A Corpus for Computational Research of Turkish Makam Music*

Burak Uyar
Bahçeşehir University
Istanbul, Turkey
burakuyar@gmail.com

Hasan Sercan Atlı
Bahçeşehir University
Istanbul, Turkey
hsercanatli@gmail.com

Sertan Şentürk
Universitat Pompeu Fabra
Barcelona, Spain
sertan.senturk@upf.edu

Barış Bozkurt
Koç University
Istanbul, Turkey
barisbozkurt0@gmail.com

Xavier Serra
Universitat Pompeu Fabra
Barcelona, Spain
xavier.serra@upf.edu

ABSTRACT

Each music tradition has its own characteristics in terms of melodic, rhythmic and timbral properties as well as semantic understandings. To analyse, discover and explore these culture-specific characteristics, we need music collections which are representative of the studied aspects of the music tradition. For Turkish makam music, there are various resources available such as audio recordings, music scores, lyrics and editorial metadata. However, most of these resources are not typically suited for computational analysis, are hard to access, do not have sufficient quality or do not include adequate descriptive information. In this paper we present a corpus of Turkish makam music created within the scope of the CompMusic project. The corpus is intended for computational research and the primary considerations during the creation of the corpus reflect some criteria, namely, *purpose, coverage, completeness, quality* and *re-usability*. So far, we have gathered approximately 6000 audio recordings, 2200 music scores with lyrics and 27000 instances of editorial metadata related to Turkish makam music. The metadata include information about makams, recordings, scores, compositions, artists etc. as well as the interrelations between them. In this paper, we also present several test datasets of Turkish makam music. Test datasets contain manual annotations by experts and they provide ground truth for specific computational tasks to test, calibrate and improve the research tools. We hope that this research corpus and the test datasets will facilitate academic studies in several fields such as music information retrieval and computational musicology.

1. INTRODUCTION

For computational studies on a specific type of musics, there is a need for corpora, which constitutes the studied aspects of the specific music. A music corpus may consist of multiple information sources such as audio recordings, music scores, lyrics and editorial metadata. We can also group the corpora into two types: research corpus and test dataset [17]. A research corpus is a data collection that represents the "real world" for a specific research problem. A test dataset is a collection for a specific research task to test, calibrate and evaluate particular methodologies.

Serra [17] provides such criteria for the design of culture specific corpora, which are specified as *purpose, coverage, completeness, quality* and *re-usability*. To elaborate, the *purpose* of the corpora should be well-defined to facilitate research tasks. The corpora should be of good *coverage* to represent the music tradition and include metadata with a high degree of *completeness* related to studied aspects of the music. The corpora should attain a certain quality and it should be *re-usable* for future research. The Turkish Makam Corpus is designed with these considerations in mind.

In this paper we present a corpus for computational research of Turkish makam music. We explain the corpus with respect to the information sources that are used to populate it, namely audio recordings, machine-readable music scores and editorial metadata. For each type of data, we discuss the *purpose, coverage, completeness, quality* and *re-usability* criteria, when applicable. We also describe the test datasets of Turkish makam music we gathered in the scope of the CompMusic Project.

The rest of the paper is as follows: Section 2 gives a brief

summary of Turkish makam music. Section 3 explains the makam music research corpus and the criteria that we used for creating this corpus. Section 4 gives a detailed information about the test datasets and Section 5 wraps up the paper with a brief conclusion.

2. TURKISH MAKAM MUSIC

Most of the melodic, rhythmic and compositional aspects of Turkish makam music can be explained by the terms *makam, usul* and *form*, respectively. *Makam*s constitute the melodic structure of most of the traditional music repertoires in Turkey. Makams are modal structures, where the melodies typically revolve around an initial tone and a final tone [8]. The final tone is typically used synonymous to tonic. The rhythmic structure of Turkish makam music is described by the *usul*s. A certain *usul* can be described by a group of strokes with different velocities, which imply the beats and downbeats in the rhythmic pattern. Turkish makam music consists of both instrumental and vocal forms. Some of these forms are related with religious music and occasionally performed in religious ceremonies. There are also non-metered improvisational forms such as *taksim* and *gazel*.

Turkish Makam Music has been predominantly an oral tradition for centuries. For this reason the performance practice is the fundamental unit of Turkish makam music. There are several theories attempting to explain the makam practice. The mainstream theory is Arel-Ezgi-Uzdilek(AEU) [1]. The AEU theory divides a whole tone into 9 equidistant intervals, which can be approximated from 53-TET (tone equal tempered) intervals [19].

Since early 20th century, a score representation extending the traditional Western music notation has also been used for Turkish makam music [10]. The extended Western notation typically follows the rules of AEU theory. Most of the scores are transcriptions written sometimes centuries after the pieces were composed. They tend to notate simple melodic lines. Makam musicians follow the scores of compositions as the guideline but they extend them considerably during the performance. These include expressive timings, adding notes and non-notated embellishments. The intonation of some intervals in the performance might differ from the notated intervals as much as a semi-tone [18]. Due to these expressive decisions, there may be high degrees of variance between different interpretations of the same piece.

3. RESEARCH CORPUS

In Compmusic project, we mainly focus on the melodic and the rhythmic characteristics of Turkish makam music. To study these aspects of the music tradition, we have been collecting audio recordings and music scores. From the audio recordings we can extract the characteristics of interpretations of compositions performed by makam musicians. The music scores, on the other hand, provide an easy-to-access medium to extract the musical elements. We additionally store editorial metadata about Turkish makam music. The metadata includes information related to the audio recordings and music scores as well as additional information such as the birth date of the artists and relevant web sources about the entities. The metadata also include the relationships between each entity so that the connections within the

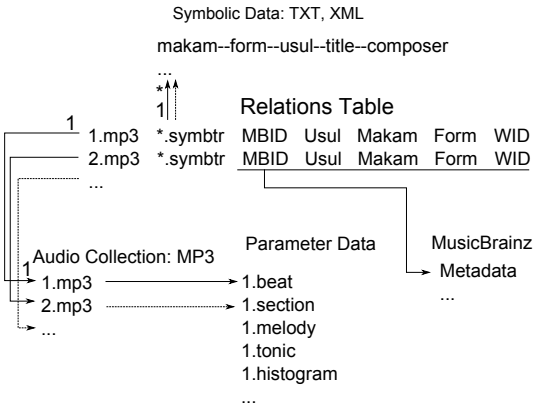

Figure 1: Block Diagram of the research corpus. A-MBID and W-MBID refers to the MBID of the audio recording and the MBID of the related work, respectively.

metadata can be exploited to access relevant information in a structured way.

For this purpose, a team with the support of more than 15 collaborators, has been working to collect and classify all the available data. In this section, we explain the audio recordings (Section 3.1), the music scores (Section 3.2) and the editorial metadata (Section 3.3) in the research corpus. The metadata related to audio recordings and the music scores are mainly explained within the corresponding source type. The corpus is discussed in terms of the *purpose, coverage, completeness, quality* and *reproducibility* of the audio recordings and the music scores[17]. In Section 3.3 we mostly focus on the overall statistics about the metadata as well as the statistics of inter-relationships.

In our corpus, we use MusicBrainz to store the metadata. MusicBrainz assigns a unique identifier, called MusicBrainz Identifier (MBID) to each entry (e.g. releases, audio recordings, artists). For more information on MBIDs please refer to `http://musicbrainz.org/doc/MusicBrainz_Identifier`.

3.1 Audio Collection

While creating the corpus, one of our major efforts has been directed to create an audio collection representative of Turkish makam music. The audio collection consists of almost 6000 stereo recordings, which are sampled at 44.1 kHz and 160 kbps and stored in MP3 audio format. This collection corresponds to 375 hours of play time. The collection includes both solo recordings and ensemble/chorus recordings. They span a time period from the start of the 20^{th} century to nowadays. The collection also covers various forms, some of which are part of the the folk (e.g. türkü) and religious repertoire (e.g. ilahi) (Table 2).

As part of the audio collection we have been also working on the annotation and extraction of various features from the audio recordings such as predominant melody, pitch histogram and tonic frequency. For predominant melody extraction, we use the Essentia implementation [3] of the

	#
Recordings	5953
Releases	340
Works	2696
Artists	536
Makams	150
Usuls	88
Forms	120

Table 1: Number of unique recordings, releases, works, artists, makams, usuls and forms in the CompMusic Turkish makam music corpus

methodology proposed in [12] [1]. The methodology proposed in [12] computes the pitch contours of given audio recordings. Then it classifies pitch contours as salient or non-salient and use salient contours to obtain the predominant melody. Since intervals with no predominant melody are very rare in Turkish makam music, we consider all pitch contours as salient. After collecting all of the pitch contours from the given audio recording, we sort them by their length. Starting from the longest pitch contour, we remove any overlapping pitch contours to obtain the predominant melody. If there are no pitch contours present for a given interval, that interval is deemed to be unpitched. From the predominant melody we compute fine-grained pitch histograms described in [4] and pitch class distributions using kernel-density estimation as explained in [14]. To obtain the annotated tonic frequencies we asked a number professional musicians to mark the time intervals where the tonic is played using the digital audio workstation of their choice. We selected the median of the predominant pitch values within these intervals as the annotated tonic frequency. So far the tonic frequencies of 3400 recordings have been marked. Along with the annotations, we also store the automatically identified tonic frequencies obtained from [4] and [14].

3.1.1 Coverage

Historically, Turkish Radio and Television Corporation (TRT) has the most representative audio productions of Turkish makam music. However, most of their audio collection is not open to public and only a small part of this collection is commercially available. Apart from TRT, there are numerous labels, which have released recordings of Turkish makam music. For these reasons it is hard to collect the overall statistics of Turkish makam music recordings.

So far, we have focused our efforts on gathering an audio collection of classical repertoire, including the available commercial recordings from TRT and other important labels. We also include several non-commercial recordings, provided that they have a good overall musical and production quality. In Table 1, we present the general statistics of the gathered audio collection.

3.1.2 Completeness

[1] `http://essentia.upf.edu/documentation/reference/std_PitchSalienceFunctionPeaks.html`

Form	#	Form	#
Şarkı	2378	Türkü	157
Taksim	1141	Ağır Semai	146
Peşrev	437	Beste	146
Saz Semaisi	390	İlahi	118
Yürük Semai	164	Other (92 forms)	836

Table 2: The most represented forms in the audio corpus and the corresponding number of audio recordings. Note that multiple compositions and improvisations might be performed in an audio recording. Hence an audio recording may have multiple forms associated with it.

	# Recordings	**% of total**
Releases	5953	100%
Works	4626	78%
Artists	5953	100%
Makams	5544	93%
Usuls	5349	90%
Forms	5770	97%

Table 3: The number of audio recordings for which the corresponding metadata is available. Note that 1141 audio recordings are improvisations, which do not have a work.

Along with the audio recordings, we also collect editorial metadata given in the album covers. In case an album cover does not provide related metada (e.g. related work, makam) we attempt to fill the missing metadata by accessing other information sources available. The procedure is as follows: if the makam, usul, form, composer information is missing, a search with the name of the recording is performed in the online score collections (such as the ones explained in 3.2.1), and the missing information is obtained from the matched score. For recording names made of form information (such as *"hicaz peşrev"*), since there can be many "hicaz peşrev", other "hicaz peşrev"s in the audio collection are listened and checked if there exists a match. If a match is found, its makam, form, usul and work information are copied.

The completeness of the audio related metadata is shown in Table 3. While checking the completeness of the artist metadata in the audio recordings, we assumed a recording is complete if it has at least one artist associated with it. Note that this does not imply a strict completeness with respect to the artist metadata since a lot of recordings (esp. ensemble recordings) do not have the complete information about the members in the album covers.

3.1.3 Quality

The audio recordings are stored in MP3 format. This format is chosen due to its small storage size compared to other audio formats. In the selection process, we did not include releases with low production quality (except the historical recordings of the grand masters) or with performances with low musical quality (e.g. MIDI accompaniment).

3.1.4 Re-Usability

The non-commercial recordings in our research corpus are freely-available. Most of these non-commercial recordings

can be downloaded from Internet Archive[2] or the respective websites where they were originally fetched from[3].

Due to copyright restrictions, we cannot distribute the commercial audio recordings and their cover arts in our collection. On the other hand, they are available for browsing and listening through Dunya [11]. Moreover the annotations on all the audio recordings and the various features extracted from them are freely distributed and can be used for computational research purposes.

3.2 Score-Collection

The existing music scores of Turkish Makam Music are mostly in physical formats, such as hand-written scores and books. There are also scores available in digital formats like JPEGs and PDFs. Typically, these types of scores are not very useful in computational research[4] since the musical elements (e.g. notes, durations, tempo, melodic structure, measure info) cannot be directly read by the machines.

In the scope of the CompMusic Project, a music score collection has been created called SymbTr [9]. The SymbTr collection consists of score-files in text format as well as the corresponding PDF and MIDI files. The naming of the text-scores follows a convention, which provides some of the key information to identify and categorize the compositions. This structure includes the *makam, form, usul, title* (for vocal compositions) and *composer name* of the music piece. Apart from the music scores, the collection also consists of *makam, usul* and *form* libraries, which provide additional structured musical information about these attributes. The symbols of each note in the music score are both given according to Arel-Ezgi-Uzdilek and 53-TET theory, as well as the corresponding pitch intervals in Holderian commas2. In addition to the note information, the nominal tempo, section information, beat information are given. In vocal compositions the lyrics are aligned to the notes in the syllable-level.

Recently, we released a second version of this score collection[5]. In this version, there are 2200 unique compositions. The general statistics are given in Table 4. In addition to the text-scores we provide the scores in MusicXML 3.0 format[6]. MusicXML is a format which can be imported/exported by the well-known score editing programs such as MuseScore, Finale and Sibelius.

Now we present the coverage, completeness and quality and re-usability of the updated SymbTr collection.

3.2.1 Coverage

To the best of our knowledge there are only two machine-readable score collections of Turkish makam music other than SymbTr, which can be used for computational research. First is the Uzun Hava Humdrum Database prepared by one of the authors of this paper [13]. This repository features the 77 music scores of *uzun hava*s, a non-metered improvisational form of Turkish folk music. Due to its specialized nature, this collection is not considered for comparison. A more relevant collection is the Türk Sanat Müziği Derlemi [2], which includes 600 compositions equally divided into 20 makams (i.e. 30 pieces per makam). It is smaller than the SymbTr collection.

To get a better means of comparison, we focus on online music score collections, in which the music scores are stored in various image formats. Although this repositories are not machine-readable, hence insuitable for computational research, they contain a much greater amount of music scores with respect to the machine-readable collections. This leads us to accept these collections as our references while measuring the coverage of our score collection.

Among these online collections, we selected TRT Tarihi Türk Müziği Arşivi (TRT-TTMA)[7] and the Türk Müziği Kültürünün Hafızası (TMKH) collections[8]. Similar to the case explained in Section 3.1.1, TRT-TTMA is arguably the most reliable resource. Currently, TRT-TTMA includes ~17,000 scores in total, all of which are manually scanned from physical scores. The scores in TRT-TTMA are sold online. TMKH is a collection created by funds from the Istanbul 2010 European Capital of Culture Organization through the European Union. The TMKH collection includes ~45,000 scanned scores (where multiple versions are available for almost each work) of the personal collections of 3 professional Turkish makam musicians/scholars. The collection is free, however there are several restrictions on the site navigation and the number of daily downloads. From these collections we crawled the metadata of the music scores to obtain the statistics (Table 4). TRT-TTMA has some duplicate entries and some compositions which are not in the context of Turkish makam music, *(e.g. church chants, operettas etc.)*. When these compositions are removed from comparison, the number of compositions are reduced to $\sim 12,000$.

Some of the names of the *makam, usul* and *form* in the collections slightly differ from each other. To match the names, we use an automatic string matching method. The algorithm we chose uses a weighted measure which consists of two edit distance measures[9]: longest common subsequence, and Damerau–Levenshtein distance, which have 0.7 and 0.3 weights respectively. The weights are determined empirically by varying them to find a configuration that results in satisfactory matches. Finally we do a manual check and remove any remaining erroneous matches.

To assess how well the SymbTr collection covers the Turkish makam music, we compare our collection against these music score collections. From each collection we report the number of compositions, composers, makams, forms and usuls. We also check how much the makams, forms and usuls in the SymbTr collection overlap with the corresponding type of metadata in other collections. We define overlap as:

[2] `http://tinyurl.com/n9omoue`
[3] e.g. `http://neyzen.com/ney_den_saz_eserleri.html`
[4] An obvious exception is optical music recognition.
[5] `https://github.com/MTG/SymbTr/releases/tag/v2.0.0`
[6] `http://www.musicxml.com/`
[7] `http://www.trtkulliyat.com/`
[8] `http://www.sanatmuziginotalari.com/`; accesible through `http://turkmusikisivakfi.org/`.
[9] The implementation is here: `https://github.com/gopalkoduri/string-matching/`

	SymbTr	TSM Derlemi	TRT-TTMA	TMKH
Compositions	2,200	600	12,035	45,368
Composers	455	230	1,447	2,674
Makams	157	20 (100%)	293 (49%)	317 (45%)
Usuls	84	46 (89%)	N/A	382 (22%)
Forms	62	6 (100%)	110 (35%)	90 (31%)

Table 4: Coverage of the score collection in the corpus. The number in paranthesis is the overlap measure 1 in percentage. N/A indicates that data is not available.

$$\mathscr{O} = \frac{|S_C \cap S_R|}{|S_R|} \tag{1}$$

where S_C is the set of the subjected attribute (*makam, usul* or *form*) from our collection (C), S_R is the set of the subjected attribute from the referance collection (R), against which we want to measure our collection's coverage and $\mathscr{O}$ is the overlap, which demonstrates how much the subjected attribute of R is represented in C.

Table 4 shows the overlap of the *makam*s, *usul*s, *form*s between our collection and the three music score collections explained above. We can observe that the SymbTr collection covers almost all of the makams, usuls and forms in the TSM Derlemi. While the number of compositions are much less than TRT-TTMA and TMKH, there is a fair number of overlapping makams, usuls and forms the the SymbTr collection with respect to TRT-TTMA and TMKH. Note that in Turkish makam music, it is common to have different titles for the scores of the same composition (first line of lyrics, the chorus etc.) and a composer might have various names (e.g. aliases, titles, added surname etc.). It is hard to obtain an accurate overlap for these attributes. Hence the overlap of the composers and the compositions are not computed.

Note that, the *makams, usuls* and *forms* listed in the score collections are not evenly distributed, some of these attributes are much more represented than the others. Hence we should also consider the coverage of these attributes with respect to their rate of presence in the reference collection. Taking these circumstances into consideration, we have modified the overlap function by adding some measure parameters as explained in detail below:

- The set (S_R) of an attribute from R has n elements.
- Each element k has an occurrence count o_k in the collection such that:

$$\sum_{k=1}^{n} o_k = |R| \tag{2}$$

  where $|R|$ indicates the number of scores in the reference collection, R. All k's are already ranked with respect to their occurances o_k in S_R such that $o_k \geq o_{k+1}$.
- The occurence ratio $o_k^{\%}$ is defined as:

$$o_k^{\%} = \frac{o_k}{|R|} \tag{3}$$

- Then we cumulatively add the ratios to find the total occurrence ratio ($O_k^{\%}$) up to k as,

$$O_k^{\%} = \frac{\sum_{k=1}^{n} o_k}{|R|} \tag{4}$$

  Notice that, $O_n^{\%} = 1$ as it includes all the score collection.
- We measure the overlap of C against $O_k^{\%}$'s.

$$\mathscr{O}_k = \frac{|S_C \cap S_R[1:k]|}{|S_R[1:k]|} \tag{5}$$

- Finally we define the attribute coverage $\mathscr{C}$ of C against R.

$$\mathscr{C} = \max(O_k^{\%}) \quad | \quad \mathscr{O}_k = 1 \tag{6}$$

By applying this modified procedure to our case, we have reached detailed results specifically different for the entities with different occurence ratios. In Figure 2, two distribution functions for $\mathscr{O}$ are provided with respect to *makams*. In these distributions, C corresponds to our collection and R corresponds TMKH and TRT-TTMA collections in each distribution. For TMKH reference, as soon as the *makams* which contribute less then 0.1% ($o_k^{\%} = 0.001$) to the collection are excluded, the coverage of TMKH by our collection increases to 100% ($\mathscr{O}_k = 1$). According to this specific k value, our collection overlaps with TMKH by 96% ($\mathscr{O} = 0.96$). In the case of TRT-TTMA, the overlap rate of our corpus reaches to 100% ($\mathscr{O}_k = 1$) at the point where the makams which constitute less than 0.6%($o_k^{\%} = 0.006$) of TRT-TTMA are excluded, providing the overlap value 76% ($\mathscr{O} = 0.76$) for TRT-TTMA. For the *form* attribute, the overlap is 98% ($\mathscr{O} = 0.98$) with ($o_k^{\%} = 0.002$) for TMKH and for TRT-TTMA the overlap is 86% ($\mathscr{O} = 0.86$) with ($o_k^{\%} = 0.01$). TRT-TTMA does not provide the *usul* information. Our corpus has an overlap of 94% ($\mathscr{O} = 0.94$) with ($o_k^{\%} = 0.003$) with THKM.

3.2.2 Completeness

As explained above, the SymbTr-scores have the *makam, usul, form* and *composer* information, as well as the beat information, nominal tempo, lyrics and section note boundaries and labels. The section boundaries in the vocal compositions indicated by single space and double spaces but the section names are not given in these cases. Therefore we can argue that SymbTr-scores are editorially complete except the section labels.

Note that the score metadata are not currently stored in MusicBrainz. However, we have been organizing the relationships between SymbTr scores and corresponding work

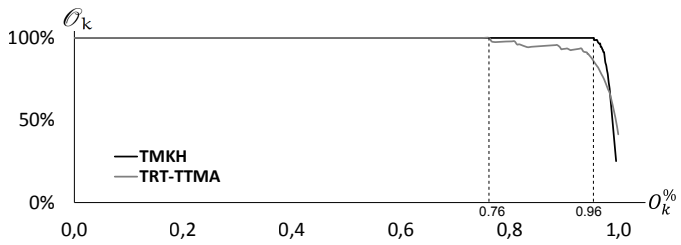

Figure 2: Overlap with respect to the *makam*, our corpus vs TMKH and TRT-TTMA. The dashed lines indiacate the coverage values, $\mathscr{C}_{\text{TRT}-\text{TTMA}} = 0.76$ and $\mathscr{C}_{\text{TMKH}} = 0.96$.

n	#works with SymbTr, related to n #recordings	total #audio recordings
1	259	259
2	164	328
3	113	339
4	81	324
5	48	240
6	32	192
7	19	133
8	10	80
9	6	54
10	3	30
11	5	55
12	1	12
Total	741	2046

Table 5: Number of works with SymbTr-scores distributed with respect to the related number of audio recordings

MBIDs, which will also be available in MusicBrainz. There is currently an intersection of 741 pieces, with 2046 corresponding recordings, between 2200 scores in our corpus and 2696 works in MusicBrainz. Detailed statistics of relation between the works in MusicBrainz and SymbTr scores can be seen in Table 5.

3.2.3 *Quality and Re-usability*

The aim of the MusicXML Makam Music Score Collection is to make this score library reachable for the interested community by providing it in a format that they can use in the software they are familiar with. In this way the community can make necessary changes on a certain work or create their own scores and share, by providing them in a machine-readable format which serves for widening the spectrum of the corpus for research. In this way, the corpus would represent the Turkish makam music tradition better and facilitate the research done on Turkish makam music.

Turkish Makam Music MusicXML Score Collection contains and supports all the accidental symbols in the notation of this music tradition. This provides users and researchers to work with the correct information of the music tradition.

The idea of having this score collection in MusicXML format solves multiple problems including the community content generation, community content examination, portability and re-usability. MusicXML is a format that can be used as a simple text file to be interpreted for research or that it can be used as a score file for the sheet music software. In both manners, the library can be used as a resource for infinitely many times as soon as the files are not changed on purpose. This is important because while organizing a community generated collection, we should use representations that facilitate users familiarity.

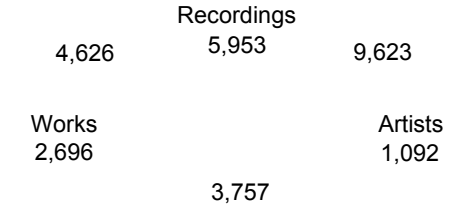

Figure 3: Number of audio recording, work and artist entities in the metadata and the number of relationships between each type of entity

3.3 Metadata

The metadata includes the general information about our corpus. It is mainly related to the audio recordings in the releases and the relevant compositions. The metadata also include other related information such as artist information, URLs to the related web pages. The metadata is interlinked with each other through relationships such that the instrument an artist played in an audio recording or the lyricist of a composition is known. We can use these relationships to navigate the concepts of Turkish makam music with ease.

So far we have collected more than 27000 entries of metadata. They are stored in Musicbrainz[10]. Figure 3 shows the number of entries related to recordings, works and artists and also the number of relationships between each entity. Among the metadata entered to Musicbrainz related to 7000 audio recordings from Turkey (pop, rock etc...), 6000 are metadata entered within the score of CompMusic Turkish makam music research corpus.

4. TEST DATASETS

Test datasets are collections arranged for the specific research problems. These datasets are typically used as the ground-truth to evaluate methodologies applied to certain problems. They can be composed of different types of data such as synthetic data or data with manual annotations.

Bozkurt et al. [5] made a review of computational analysis literature for Turkish makam music. The datasets that we mention in this section are useful for some of the research tasks discussed in this paper such as structure analysis, automatic tonic identification, automatic ornamentation segmentation, melodic phrase segmentation.

In our test datasets we have manual annotations by the experts of Turkish makam music tradition. The details of test datasets are discussed in section 4.1, 4.4, 4.5 and 4.2.

[10] `http://musicbrainz.org/collection/544f7aec-dba6-440c-943f-103cf344efbb`

4.1 Melodic Segmentation Test Dataset

Karaosmanoğlu and Bozkurt have studied the problem of usul and makam driven automatic melodic segmentation for Turkish Music [6]. For this research, 899 SymbTr-scores were manually annotated into melodic segments by 3 experts. There are a total of 31362 phrase annotations in this dataset [11].

4.2 Tonic Test Dataset

For score-informed tonic identification, we annotated the tonic frequency of 257 audio recordings associated with 57 compositions in total [14]. The SymbTr-score of the related composition for each audio recordings is given in the metadata.

Additionally, tonic frequencies of 3400 audio recordings have been manually annotated by the musicians.[12]

4.3 Section Test Dataset

For section linking experiments, we have also annotated the start and end of each section (as given in the corresponding SymbTr-score) in the same 257 audio recordings mentioned above (Section 4.2) [16]. The number of section annotations in the test dataset is 2095.

4.4 Audio-Score Alignment Test Dataset

For the initial experiments in audio-score alignment of Turkish makam music, we collected 6 audio recordings of 4 peşrev compositions [15]. The audio recordings in the dataset have the annotated tonic frequencies, 51 section annotations and 3896 note annotations in total. The note annotations follow the note sequences in the corresponding SymbTr-scores.

4.5 Audio-Lyrics Aligment Test Dataset

Dzhambazov has worked on automatic lyrics-to-audio alignment in Turkish Makam Music [7]. 10 şarkı were manually divided into sections and aligned to the recordings[13].

5. CONCLUSION

In this paper a research corpus of Turkish Makam Music is presented. The corpus is created under the considerations to meet some criteria: purpose, quality, completeness, coverage and re-usability. We also present some test datasets, which have been used to test and calibrate some computational tasks [6, 7, 14, 15, 16]. We hope that this research corpus and the test datasets will facilitate academic studies in several fields such as music information retrieval and computational musicology.

6. ACKNOWLEDGMENTS

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).

[11] http://akademik.bahcesehir.edu.tr/~bbozkurt/112E162_en.html

[12] http://compmusic.upf.edu/node/230

[13] http://compmusic.upf.edu/node/226

7. REFERENCES

[1] H. S. Arel. *Türk Musikisi Nazariyatı (The Theory of Turkish Music)*. ITMKD yayınları, 1968.

[2] N. B. Atalay and S. Yöre. Türk sanat müziği derlemi. www.tsmderlemi.com, 2011.

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

[4] B. Bozkurt. An automatic pitch analysis method for Turkish maqam music. *Journal of New Music Research*, 37(1):1–13, 2008.

[5] B. Bozkurt, R. Ayangil, and A. Holzapfel. Computational analysis of turkish makam music: Review of state-of-the-art and challenges. *Journal of New Music Research*, 43(1):3–23, 2014.

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

[7] G. Dzhambazov, S. Şentürk, and X. Serra. Automatic lyrics-to-audio alignment in classical turkish music. In *4th International Workshop on Folk Music Analysis*, Istanbul, Turkey, 12/06/2014 2014.

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

[9] 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)*, pages 223–228, 2012.

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

[11] A. Porter, M. Sordo, and X. Serra. Dunya: A system for browsing audio music collections exploiting cultural context. In *Proceedings of 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, 11 2013.

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

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

[14] 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 (ISMIR)*, pages 175–180, Curitiba, Brazil, 11 2013.

[15] S. Şentürk, S. Gulati, and X. Serra. Towards alignment of score and audio recordings of ottoman-turkish makam music. In *4th International Workshop on Folk Music Analysis*, Istanbul, Turkey, 06 2014.

[16] S. Şentürk, A. Holzapfel, and X. Serra. Linking scores and audio recordings in makam music of Turkey. *Journal of New Music Research*, 43:34–52, 03/2014

2014.

[17] X. Serra. Creating research corpora for the computational study of music: the case of the compmusic project. In *AES 53rd International Conference on Semantic Audio*, London, UK, 27/01/2014 2014. AES, AES.

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

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