactions, tuning or intonation information of the *makam* but only the basic melodic progression. The symbols in the music notation typically follow the rules of AEU music theory (Popescu-Judetz, 1996).

4. METHODOLOGY

We define *audio-score alignment* as synchronisation of the musical events in the score of a composition with the corresponding events in the audio recording of the same composition. In this paper we deal with two levels of granularity in the alignment: **1)** Section level, **2)** Note level. Our method addresses the some of main challenges of computational analysis of Ottoman-Turkish *makam* musics namely the transpositions, structural differences and the tuning.

Given a machine-readable score of a composition with the note symbols, the note durations and the annotated section information (i.e. the beginning and ending notes of each section) and an audio recording of the same performance, our method first extracts a prominent pitch of the audio recording [1] and synthesises a prominent pitch from the basic melody of each section indicated in the score such. In the synthesised prominent pitch of each section, the tonic symbol is assigned to 0 and the other note symbols are mapped to the melodic intervals defined by the AEU theory. The audio prominent pitch is also normalised such that the tonic frequency is assigned to 0. We then compute a distance matrix per section in the score between the synthesised prominent pitch of the section and the audio prominent pitch. We convert the distance matrix to a binary similarity matrix by assigning 1 to each point having a pitch distance less than 2.5 Holdrian commas ($\approx$ 56.6cents, a little bit higher than a quarter tone) and 0 otherwise. Binarisation takes care of the tuning and intonation differences between the synthesised and the performed pitches. In the binary matrices some pixels are distributed such that they form blobs similar to diagonal line segments. These line segments hint the locations of the sections in the audio. We use Hough transform (Duda & Hart, 1972), a common line detection algorithm to detect these lines and hence link the score sections with their corresponding time-intervals in the audio recording using the methodology. [2]

Note that for the above methodology to work the tonic frequency of the audio recording should be correctly identified. To handle this issue, we compute a pitch class distribution from the audio prominent pitch and extract the peaks in the distribution as tonic candidates. Assuming each peak as the tonic, attempts to linking the repetitive section using the methodology explained above. The tonic is identified as the pitch class which produces the most confident links. [3]

After we link the score sections with the audio recording, the time-interval of each section link is extended by 3 seconds to deal with the tempo differences. Then, we attempt to align the note events within each section with the corresponding time-intervals of the section links in the audio recording. For the preliminary experiments, we aim to present how well a standard alignment technique applied to other musics performs on makam musics.

We use dynamic time warping (DTW), which is a standard technique for audio-score alignment (Müller, 2007, Chapter 4). In order to avoid pathological warping during alignment, a known issue in DTW computation, we apply local constraints as discussed in (Sakoe & Chiba, 1978). We select the step size condition as $\{(2,1), (1,1), (1,2)\}$, which is a common local constraint in audio-score alignment (Müller, 2007, Chapter 4). In addition to the local constraint we also apply global constraint as discussed in (Sakoe & Chiba, 1978). The bandwidth of the global constraint is selected as 20% of the query length. Further, we also leverage the condition for the alignment path to be from the start of the strings to the end by implementing a subsequence version of the DTW as described in (Müller, 2007, Chapter 4).

For each section link, we apply subsequence DTW between the features audio prominent pitch extracted from each section link in the audio recording and the corresponding section in score. The extracted features are prominent pitch from the audio recording and synthetic prominent pitch from the section as explained above. We octave-wrap the features to deal with the octave errors and use City Block Distance (L1) for the distance computation at each step in DTW. Octave-wrapped City Block distance was previously shown as an effective and intuitive distance metric for comparing pitch values (Şentürk et al., 2014).

5. DATA COLLECTION

For the initial experiments, we collected 6 audio recordings of 4 *peşrev* compositions from the classical Ottoman-Turkish tradition. [4] The recordings are performed in a variety of transpositions. There are 51 sections in the audio recordings in total. The duration of the sections are 36.1 seconds on average with a standard deviation of 16.2 seconds.

The scores for each composition are obtained from the SymbTr collection (Karaosmanoğlu, 2012). The SymbTr-scores are machine-readable files, which contains note values on 53-TET (tone-equal tempered) resolution and note durations. These SymbTr-scores are divided into sections that represent structural elements in *makam* music. The beginning and ending notes of each section are indicated in the instrumental SymbTr-scores. These scores also follow the section sequence of the composition. [5]

[1] We use the Essentia implementation (Bogdanov et al., 2013) of the Melodia algorithm (Salamon & Gómez, 2012)

[2] For detailed information on the section linking methodology, we refer the reader to (Şentürk et al., 2014)

[3] For detailed information on the tonic identification methodology, we refer the reader to (Şentürk et al., 2014)

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

[5] The SymbTr-scores, the audio and composition metadata, the annotations and the complete results are available at *http://compmusic.upf.edu/node/218*.

symbTr-score	Audio MBID	Instrumentation	#Anno	t_p	f_p	f_n	$F_1\%$
beyati–pesrev–hafif—-seyfettin_osmanoglu	70a235be-074d-4b9b-8f94-b1860d7be887	ensemble	906	790	116	116	87.2
huseyni–pesrev–muhammes—-lavtaci_andon	8b78115d-f7c1-4eb1-8da0-5edc564f1db3	ensemble	614	482	132	132	78.5
	9442e4cf-0cb3-4cb3-a060-77aa37392501	ney & percussion	302	260	45	42	85.7
rast–pesrev–devrikebir—-giriftzen_asim_bey	31bf3d56-03d8-484e-b63c-ae5ae9a6e733	tanbur	658	374	306	281	56.0
	5c14ad3d-a97a-4e04-99b6-bf27f842f909	ney	673	418	262	255	61.8
segah–pesrev–devrikebir—-yusuf_pasa	e49f33b8-cf8a-4ca9-88cf-9a994dbad1c0	ney & kanun	743	267	490	476	35.6

Table 1: Results of note-level alignment per experiment

6. EXPERIMENTS

Given a score of a composition and an audio recording of the same composition, we align the note onsets in the symbTr-score of a composition with the corresponding audio performance of the same composition using the methodology explained in Section 4 and obtain aligned note onsets in the audio recording.

As the ground truth, we use the manual note annotations collected for the evaluation of (Benetos & Holzapfel, 2013). For the data collection explained in Section 4, the total number of the note annotations in the audio recordings are 3896. These annotations typically follow the note sequence in the symbTr. Note that there are 3 inserted and 49 omitted notes in the annotations with respect to the symbTr-scores.

To evaluate the tonic identification, we compare the distance between the pitch class of the estimated tonic and the pitch class of the annotated tonic as explained in (Şentürk et al., 2013). If the distance is less than 1 Hc, the estimation is marked as correct.

To evaluate section linking, we check the time distance between the time interval of annotated sections and sections links as explained in (Şentürk et al., 2014). A section link is marked as a true positive, if an annotation in the audio recording and the link has the same section label, and the link is aligned with the annotation, allowing a tolerance of ± 3 seconds. All links that do not satisfy these two conditions are considered as false positives. If a section annotation does not have any links in the vicinity of ± 3 seconds, it is marked as false negative.

To evaluate the note-level alignment, we compare the aligned onset and the corresponding annotated onset. We consider the aligned onset as a true positive if the distance is less than ± 200ms. If the distance is higher than ± 200ms, the aligned onset and the annotated onset are labeled as false positive and false negative, respectively. The insertions are ignored in the evaluation. If the aligned note corresponding to an omitted annotation is not rejected (i.e. the duration is non-zero), it is deemed as a false positive.

From these quantities we compute the F_1-scores for section linking and note-level alignment separately as:

$$P = \frac{t_p}{t_p + f_p}, \quad R = \frac{t_p}{t_p + f_n}, \quad F_1 = 2\,\frac{P\,R}{P + R} \qquad (1)$$

t_p, f_n, f_p, P, R and F_1 stand for number of true positives, number of false negatives, number of false positives, precision, recall and F_1-score, respectively.

7. RESULTS AND DISCUSSION

Across all the experiments, the tonic is identified correctly (100% accuracy in tonic identification) and all the sections were linked perfectly ($F_1 = 100\%$ for section linking). In the note level our methodology is able to align 2591 notes out of 3896 notes correctly, yielding to an F_1-score of 66.1%. The mean, median and standard deviation of the time-distance between the aligned note and the corresponding annotation are 299, 93 and 498 milliseconds, respectively. Moreover, 89.2% of the notes are aligned with a margin of ± 1 second, implying that DTW does not lose track of the melody.

Previously in (Şentürk et al., 2013) and (Şentürk et al., 2014) we showed that our linking methodology is highly reliable for tonic identification and section linking. The results in this paper also comply with these previous findings.

To understand the common mistakes in the note-level, we examined the aligned notes against annotated notes. Table 1 shows the results per experiment. The expressive embellishments in the performance (*portamento*s, *legato*s, *trill*s etc.) are common reasons of misalignment. For example, DTW infers portamentos as an insertion and the note onsets are aligned around the time when the portamento reaches to the stable note pitch. Similarly when there is a melodic interval less than a whole tone, a trill might cause a note onset to be marked earlier. Since these embellishments are not shown in the score, standard (subsequence) DTW was expected to fail. While we can argue that the section-level alignment is accurateXX, the results in the note-level alignment show that there is still more room for improvement for note-level alignment.

From Table 1, it can be seen that the note-level alignment fails for most of the notes in the audio recording of *Segah Peşrev* within the ± 200ms tolerance. This is a recording with ney and kanun, which consists of heterophonic interactions such as embellishments played by a single musician and time differences in note onsets between the performers. Due to such cases, the time distance between aligned onset and the annotated is typically larger than 200ms. Note that 75% of the notes are still aligned correctly within a tolerance of ± 1 second.

8. CONCLUSION

In this paper, we propose a method to align scores of *makam* musics with their associated audio recordings. Our system is able to handle the transpositions and structural repetitions and omissions in the audio recordings, which are common phenomenon in *makam* musics. The results obtained from the data collection present a proof-of-concept that a standard technique such as DTW can be effective for audio-score alignment for *makam* musics in the note level. Nevertheless, we need incorporate additional steps to handle non-notated embellishments and note omissions, insertions and repetitions.

Currently method relies on manual section segmentations in music scores. Manual segmentation of the score is not an difficult task compared to the note-level audio-score alignment itself. Nevertheless, it might be desirable to use other methodologies that do not require structural segmentations (e.g. (Gasser et al., 2013)), especially when we are working on large audio-score collections.

While we didn't have such an example in our data collection, there can be also omissions, insertions and repetition of phrases inside the sections. Currently, our methodology cannot handle such cases. In the future we want to use the JumpDTW proposed by (Fremerey et al., 2010) to handle phrase omissions, insertions and repetitions. Another approach might be segmentation of the symbolic score into melodic phrases and link extracted phrases from score with the corresponding audio recording. Recently, Bozkurt et al. (Bozkurt et al., pted) came with a method for segmenting music scores into melodic phrases according to the makam and usual information. Our initial experiments using the extracted phrases show that phrase linking is highly accurate. We observed that the erroneously linked phrases are almost identical to the true phrase, differing by very few pitches or durations, hence note-level alignment does not suffer a large number of errors.

We are extending the data collection to cover more examples from the CompMusic collection. In audio recordings with heterophonic interactions (such as the audio recording of *Segah Peşrev*) there is an ambiguity of the exact timings in the note onsets. To study the implications we plan to make several annotators, annotate the notes in the same set of scores and audio recordings. We will jointly compare the onset markings from each annotator with the aligned onsets produced by the future iterations of our automatic audio-score alignment method.

9. ACKNOWLEDGEMENTS

We would like to thank André Holzapfel for providing the note annotations. 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).

10. REFERENCES

Benetos, E. & Holzapfel, A. (2013). Automatic transcription of Turkish makam music. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*.

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

Bozkurt, B., Karaosmanoğlu, M. K., Karaçalı, B., & Ünal, E. (accepted). Usul and makam driven automatic melodic segmentation for Turkish music. *Journal of New Music Research*.

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

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

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

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

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

Gasser, M., Grachten, M., Arzt, A., & Widmer, G. (2013). Automatic alignment of music performances with structural differences. In *Proceedings of 14th International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 607–612)., Curitiba, Brazil.

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

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

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

Sakoe, H. & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *26*(1), 43–49.

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

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

Şentürk, S., Holzapfel, A., & Serra, X. (2014). Linking scores and audio recordings in makam music of Turkey. *Journal of New Music Research*, *43*, 34–52.